SIAM J. OPTIMIZATION
Vol. 2, No. 1, pp. 1-6, February 1992

001

CLASSIFICATION OF CRITICAL STATIONARY POINTS IN UNCONSTRAINED OPTIMIZATION*

STEFAN SCHÄFFLER†

Abstract. A stationary point of an unconstrained optimization problem is called critical if the Hessian matrix at this point is positive semidefinite. Such a point cannot be classified using second-order optimality conditions. In this paper the problem of classifying a critical stationary point of a given objective function is reduced to the application of higher-order optimality conditions for a special auxiliary function.

Key words. optimality conditions, critical stationary points, positive-semidefinite Hessian

AMS(MOS) subject classifications. 65K05, 90C30

1. Introduction. We consider the following unconstrained minimization problem:

$$\min_{\mathbf{x}} \{f(\mathbf{x})\}, \qquad f:\mathbb{R}^n \to \mathbb{R}, \qquad f \in C^p, \qquad p > 2. \tag{1.1}$$

It is assumed that we have computed a critical stationary point $\mathbf{x}^*$ of problem (1.1), which is defined as follows.

DEFINITION 1.1. For problem (1.1) a point $\mathbf{x}^*$ is called a critical stationary point if

(a) $\nabla f(\mathbf{x}^*) = \mathbf{0}$,

(b) $\nabla^2 f(\mathbf{x}^*)$ is positive semidefinite.

Using the second-order optimality conditions (see, e.g., [6]) it is not possible to decide whether $\mathbf{x}^*$ is an (isolated) minimizer of f or not.

In this paper we use a catastrophe theoretical approach for analyzing critical stationary points. The main results are given in the next section. In § 3 the applications of these results in combination with higher-order optimality conditions are shown. Finally, examples are discussed.

2. Main results. Let $\mathbf{x}^*$ be a critical stationary point of f. Without loss of generality we introduce the following simplification:

$$\nabla^2 f(\mathbf{x}^*) = \begin{pmatrix} 1 & & & \mathbf{0} \\ & \ddots & & \\ & & 1 & \\ \mathbf{0} & & & \mathbf{0} \end{pmatrix} \begin{matrix} \}\, r \quad (0 \leqq r < n) \\ \\ \}\, n-r \end{matrix}, \tag{2.1}$$

and we define

$$\mathbf{x}^r := \begin{pmatrix} x_1 \\ \vdots \\ x_r \end{pmatrix}, \qquad \mathbf{x}^{n-r} := \begin{pmatrix} x_{r+1} \\ \vdots \\ x_n \end{pmatrix},$$

where r is called the rank of f. The following theorem is the main theoretical instrument for analyzing f at the critical stationary point $\mathbf{x}^*$.

* Received by the editors June 25, 1990; accepted for publication November 29, 1990.

† Institut für Angewandte Mathematik und Statistik, Technische Universität München, Arcisstrasse 21, 8000 München 2, Germany.

THEOREM 2.1. *Consider problem* (1.1). *Let* $\mathbf{x}^*$ *be a critical stationary point of f and let* $r>0$ *be the rank of f. Then there exist functions*

$$g: U(\mathbf{x}^{*n-r})\subset\mathbb{R}^{n-r}\to\mathbb{R}^r,\qquad g\in C^{p-1},\qquad g \text{ unique},$$

where $U(\mathbf{x}^{*n-r})$ *is an open neighbourhood of* $\mathbf{x}^{*n-r}$, *and*

$$\phi: U(\mathbf{x}^{*r})\times U(\mathbf{x}^{*n-r})\subset\mathbb{R}^n\to\mathbb{R}^r,\qquad \phi\in C^{p-3},$$

where $U(\mathbf{x}^{*r})$ *is an open neighbourhood of* $\mathbf{x}^{*r}$, *and the following statements hold*:

(a)
$$f(\mathbf{x}^r,\mathbf{x}^{n-r})=\|\phi(\mathbf{x}^r+\mathbf{x}^{*r}-g(\mathbf{x}^{n-r}),\mathbf{x}^{n-r})\|_2^2+f(g(\mathbf{x}^{n-r}),\mathbf{x}^{n-r})$$
for all $(\mathbf{x}^r,\mathbf{x}^{n-r})\in U(\mathbf{x}^{*r})\times U(\mathbf{x}^{*n-r})$.

(b) *For each fixed* $\mathbf{x}_0^{n-r}\in U(\mathbf{x}^{*n-r})$, *the following inequality is valid for all* $\mathbf{x}^r\in U(\mathbf{x}^{*r})$:

$$f(\mathbf{x}^r,\mathbf{x}_0^{n-r})\geqq f(g(\mathbf{x}_0^{n-r}),\mathbf{x}_0^{n-r}),$$

where equality holds if and only if $\mathbf{x}^r=g(\mathbf{x}_0^{n-r})$.

(c) *Using* $\bar f: U(\mathbf{x}^{*n-r})\to\mathbb{R}$, $\bar f(\mathbf{x}^{n-r})=f(g(\mathbf{x}^{n-r}),\mathbf{x}^{n-r})$, *we obtain*

$$\nabla\bar f(\mathbf{x}^{*n-r})=\mathbf{0},\qquad \nabla^2\bar f(\mathbf{x}^{*n-r})=\mathbf{0}.$$

Proof. (a) This part of the proof is based on a proof of the splitting lemma given by Castrigiano and Hayes-Widmann [1]. From the assumptions, we know that $\nabla_{\mathbf{x}^r}f(\mathbf{x}^{*r},\mathbf{x}^{*n-r})=\mathbf{0}$ and $\nabla^2_{\mathbf{x}^r}f(\mathbf{x}^{*r},\mathbf{x}^{*n-r})=\boldsymbol{E}_r$, where $\boldsymbol{E}_r$ denotes the r-dimensional identity matrix. Using the implicit function theorem we obtain the existence of the open neighbourhood $U(\mathbf{x}^{*n-r})$ and the unique function $g: U(\mathbf{x}^{*n-r})\to\mathbb{R}^r$, with

$$\nabla_{\mathbf{x}^r}f(g(\mathbf{x}^{n-r}),\mathbf{x}^{n-r})\equiv\mathbf{0}\quad\text{on } U(\mathbf{x}^{*n-r}). \tag{2.2}$$

This equation shows that $g\in C^{p-1}$.

Now we consider the C^{p-1} function

$$h(\mathbf{y}^r,\mathbf{x}^{n-r}):=f(\mathbf{y}^r-\mathbf{x}^{*r}+g(\mathbf{x}^{n-r}),\mathbf{x}^{n-r})-\bar f(\mathbf{x}^{n-r}),\qquad \mathbf{y}^r\in\mathbb{R}^r.$$

For all $\mathbf{x}^{n-r}\in U(\mathbf{x}^{*n-r})$ we obtain

$$h(\mathbf{x}^{*r},\mathbf{x}^{n-r})=0,\qquad \nabla_{\mathbf{y}^r}h(\mathbf{x}^{*r},\mathbf{x}^{n-r})=\mathbf{0}.$$

Hence, using Taylor expansion the function h is given by

$$h(\mathbf{y}^r,\mathbf{x}^{n-r})=(\mathbf{y}^r-\mathbf{x}^{*r})^T\mathbf{K}(\mathbf{y}^r,\mathbf{x}^{n-r})(\mathbf{y}^r-\mathbf{x}^{*r}),$$

where $\mathbf{K}(\mathbf{y}^r,\mathbf{x}^{n-r})$ is a C^{p-3} function with

$$\mathbf{K}(\mathbf{x}^{*r},\mathbf{x}^{*n-r})=\tfrac{1}{2}\mathbf{E}_r. \tag{2.3}$$

Thus, we have the following representation of h for all $(\mathbf{y}^r,\mathbf{x}^{n-r})\in U\times U(\mathbf{x}^{*n-r})$, where U is some open neighbourhood of $\mathbf{x}^{*r}$:

$$h(\mathbf{y}^r,\mathbf{x}^{n-r})=\|\phi(\mathbf{y}^r,\mathbf{x}^{n-r})\|_2^2,$$

with some C^{p-3} function ϕ: $U\times U(\mathbf{x}^{*n-r})\to\mathbb{R}^n$. Now we use the variable transformation $\chi: U\times U(\mathbf{x}^{*n-r})\to U(\mathbf{x}^{*r})\times U(\mathbf{x}^{*n-r})$,

$$(\mathbf{x}^r,\mathbf{x}^{n-r})=\chi(\mathbf{y}^r,\mathbf{x}^{n-r})=(\mathbf{y}^r-\mathbf{x}^{*r}+g(\mathbf{x}^{n-r}),\mathbf{x}^{n-r}),$$

where $U(\mathbf{x}^{*r})$ is some open neighbourhood of $\mathbf{x}^{*r}$ and χ is a diffeomorphism (U and $U(\mathbf{x}^{*r})$ are chosen sufficiently small). Furthermore, there exists an open neighbourhood $V=V_1\times V_2$ of $(\mathbf{x}^{*r},\mathbf{x}^{*n-r})$, $V_1\subset U(\mathbf{x}^{*r})$, $V_2\subset U(\mathbf{x}^{*n-r})$, with

$$\phi(\mathbf{y}^r,\mathbf{x}^{n-r})=\mathbf{0}\ \Leftrightarrow\ \mathbf{y}^r=\mathbf{x}^{*r}\quad\text{for all } (\mathbf{y}^r,\mathbf{x}^{n-r})\in\chi^{-1}(V_1\times V_2),$$

and thus

$$\phi(\mathbf{x}^r+\mathbf{x}^{*r}-g(\mathbf{x}^{n-r}),\mathbf{x}^{n-r})=\mathbf{0}\Leftrightarrow(\mathbf{x}^r,\mathbf{x}^{n-r})=(g(\mathbf{x}^{n-r}),\mathbf{x}^{n-r})\tag{2.4}$$

for all $(\mathbf{x}^r,\mathbf{x}^{n-r})\in V$ (see (2.3)).

Finally, we obtain for all $(\mathbf{x}^r,\mathbf{x}^{n-r})\in U(\mathbf{x}^{*r})\times U(\mathbf{x}^{*n-r})$:

$$h(\mathbf{x}^r+\mathbf{x}^{*r}-g(\mathbf{x}^{n-r}),\mathbf{x}^{n-r})=f(\mathbf{x}^r,\mathbf{x}^{n-r})-\bar{f}(\mathbf{x}^{n-r})=\|\phi(\mathbf{x}^r+\mathbf{x}^{*r}-g(\mathbf{x}^{n-r}),\mathbf{x}^{n-r})\|_2^2.$$

(b) If $V_1\neq U(\mathbf{x}^{*r})$ and $V_2\neq U(\mathbf{x}^{*n-r})$, then we redefine $U(\mathbf{x}^{*r})$ and $U(\mathbf{x}^{*n-r})$ as follows: $U(\mathbf{x}^{*r})=V_1$ and $U(\mathbf{x}^{*n-r})=V_2$. Hence, the proof is finished with (2.4).

(c) We obtain, by differentiation of identity (2.2) with respect to $\mathbf{x}^{n-r}$,

$$\nabla_{\mathbf{x}^{n-r}}g(\mathbf{x}^{*n-r})=\mathbf{0}.$$

Using this result, the rest of the proof is obvious. □

Now we prove an important corollary of Theorem 2.1.

COROLLARY 2.2. *Consider the function* $\bar{f}$ *defined in Theorem* 2.1(c); *then* $(\mathbf{x}^{*r},\mathbf{x}^{*n-r})$ *is an (isolated) local minimizer of* f *if and only if* $\mathbf{x}^{*n-r}$ *is an (isolated) local minimizer of* $\bar{f}$.

Proof. Under the assumption that $(\mathbf{x}^{*r},\mathbf{x}^{*n-r})$ is an (isolated) minimizer of f, we obtain an open neighbourhood N of $(\mathbf{x}^{*r},\mathbf{x}^{*n-r})$, $N\subset U(\mathbf{x}^{*r})\times U(\mathbf{x}^{*n-r})$, with

$$f(\mathbf{x}^r,\mathbf{x}^{n-r})\geqq f(\mathbf{x}^{*r},\mathbf{x}^{*n-r})\quad\text{for all } (\mathbf{x}^r,\mathbf{x}^{n-r})\in N.$$

(If $(\mathbf{x}^{*r},\mathbf{x}^{*n-r})$ is an isolated local minimizer of f, then we obtain the above inequality with "$>$" instead of "$\geqq$" for all $(\mathbf{x}^r,\mathbf{x}^{n-r})\in N\backslash\{(\mathbf{x}^{*r},\mathbf{x}^{*n-r})\}$.) Now consider the function $\pi: U(\mathbf{x}^{*n-r})\to\mathbb{R}^n$:

$$\pi(\mathbf{x}^{n-r})=(g(\mathbf{x}^{n-r}),\mathbf{x}^{n-r}).$$

π is a continuous function and therefore $W:=\pi^{-1}(N)$ is an open neighbourhood of $\mathbf{x}^{*n-r}$. Hence, we obtain for all $\mathbf{x}^{n-r}\in W\subset U(\mathbf{x}^{*n-r})$:

$$\bar{f}(\mathbf{x}^{n-r})=f(\pi(\mathbf{x}^{n-r}))\geqq f(\mathbf{x}^{*r},\mathbf{x}^{*n-r})=\bar{f}(\mathbf{x}^{*n-r}).$$

(If $(\mathbf{x}^{*r},\mathbf{x}^{*n-r})$ is an isolated local minimizer of f, then we obtain the above inequality with "$>$" instead of "$\geqq$" for all $\mathbf{x}^{n-r}\in W\backslash\{\mathbf{x}^{*n-r}\}$.)

Now let $\mathbf{x}^{*n-r}$ be an (isolated) minimizer of $\bar{f}$; then there exists an open neighbourhood $\bar{W}\subset U(\mathbf{x}^{*n-r})$ of $\mathbf{x}^{*n-r}$, with:

$$\bar{f}(\mathbf{x}^{n-r})\geqq\bar{f}(\mathbf{x}^{*n-r})\quad\text{for all } \mathbf{x}^{n-r}\in\bar{W}.$$

(If $\mathbf{x}^{*n-r}$ is an isolated local minimizer of $\bar{f}$, then we obtain the above inequality with "$>$" instead of "$\geqq$" for all $\mathbf{x}^{n-r}\in\bar{W}\backslash\{\mathbf{x}^{*n-r}\}$.)

With Theorem 2.1 we obtain for all $(\mathbf{x}^r,\mathbf{x}^{n-r})\in U(\mathbf{x}^{*r})\times\bar{W}$:

$$f(\mathbf{x}^r,\mathbf{x}^{n-r})=\|\phi(\mathbf{x}^r+\mathbf{x}^{*r}-g(\mathbf{x}^{n-r}),\mathbf{x}^{n-r})\|_2^2+\bar{f}(\mathbf{x}^{n-r})\geqq f(\mathbf{x}^{*r},\mathbf{x}^{*n-r}).$$

(If $\mathbf{x}^{*n-r}$ is an isolated local minimizer of $\bar{f}$, then we obtain for all

$$(\mathbf{x}^r,\mathbf{x}^{n-r})\in(U(\mathbf{x}^{*r})\times\bar{W})\backslash\{(\mathbf{x}^{*r},\mathbf{x}^{*n-r})\}$$

the above inequality with "$>$" instead of "$\geqq$".) □

The results of Corollary 2.2 allow us the classification of the critical stationary point $\mathbf{x}^{*n-r}$ of $\bar{f}$ (with $\nabla^2\bar{f}(\mathbf{x}^{*n-r})=\mathbf{0}$) instead of $(\mathbf{x}^{*r},\mathbf{x}^{*n-r})$. If rank $f=0$, then we define $\bar{f}\equiv f$. Hence, higher-order optimality conditions are available.

3. Higher-order optimality conditions. Consider a function $h:\mathbb{R}^n\to\mathbb{R}$, $h\in C^k$, $k>2m$, $m\in\mathbb{N}$. Let the kth derivative of h at any point $\bar{\mathbf{x}}$ be given by the tensor $\nabla^k h(\bar{\mathbf{x}})=(t_{i_1,\dots,i_k})$, $i_j=1,\cdots,n$ for $j=1,\cdots,k$, where

$$(3.1)\qquad t_{i_1,\cdots,i_k}=\frac{\partial^k h}{\partial x_{i_1},\cdots,\partial x_{i_k}}(\bar{\mathbf{x}}).$$

The computation of this tensor at each fixed but arbitrary chosen point is possible using automatic differentiation (see, e.g. [3], [4]). For each fixed $\mathbf{s}\in\mathbb{R}^n$ we define

$$(3.2)\qquad \nabla^k h(\bar{\mathbf{x}})*\mathbf{s}:=\sum_{i_1=1}^{n}\left(\cdots\left(\left(\sum_{i_k=1}^{n}t_{i_1,\cdots,i_k}\cdot s_{i_k}\right)\cdot s_{i_{(k-1)}}\right)\cdots\right)\cdot s_{i_1}.$$

Now we are able to prove the following higher-order optimality conditions.

THEOREM 3.1 (necessary condition). *Let* $\mathbf{x}^*$ *be a local minimizer of* $h\in C^k$, $k>2m$, $m\in\mathbb{N}$ *such that*

$$\nabla^i h(\mathbf{x}^*)=\mathbf{0}\quad \textit{for all } i=1,\cdots,2m,$$

then

$$\nabla^{2m+1}h(\mathbf{x}^*)=\mathbf{0}.$$

Proof. From (3.2) we know that $\nabla^{2m+1}h(\mathbf{x}^*)*\mathbf{x}$ is a polynomial function in $\mathbf{x}\in\mathbb{R}^n$. Using the special symmetry of the tensor $\nabla^{2m+1}h(\mathbf{x}^*)$ (see (3.1)) and the fact that a polynomial function in several variables is the zero function if and only if all coefficients are zero, we obtain:

$$\nabla^{2m+1}h(\mathbf{x}^*)*\mathbf{x}\equiv 0\quad\text{for all }\mathbf{x}\in\mathbb{R}^n\quad\text{iff }\nabla^{2m+1}h(\mathbf{x}^*)=\mathbf{0}.$$

Now we consider the C^k function $\zeta:\mathbb{R}\to\mathbb{R}$, $\zeta(\sigma)=h(\mathbf{x}^*-\sigma\mathbf{s})$, $\mathbf{s}\in\mathbb{R}^n$, $\mathbf{s}\neq\mathbf{0}$. We obtain:

$$\frac{d^i\zeta}{d\sigma^i}(0)=0\quad\text{for all } i=1,\cdots,2m,$$

$$\frac{d^{2m+1}\zeta}{d\sigma^{2m+1}}(0)=-\nabla^{2m+1}h(\mathbf{x}^*)*\mathbf{s}.$$

Hence, if $\nabla^{2m+1}h(\mathbf{x}^*)\neq\mathbf{0}$, then there exists an $\mathbf{s}\in\mathbb{R}^n$, $\mathbf{s}\neq\mathbf{0}$ with

$$\frac{d^{2m+1}\zeta}{d\sigma^{2m+1}}(0)<0.$$

Thus, the function h decreases starting from $\mathbf{x}^*$ along the direction of $\mathbf{s}$. This is a contradiction to the assumption of Theorem 3.1. □

Now we come back to the problem of classifying the critical stationary point $\mathbf{x}^*$ of f (see (1.1)). We know that $\mathbf{x}^*$ is an (isolated) local minimizer of f if and only if $\mathbf{x}^{*n-r}$ is an (isolated) local minimizer of $\bar{f}$, with

$$\nabla\bar{f}(\mathbf{x}^{*n-r})=\mathbf{0},$$

$$\nabla^2\bar{f}(\mathbf{x}^{*n-r})=\mathbf{0}\quad\text{(see Theorem 2.1 and Corollary 2.2).}$$

If $f\in C^p$, then $\bar{f}\in C^{p-1}$. The derivatives of $\bar{f}$ at $\mathbf{x}^{*n-r}$ can be computed using (2.2) and implicit automatic differentiation (see [5]). Hence, for $p\geqq 4$, Theorem 3.1 is applicable for classifying $\mathbf{x}^{*n-r}$ and consequently $\mathbf{x}^*$. If $\bar{f}\in C^{2m+2}$, $m\in\mathbb{N}$, and

$\nabla^i \bar{f}(\mathbf{x}^{*n-r}) = \mathbf{0}$ for all $i = 1, \cdots, 2m+1$, then the following two optimality conditions are applicable.

THEOREM 3.2 (necessary condition). *Let* $\mathbf{x}^*$ *be a local minimizer of* $h \in C^k$, $k > 2m+1$, $m \in \mathbb{N}$ *such that*

$$\nabla^i h(\mathbf{x}^*) = \mathbf{0} \quad \textit{for all } i = 1, \cdots, 2m+1,$$

then

$$\nabla^{2m+2} h(\mathbf{x}^*) * \mathbf{x} \geqq \mathbf{0} \quad \textit{for all } \mathbf{x} \in \mathbb{R}^n.$$

Proof. Let us consider the C^{2m+2} function $\zeta\colon \mathbb{R} \to \mathbb{R}$, $\zeta(\sigma) = h(\mathbf{x}^* - \sigma \mathbf{s})$, $\mathbf{s} \in \mathbb{R}^n$, $\mathbf{s} \neq \mathbf{0}$. We obtain:

$$\frac{d^i \zeta}{d\sigma^i}(0) = 0 \quad \text{for all } i = 1, \cdots, 2m+1,$$

$$\frac{d^{2m+2} \zeta}{d\sigma^{2m+2}}(0) = \nabla^{2m+2} h(\mathbf{x}^*) * \mathbf{s}.$$

If there exists any $\mathbf{s} \in \mathbb{R}^n$, $\mathbf{s} \neq \mathbf{0}$ with $\nabla^{2m+2} h(\mathbf{x}^*) * \mathbf{s} < 0$, then h decreases starting from $\mathbf{x}^*$ along the direction of $\mathbf{s}$. This contradicts the assumption of Theorem 3.2. □

THEOREM 3.3 (sufficient condition). *Let* $h \in C^k$, $k > 2m+1$, $m \in \mathbb{N}$, *and* $\mathbf{x}^*$ *be a point such that*

$$\nabla^i h(\mathbf{x}^*) = \mathbf{0} \quad \textit{for all } i = 1, \cdots, 2m+1$$

and

$$\nabla^{2m+2} h(\mathbf{x}^*) * \mathbf{x} > \mathbf{0} \quad \textit{for all } \mathbf{x} \in \mathbb{R}^n, \qquad \mathbf{x} \neq \mathbf{0};$$

then $\mathbf{x}^*$ *is an isolated minimizer of* h.

Proof. From the assumptions we obtain the existence of some $\alpha > 0$ with

$$\nabla^{2m+2} h(\mathbf{x}^*) * \mathbf{s} \geqq \alpha \quad \text{for all } \mathbf{s} \in \mathbb{R}^n, \qquad \|\mathbf{s}\|_2^2 = 1.$$

Since $\nabla^{2m+2}\mathrm{h}(\mathbf{x})$ is continuous, there exists some $\varepsilon > 0$ and an open neighbourhood $U_\varepsilon(\mathbf{x}^*) := \{\mathbf{x} \in \mathbb{R}^n;\ \|\mathbf{x} - \mathbf{x}^*\|_2^2 < \varepsilon\}$, with

(3.3) $$\nabla^{2m+2} h(\mathbf{x}) * \mathbf{s} > 0 \quad \text{for all } \mathbf{x} \in U_\varepsilon(\mathbf{x}^*), \qquad \mathbf{s} \in \mathbb{R}^n, \qquad \|\mathbf{s}\|_2^2 = 1.$$

Using Taylor expansion of $\zeta\colon \mathbb{R} \to \mathbb{R}$, $\zeta(\sigma) = h(\mathbf{x}^* - \sigma \mathbf{s})$, $\mathbf{s} \in \mathbb{R}^n$, $\|\mathbf{s}\|_2^2 = 1$, we obtain:

$$h(\mathbf{x}^* - \sigma \mathbf{s}) = h(\mathbf{x}^*) + \frac{1}{(2m+2)!} \nabla^{2m+2} h(\bar{\mathbf{x}}) * \mathbf{s} \cdot \sigma^{2m+2},$$

where $\bar{\mathbf{x}} = \mathbf{x}^* - t \cdot \sigma \cdot \mathbf{s}$, $0 \leqq t \leqq 1$. For $0 < \sigma < \varepsilon$, it follows from (3.3) that $h(\mathbf{x}^* - \sigma \mathbf{s}) > h(\mathbf{x}^*)$ for all $s \in \mathbb{R}^n$, $\|\mathbf{s}\|_2^2 = 1$. □

4. Examples. First we investigate the function $f\colon \mathbb{R}^2 \to \mathbb{R}$, $f(\mathbf{x}) = \frac{1}{2}(x_1 - x_2^2)(x_1 - 2x_2^2)$. The point $(0, 0)$ is a critical stationary point of f with rank 1. An interesting property of f can be observed: f is increasing along each direction starting from $(0, 0)$, but $(0, 0)$ is a saddlepoint of f. We obtain:

$$\mathbf{x}^r = x_1,$$

$$\mathbf{x}^{n-r} = x_2,$$

$$(\mathbf{x}^{*r}, \mathbf{x}^{*n-r}) = (0, 0),$$

$$g(x_2) = \frac{3}{2} x_2^2,$$

$$\bar{f}(x_2) = -\frac{1}{8} x_2^4.$$

Thus, the function f decreases along the curve $(3/2x_2^2, x_2)$, $x_2 \in \mathbb{R}$. A point $(\bar{x}_1, \bar{x}_2)$ with $f(\bar{x}_1, \bar{x}_2) < 0$ can be computed numerically by computation of $\bar{x}_2$ such that $\bar{f}(\bar{x}_2) < 0$. In this case the fixed-point theorem of Banach, for the computation of the function values of g and $\bar{f}$ (see (2.2)), and implicit automatic differentiation, for the computation of a descent direction for $\bar{f}$ in $x_2 = 0$, are necessary. The relationship between the implicit function theorem and the fixed-point theorem of Banach for numerical analysis is given in [2].

Now we consider the function $f: \mathbb{R}^2 \to \mathbb{R}$,

$$f(\mathbf{x}) = \tfrac{1}{2}(\tfrac{1}{4}(x_1^2 + x_2^2) + x_1)^2.$$

The point $(0, 0)$ is a critical stationary point of f with rank 1. We obtain:

$$\begin{aligned} \mathbf{x}^r &= x_1, \\ \mathbf{x}^{n-r} &= x_2, \\ (\mathbf{x}^{*r}, \mathbf{x}^{*n-r}) &= (0, 0), \\ g(x_2) &= -2 + (4 - x_2^2)^{1/2}, \quad \text{for all } x_2 \in \,]-2, 2[, \\ \bar{f}(x_2) &\equiv 0. \end{aligned}$$

Thus, the function f has not isolated minimizers along the curve

$$(-2 + (4 - x_2^2)^{1/2}, x_2), \qquad x_2 \in \,]-2, 2[.$$

This can be investigated numerically only by evaluating the function $\bar{f}$ in a neighbourhood of zero.

REFERENCES

[1] D. CASTRIGIANO AND S. HAYES-WIDMANN, *Catastrophe Theory*, Addison-Wesley, Reading, MA, 1991.

[2] C. H. EDWARDS, *Advanced Calculus of Several Variables*, Academic Press, New York, 1973.

[3] H. FISCHER, *Some Aspects of Automatic Differentiation*, in Numerical Methods and Approximation Theory III, G. V. Milovanovic, ed., University of Nis Press, Nis, Yugoslavia, 1988, pp. 199-208.

[4] ———, *Automatic differentiation of characterizing sequences*, J. Comput. Appl. Math., 28 (1989), pp. 181-185.

[5] ———, *Automatic differentiation of implicitly defined functions*, Tech. Report Nr. 112A, Institut für Angewandte Mathematik und Statistik, Technische Universität München, München, Germany, 1989.

[6] R. FLETCHER, *Practical Methods of Optimization*, Vol. 1, John Wiley & Sons, New York, 1980.

SIAM J. OPTIMIZATION
Vol. 2, No. 1, pp. 7-20, February 1992

002

A POTENTIAL REDUCTION ALGORITHM ALLOWING COLUMN GENERATION*

YINYU YE†

Abstract. Using the Dantzig-Wolfe decomposition technique, a potential reduction algorithm allowing column generation for the linear feasibility (LF) problem is developed. The point of departure is a simple containing polytope and its analytic center. In each iteration, an inequality violated at the current center is selected, used to cut the polytope, and then used to find the new center for the shrunken polytope. The potential value associated with the containing polytope is reduced by a constant at each step, and the algorithm is terminated in a polynomial time that depends only on the number and data length of the inequalities generated in the iterative process.

Key words. linear inequality, linear programming, analytic center, potential function, column generation

AMS(MOS) subject classification. 90C05

Introduction. While various interior point algorithms (for example, Karmarkar [7] and Renegar [11]) have been developed for linear programming, they have a common drawback: they need the complete knowledge of the linear system, and their computational complexity depends on all of the constraints in the full system, although some of them are merely useless for defining the solution.

On the other hand, the decomposition algorithm (Dantzig and Wolfe [1]) and the ellipsoid method (Khachiyan [8]) do not need knowledge of the full system in advance—they allow column generation: if a constraint (or column) is needed during their course, it is then called and added to the process. This column generation technique permits a great deal of flexibility for solving linear programs, such as semi-infinite programming and combinatorial optimization problems, in which the number of inequalities is very large or some constraints are not explicitly known.

Based on the theorem developed in [18] for the analytic center and its associated potential function (Huard and Liêu [6] and Sonnevend [14]), we develop a decomposition and potential reduction algorithm allowing column generation for the linear feasibility (LF) problem that is equivalent to linear programming (LP). We start with a simple containing polytope and its known center. In each iteration, we select an inequality violated at the center, use it to cut the polytope, and then find the new (approximate) center for the shrunken but containing polytope. The potential value associated with the polytope is reduced by a constant at each iteration, and the algorithm is terminated in a polynomial time that depends only on the number and data length of the inequalities generated during the iterative process. The algorithm either finds a feasible point or detects the infeasibility of the problem. It does not need knowledge of the full linear system.

In the context of LP, this approach sounds much like a generalized simplex-like method. We select a column with negative reduced cost to enter the basis, but we never delete a column from the basis (one may do so in practice). The iterative progress is measured by a potential function that somehow represents the volume of the containing polytope, and it is monotonically reduced by a constant as the containing polytope shrinks.

* Received by the editors November 6, 1989; accepted for publication (in revised form) February 11, 1991. This research was supported in part by National Science Foundation grant DDM-8922636 and the College Summer Grant of College of Business Administration, University of Iowa, Iowa City, Iowa.

† Department of Management Sciences, University of Iowa, Iowa City, Iowa 52242.

In fact, several authors (Goffin and Vial [2] and Mitchell [9]) have proposed using Karmarkar's projective algorithm as the LP solver in the column generation method. They have reported encouraging computational results. In their approaches, no special efforts were devoted to adding a cut through the analytic center. Sonnevend [14] suggested doing so but gave no algorithmic analysis. Recently, we have learned that Goffin, Haurie, and Vial [3] also proposed a decomposition algorithm by adding cuts through the centers. They reported good performance of their algorithm. None of the above authors give complexity results for their approaches. Obviously, the number of sub-LP problems solved here is q, the number of cuts added to the system. Therefore, the complexity of the general decomposition algorithm is $O(q\chi)$, where χ is the complexity of a linear program with $O(q)$ inequalities. Since the best-known complexity χ is $O(\sqrt{q}L)$ iterations, their approaches yield $O(q\sqrt{q}L)$ iterations for solving a linear program.

In this paper, we show that our decomposition potential reduction algorithm needs only $O(qL)$ iterations. This reduces the above complexity by a factor $\sqrt{q}$. (This result was pointed out by one of the referees.) Although q in the decomposition algorithm (regardless of which LP solver is being used) is generally small for solving most practical problems, it can very well be exponential in number. Thus we emphasize that the complexity of our approach remains *worse* than the ellipsoid method for some LP problems with large numbers of inequalities. At the end of this paper, we will discuss some directions which might resolve this problem. In fact, Vaidya [17] recently developed another decomposition method using a "volumetric center" and obtained a better theoretical result than ours.

1. A combinatorial property of analytic centers. Denote by y^a the analytic center of a polytope

$$\Omega = \{y \in R^m \colon s = c - A^T y \geqq 0\},$$

where $c \in R^n$ and $A \in R^{m \times n}$, and define the potential value of Ω as the potential function value at y^a, which is the logarithmic product of the inequality slacks (distances) $s^a = c - A^T y^a$, i.e.,

$$P(\Omega) = \sum_{j=1}^{n} \ln (s_j^a) = \sum_{j=1}^{n} \ln (c_j - a_j^T y^a), \tag{1}$$

where a_j is the jth column of A. In fact, y^a is the point that maximizes the potential function among all $y \in \Omega$, and $P(\Omega)$ is related to the volume of the largest ellipsoid inscribing Ω in terms of the slacks of the inequality system (see related topics in Todd [16] and Ye [20]).

Now one of the inequalities, say the last one, of $c - A^T y \geqq 0$ needs to be updated: change $c_n - a_n^T y \geqq 0$ to $a_n^T y^a - a_n^T y \geqq 0$, i.e., a hyperplane parallel to the last inequality that cuts through the center y^a and divides Ω into two bodies. Let

$$\Omega^+ = \{y \in R^m \colon c_j - a_j^T y \geqq 0, \quad j = 1, 2, \cdots, n-1, a_n^T y^a - a_n^T y \geqq 0\}$$

and let $\bar{y}^a$ be the analytic center of Ω^+. Then the potential value for the new polytope Ω^+ is

$$P(\Omega^+) = \sum_{j=1}^{n-1} \ln (c_j - a_j^T \bar{y}^a) + \ln (a_n^T y^a - a_n^T \bar{y}^a).$$

In [18], we showed that

$$P(\Omega^+) \leqq P(\Omega) - 1. \tag{2}$$

Inequality (2) resembles properties that hold for the center of gravity (Grünbaum [5] and Mityagin [10]) and the center of the maximum-volume inscribing ellipsoid (Tarasov, Khachiyan, and Érlikh [15]). In fact, a more general inequality can be developed. Let the hyperplane cut the polytope not necessarily through the center y^a, i.e., let

$$\Omega^+_\beta = \{y \in R^m \colon c_j - a_j^T y \geqq 0, j = 1, 2, \cdots, n-1, \beta s_n^a + a_n^T y^a - a_n^T y \geqq 0\},$$

where $\beta \geqq 0$. Denote by $\bar{y}^a$ the analytic center of Ω^+_β. Then we have Theorem 1.

THEOREM 1.

$$P(\Omega^+_\beta) \leqq P(\Omega) - (1-\beta). \tag{3}$$

Proof. The proof is similar to the one in [18]. Since y^a is the analytic center of Ω, there exists $x^a > 0$ such that

$$X^a s^a = X^a(c - A^T y^a) = e \quad \text{and} \quad Ax^a = 0, \tag{4}$$

where e is the vector (with varying dimensions) of all ones, and X designates the diagonal matrix of x. Since $\bar{c}_j = c_j$ for $j = 1, \cdots, n-1$, $\bar{c}_n = \beta s_n^a + a_n^T y^a$, and $\bar{s}^a = \bar{c} - A^T \bar{y}^a$, we have

$$\begin{aligned} e^T X^a \bar{s}^a &= e^T X^a(\bar{c} - A^T \bar{y}^a) = e^T X^a \bar{c} \\ &= e^T X^a c - (1-\beta) x_n^a s_n^a = n - 1 + \beta. \end{aligned}$$

Thus,

$$\begin{aligned} \frac{\exp P(\Omega^+_\beta)}{\exp P(\Omega)} &= \prod_{j=1}^n \frac{\bar{s}_j^a}{s_j^a} = \prod_{j=1}^n \bar{s}_j^a x_j^a \\ &\leqq \left(\frac{1}{n} \sum_{j=1}^n \bar{s}_j^a x_j^a\right)^n = \left(\frac{n-1+\beta}{n}\right)^n \leqq \exp(\beta - 1). \end{aligned}$$ □

When an additional hyperplane (say the $(n+1)$th) is added to the system, the new convex body is defined by

$$\Omega^+_\beta = \{y \in R^m \colon c_j - a_j^T y \geqq 0, j = 1, 2, \cdots, n, \beta \bar{r} + a_{n+1}^T y^a - a_{n+1}^T y \geqq 0\},$$

where $\beta \geqq 0$ and

$$\bar{r} = \sqrt{a_{n+1}^T (A(X^a)^2 A^T)^{-1} a_{n+1}}. \tag{5}$$

Then we have the second inequality, in Theorem 2.

THEOREM 2.

$$P(\Omega^+_\beta) \leqq P(\Omega) + \ln(4\bar{r}) - (1.5 - \beta).$$

Proof. Again, x^a and y^a satisfy (4). Note $c_{n+1} = \beta \bar{r} + a_{n+1}^T y^a$. Let $\bar{s}^a = c - A^T \bar{y}^a > 0$ be the first n slacks at the new center $\bar{y}^a$. Then we have

$$\begin{aligned} \bar{s}_{n+1}^a &= a_{n+1}^T (y^a - \bar{y}^a) + \beta \bar{r} \\ &= a_{n+1}^T (A(X^a)^2 A^T)^{-1} (A(X^a)^2 A^T)(y^a - \bar{y}^a) + \beta \bar{r} \\ &= a_{n+1}^T (A(X^a)^2 A^T)^{-1} A(X^a)^2 (A^T y^a - A^T \bar{y}^a) + \beta \bar{r} \\ &= a_{n+1}^T (A(X^a)^2 A^T)^{-1} A(X^a)^2 (-c + A^T y^a + c - A^T \bar{y}^a) + \beta \bar{r} \\ &= a_{n+1}^T (A(X^a)^2 A^T)^{-1} A(X^a)^2 (\bar{s}^a - s^a) + \beta \bar{r} \\ &= a_{n+1}^T (A(X^a)^2 A^T)^{-1} AX^a (X^a \bar{s}^a - e) + \beta \bar{r} \\ &\leqq \| a_{n+1}^T (A(X^a)^2 A^T)^{-1} AX^a \| \, \| X^a \bar{s}^a - e \| + \beta \bar{r} \\ &= (\| X^a \bar{s}^a - e \| + \beta) \bar{r}. \end{aligned} \tag{6}$$

We also have

$$e^T X^a \bar{s}^a = e^T X^a (c - A^T \bar{y}^a) = e^T X^a c = n. \tag{7}$$

Thus, from (6),

$$\frac{\exp P(\Omega_\beta^+)}{\bar{r} \exp P(\Omega)} = \frac{\bar{s}_{n+1}^a}{\bar{r}} \prod_{j=1}^n \frac{\bar{s}_j^a}{s_j^a} = \frac{\bar{s}_{n+1}^a}{\bar{r}} \prod_{j=1}^n \bar{s}_j^a x_j^a$$

$$\leqq (\|X^a \bar{s}^a - e\| + \beta) \prod_{j=1}^n \bar{s}_j^a x_j^a. \tag{8}$$

Let $\alpha = X^a \bar{s}^a > 0$. Then, to evaluate the right side of (8) together with (7), we face the problem

$$\begin{aligned} &\text{maximize} \quad \psi(\alpha) = (\|\alpha - e\| + \beta) \prod_{j=1}^n \alpha_j \\ &\text{subject to} \quad e^T \alpha = n \text{ and } \alpha > 0. \end{aligned}$$

Let $\|\alpha - e\|$ be fixed at some μ. Then we have a related problem:

$$\begin{aligned} &\text{maximize} \quad (\mu + \beta) \prod_{j=1}^n \alpha_j \\ &\text{subject to} \quad e^T \alpha = n, \qquad \|\alpha - e\|^2 = \mu^2 \text{ and } \alpha > 0. \end{aligned}$$

This maximum is achieved, without loss of generality, at $\alpha_1 = \delta > 1$ and

$$\alpha_2 = \cdots = \alpha_n = \frac{n - \delta}{n - 1} > 0.$$

This can be directly derived from the lemma of Schrijver ([13, p. 192]), where he considers the minimum of the problem. Hence,

$$\begin{aligned} \psi(\alpha) &\leqq \left((\delta - 1)\sqrt{\frac{n}{n-1}} + \beta \right) \delta \left(\frac{n - \delta}{n - 1} \right)^{n-1} \\ &= 4\sqrt{\frac{n}{n-1}} \, \frac{\delta - 1 + \sqrt{(n-1)/n}\,\beta}{2} \, \frac{\delta}{2} \left(\frac{n - \delta}{n - 1} \right)^{n-1} \\ &\leqq 4\sqrt{\frac{n}{n-1}} \left(\frac{n - .5 + \sqrt{(n-1)/n}\,\beta}{n+1} \right)^{n+1} \\ &\leqq \frac{4}{\exp(1.5 - \sqrt{(n-1)/n}\,\beta)} < \frac{4}{\exp(1.5 - \beta)}. \end{aligned}$$

From (8),

$$\frac{\exp P(\Omega_\beta^+)}{\exp P(\Omega)} \leqq \frac{4\bar{r}}{\exp(1.5 - \beta)}. \qquad \square$$

2. The potential algorithm allowing column generation. Now consider finding a feasible point in the region

$$\Omega = \{ y \in R^m \colon c - A^T y \geqq 0 \text{ and } 0 \leqq y \leqq e \},$$

where the components of A and c are rationals.

We assume that $\|a_j\|=1$, i.e., $c_j - a_j^T y$ represents the real distance from y to the hyperplane $\{y: c_j - a_j^T y = 0\}$. We also assume that Ω has a nonempty interior. More specifically, for any subsystem,

$$\bar{\Omega} = \{y \in R^m: \bar{c} - \bar{A}^T y \geqq 0 \text{ and } 0 \leqq y \leqq e\},$$

where $\bar{c}$ (or $\bar{A}$) is a subvector (or a submatrix) of c (or A) and there exists a point such that

$$\bar{c} - \bar{A}^T y \geqq 2^{-L} e \quad \text{and} \quad 2^{-L} e \leqq y \leqq (1-2^{-L})e$$

for some fixed $L > 1$. Thus,

$$P(\bar{\Omega}) \geqq -qL, \tag{9}$$

where q is the number of inequalities in $\bar{\Omega}$.

The latter assumption is not critical since any nonempty integral linear inequality system can be equivalently represented by a linear strict-inequality system with a nonempty interior. The inequalities $0 \leqq y \leqq e$ are also without loss of generality since a feasible point, if it exists, must be bounded. Thus we can explicitly add lower and upper bounds for the variables, then transform and scale them to 0 and 1, respectively.

We now describe an algorithm using perfect center pairs.

ALGORITHM USING PERFECT CENTERS

Initialization.

Let

$$A^0 = (I, -I) \in R^{m \times 2m}, \qquad c^0 = \begin{pmatrix} e \\ 0 \end{pmatrix} \in R^{2m},$$

and

$$y^0 = 0.5e \in R^m, \qquad s^0 = c^0 - (A^0)^T y^0 = 0.5e \in R^{2m} \quad \text{and} \quad x^0 = 2e \in R^{2m}.$$

Obviously,

$$A^0 x^0 = 0 \quad \text{and} \quad \|X^0 s^0 - e\| = 0;$$

in other words, y^0 and x^0 are the center pair of

$$\Omega^k = \{y \in R^m: c^k - (A^k)^T y \geqq 0\} \quad \text{for } k = 0.$$

The kth Iteration.

At the kth iteration, we check to see if the current center y^k satisfies all inequalities of Ω. If y^k does, then we terminate the algorithm with a feasible point; otherwise, there exists one $\bar{j}$ such that $c_{\bar{j}} - a_{\bar{j}}^T y^k < 0$. Two cases may occur:

1. $a_{\bar{j}}$ is already included in A^k, or
2. $a_{\bar{j}}$ is not.

In case 1, we update

$$c_{\bar{j}}^{k+1} = \beta(c_{\bar{j}}^k - a_{\bar{j}}^T y^k) + a_{\bar{j}}^T y^k = \beta s_{\bar{j}}^k + a_{\bar{j}}^T y^k,$$
$$c_j^{k+1} = c_j^k \quad \text{for } j \neq \bar{j},$$

and

$$A^{k+1} = A^k.$$

Let the center of Ω^{k+1} be y^{k+1}. Then, from Theorem 1, we have

$$P(\Omega^{k+1}) \leqq P(\Omega^k) - (1-\beta). \tag{10}$$

In case 2, we update

$$c^{k+1}=\begin{pmatrix} c^k \\ \beta\bar{r}+a_{\bar{j}}^Ty^k \end{pmatrix} \quad \text{and} \quad A^{k+1}=(A^k, a_{\bar{j}}),$$

where $\bar{r}=\sqrt{a_{\bar{j}}^T(A^k(X^k)^2(A^k)^T)^{-1}a_{\bar{j}}}$. Since $0<y^k<e$,

$$A^k(X^k)^2(A^k)^T=A^k(S^k)^{-2}(A^k)^T \geqq (Y^k)^{-2}+(I-Y^k)^{-2} \geqq 8I,$$

i.e., $A^k(S^k)^{-2}(A^k)^T-8I$ is positive semidefinite. Hence,

$$\bar{r}^2=a_{\bar{j}}^T(A^k(X^k)^2(A^k)^T)^{-1}a_{\bar{j}} \leqq \frac{\|a_{\bar{j}}\|^2}{8}=\frac{1}{8}.$$

Let the center of Ω^{k+1} be y^{k+1}. Then, from Theorem 2, we have

$$P(\Omega^{k+1}) \leqq P(\Omega^k)+\ln(\sqrt{2})-(1.5-\beta). \tag{11}$$

Center Updating.

Compute the center pair y^{k+1} and x^{k+1} of Ω^{k+1}, using Newton's method from y^k and $\bar{x}$ as described below and go to the $(k+1)$th iteration.

In both cases of the algorithm the potential function is reduced by a constant (for example, 0.15) for an appropriate fixed $\beta(\beta=\frac{1}{2}$ in case 1 and $\beta=1$ in case 2). Next, we show that the new center pair y^{k+1} and x^{k+1} of Ω^{k+1} can be "easily" computed. In fact, we show that y^k is still in the "quadratic convergence" region of Ω^{k+1}, i.e., y^{k+1} can be updated from y^k using Newton's method with a quadratic convergence order.

LEMMA 1. *In both cases of the algorithm there exists an $\bar{x}>0$ such that*

$$A^{k+1}\bar{x}=0 \quad \textit{and} \quad \|\bar{X}\bar{s}-e\| \leqq \delta<1,$$

where $\bar{s}=c^{k+1}-(A^{k+1})^Ty^k \in R^n$.

Proof. In case 1, we have $\bar{s}_j=s_j^k$ for $j\neq\bar{j}$ and $\bar{s}_{\bar{j}}=\beta s_{\bar{j}}^k$. Let $\bar{x}=x^k>0$. Then,

$$A^{k+1}\bar{x}=A^kx^k=0,$$

and for $0<\beta<1$,

$$\|\bar{X}\bar{s}-e\|=(1-\beta)<1.$$

In case 2, we have

$$\bar{s}=\begin{pmatrix} s^k \\ \beta\bar{r} \end{pmatrix}.$$

Let $\alpha=\beta/(1+\beta^2)<1$ and

$$\Delta x=-\frac{\alpha}{\bar{r}}X^k(A^k)^T(A^k(X^k)^2(A^k)^T)^{-1}a_{\bar{j}}.$$

Then

$$\|\Delta x\|=\frac{\alpha}{\bar{r}}\bar{r}=\alpha<1.$$

Now let

$$\bar{x}=\begin{pmatrix} X^k(e+\Delta x) \\ \alpha/\bar{r} \end{pmatrix}>0.$$

Then

$$A^{k+1}\bar{x} = A^k X^k \Delta x + \frac{\alpha}{\bar{r}} a_{\bar{j}} = 0$$

and

$$\|\bar{X}\bar{s} - e\|^2 = \|\Delta x\|^2 + (1-\alpha\beta)^2 = \alpha^2 + (1-\alpha\beta)^2 = \frac{1}{(1+\beta^2)}. \qquad \square$$

It has been shown by Roos and Vial [12] that if the starting $\bar{x}$ and $\bar{s}$ are in the "quadratic convergence" region described in Lemma 1, the new center pair y^{k+1} and x^{k+1} of Ω^{k+1} can be generated in $O(\ln L)$ Newton's steps. In particular, we can repeatedly solve the following system of linear equations for Δx and Δy:

$$\bar{X}\Delta s + \bar{S}\Delta x = e - \bar{X}\bar{s},$$

$$A^{k+1}\,\Delta x = 0 \quad \text{and} \quad \Delta s = c^{k+1} - (A^{k+1})^T\,\Delta y,$$

and let

$$\bar{x} = \bar{x} + \Delta x \quad \text{and} \quad \bar{s} = \bar{s} + \Delta s.$$

Note that a new solution generated from the system always has

$$(\bar{s})^T\bar{x} = n. \tag{12}$$

We now derive the following convergence theorem.

THEOREM 3. *In $O(qL)$ iterations and $O(qL \ln L)$ Newton's steps, the perfect center algorithm generates a feasible point in Ω, where q is the number of inequalities in the final system of $\{\Omega^k\}$.*

Proof. Note that $P(\Omega^0) \leqq 0$. In each iteration, we either update an inequality or add an inequality. The potential function is reduced by a constant due to (10) and (11). Let $\bar{\Omega}$ be the subsystem of Ω corresponding to the final Ω^k, i.e.,

$$\bar{\Omega} = \{y\colon c(c^k) - (A^k)^T y \geqq 0\},$$

where $c(c^k)$ is the subvector of c having the same indices as c^k. Then, we must have

$$c(c^k) \leqq c^k \quad \text{and} \quad \bar{\Omega} \subset \Omega^k,$$

which leads from (9) to

$$-qL \leqq P(\bar{\Omega}) \leqq P(\Omega^k). \tag{13}$$

However, after $O(qL)$ iterations,

$$P(\Omega^k) \leqq -qL,$$

which contradicts (13). □

3. An algorithm using approximate centers. In this section, we show that using the perfect center is unnecessary. This issue has been discussed by Renegar [11] and Sonnevend [14] in a path-following algorithm for linear programming. Similarly, we now use approximate centers instead of perfect centers in our potential algorithm. In fact, the step in case 1 of our algorithm is similar to the step in Renegar's algorithm in which only the objective hyperplane is updated iteratively.

Without involving too many error analyses, we prove the following basic theorem to show that an approximate center suffices to terminate the algorithm.

LEMMA 2. *Let* $\Omega = \{y \in R^m: s = c - A^T y \geqq 0 \in R^n\}$, *and let an interior pair* x^k *and* y^k $(s^k = c - A^T y^k)$ *of* Ω *satisfy*

$$Ax^k = 0, \qquad (s^k)^T x^k = n, \quad \textit{and} \quad \|X^k s^k - e\| \leqq \gamma, \tag{14}$$

where $\gamma < 1$. *Then,*

$$P(\Omega) \geqq \sum_{j=1}^n \ln(s_j^k) \geqq P(\Omega) - \frac{\gamma^2}{2(1-\gamma)}.$$

Proof. From Lemma 1 of Ye [19],

$$\sum_{j=1}^n \ln(x_j^k s_j^k) \geqq e^T X^k s^k - n - \frac{\gamma^2}{2(1-\gamma)} = -\frac{\gamma^2}{2(1-\gamma)}.$$

Denote by x^a and y^a $(s^a = c - A^T y^a)$ the center pair of Ω. Noting that $X^a s^a = e$, we have

$$\sum_{j=1}^n \ln(x_j^a s_j^a) - \sum_{j=1}^n \ln(x_j^k s_j^k) \leqq \frac{\gamma^2}{2(1-\gamma)}. \tag{15}$$

The left-hand side of (15) can be written as

$$\sum_{j=1}^n \ln(x_j^a) - \sum_{j=1}^n \ln(x_j^k) + \sum_{j=1}^n \ln(s_j^a) - \sum_{j=1}^n \ln(s_j^k).$$

Since y^a maximizes the potential function over the interior of Ω, we have

$$\sum_{j=1}^n \ln(s_j^a) - \sum_{j=1}^n \ln(s_j^k) \geqq 0.$$

On the other hand, one can verify that x^a is the maximizer of

$$\begin{aligned} &\text{maximize} && \sum_{j=1}^n \ln(x_j) \\ &\text{subject to} && x \in \{x: Ax = 0, c^T x = n, x > 0\}. \end{aligned}$$

Thus,

$$\sum_{j=1}^n \ln(x_j^a) - \sum_{j=1}^n \ln(x_j^k) \geqq 0.$$

Due to (15), we have the desired result. □

Lemma 2 indicates that the potential value at an approximate center of Ω, characterized by the condition (14), differs from the exact potential value by a small constant. Therefore, for an approximate center y^k at the kth iteration of the algorithm, we replace (13) with

$$\sum \ln(s_j^k) \geqq P(\Omega^k) - \frac{\gamma^2}{2(1-\gamma)} \geqq -qL - \frac{\gamma^2}{2(1-\gamma)},$$

which can be used to terminate the algorithm after $k = O(qL)$ iterations.

ALGORITHM USING APPROXIMATE CENTERS

The kth iteration of the algorithm can be modified as follows. In case 1, we update

$$c_{\bar{j}}^{k+1} = s_{\bar{j}}^k - (1-\beta)/x_{\bar{j}}^k + a_{\bar{j}}^T y^k,$$
$$c_j^{k+1} = c_j^k \quad \text{for } j \neq \bar{j},$$

and

$$A^{k+1} = A^k.$$

In case 2, we update

$$c^{k+1} = \begin{pmatrix} c^k \\ \beta\bar r + a_{\bar j}^T y^k \end{pmatrix} \quad \text{and} \quad A^{k+1} = (A^k, a_{\bar j}),$$

where $\bar r = \sqrt{a_{\bar j}^T(A^k(X^k)^2(A^k)^T)^{-1}a_{\bar j}}$. Since $0 < y^k < e$,

$$\begin{aligned} A^k(X^k)^2(A^k)^T &= A^k(S^k)^{-1}(S^kX^k)^2(S^k)^{-1}(A^k)^T \\ &\geqq (\min x_j^k s_j^k)^2((Y^k)^{-2} + (I - Y^k)^{-2}) \\ &= (1-\gamma)^2((Y^k)^{-2} + (I - Y^k)^{-2}) \geqq 8(1-\gamma)^2 I, \end{aligned}$$

i.e., $A^k(S^k)^{-2}(A^k)^T - 8(1-\gamma)^2 I$ is positive semidefinite. Hence,

$$\bar r^2 = a_{\bar j}^T(A(X^k)^2A^T)^{-1}a_{\bar j} \leqq \frac{\|a_{\bar j}\|^2}{8(1-\gamma)^2} = \frac{1}{8(1-\gamma)^2}. \tag{16}$$

We now show that Lemma 1 is still valid.

LEMMA 3. *Let x^k and s^k be given in* (14). *Then, in both cases of the modified algorithm, there exists an $\bar x > 0$ such that*

$$A^{k+1}\bar x = 0 \quad \textit{and} \quad \|\bar X\bar s - 1\| \leqq \lambda \quad \textit{for some constant } \lambda < 1,$$

where $\bar s = c^{k+1} - (A^{k+1})^T y^k \in R^n$.

Proof. In case 1, we have $\bar s_j = s_j^k$ for $j \neq \bar j$ and $\bar s_{\bar j} = s_{\bar j}^k - (1-\beta)/x_{\bar j}^k$. Let $\bar x = x^k > 0$. Then,

$$A^{k+1}\bar x = A^k x^k = 0$$

and

$$\begin{aligned} \|\bar X\bar s - e\| &= \|X^k s^k - e + \bar X\bar s - X^k s^k\| \\ &\leqq \|X^k s^k - e\| + \|\bar X\bar s - X^k s^k\| \\ &\leqq \gamma + (1-\beta). \end{aligned}$$

In case 2, we have

$$\bar s = \begin{pmatrix} s^k \\ \beta\bar r \end{pmatrix}.$$

Let $\alpha < 1$ and

$$\Delta x = -\frac{\alpha}{\bar r} X^k(A^k)^T(A^k(X^k)^2(A^k)^T)^{-1}a_{\bar j}.$$

Then, again,

$$\|\Delta x\| = \frac{\alpha}{\bar r}\bar r = \alpha < 1.$$

Now let

$$\bar x = \begin{pmatrix} X^k(e + \Delta x) \\ \alpha/\bar r \end{pmatrix} > 0.$$

Then,

$$A^{k+1}\bar{x} = A^k X^k \Delta x + \frac{\alpha}{\bar{r}} a_{\bar{j}} = 0,$$

and

$$\begin{aligned}\|\bar{X}\bar{s} - e\|^2 &= \|X^k S^k(1+\Delta x) - e\|^2 + (1-\alpha\beta)^2\\ &\leqq (\|X^k S^k - e\| + \|X^k S^k \Delta x\|)^2 + (1-\alpha\beta)^2\\ &\leqq (\|X^k S^k - e\| + \|X^k S^k\|\,\|\Delta x\|)^2 + (1-\alpha\beta)^2\\ &\leqq (\gamma + (1+\gamma)\alpha)^2 + (1-\alpha\beta)^2\\ &\leqq 1 + \gamma^2 - \frac{(\beta - \gamma(1+\gamma))^2}{(1+\gamma)^2 + \beta^2}\end{aligned}$$

for

$$\alpha = \frac{\beta - \gamma(1+\gamma)}{(1+\gamma)^2 + \beta^2}.$$

Letting $\beta = \frac{1}{2}$ in case 1 and $\beta = 1$ in case 2, and letting γ be small enough but a constant, we then have the desired result. □

Lemma 3 shows that, even though it is not perfectly centered, y^k is still in the "quadratic convergence" region of Ω^{k+1}. An approximate center pair y^{k+1} and x^{k+1} can be updated from y^k and $\bar{x}$ in a constant number of Newton's steps. We now verify that the potential function is still reduced by a constant for a small γ.

LEMMA 4. *Let x^k and s^k be given in* (14) *and let $\bar{x}^a$ and $\bar{s}^a$ ($\bar{y}^a$) be the center pair for Ω^{k+1} defined in the modified algorithm. Then, in both cases of the modified algorithm,*

$$P(\Omega^{k+1}) \leqq P(\Omega^k) - \delta \quad \textit{for some constant } \delta > 0.$$

Proof. The proof for case 1 is similar to Theorem 1. Note that we still have

$$e^T X^k \bar{s}^a = e^T X^k s^k - (1-\beta) = n - 1 + \beta.$$

Since $P(\Omega^k) \geqq \sum_{j=1}^n \ln(s_j^k)$,

$$\begin{aligned}\frac{\exp P(\Omega^{k+1})}{\exp P(\Omega^k)} &\leqq \prod_{j=1}^n \frac{\bar{s}_j^a}{s_j^k} = \prod_{j=1}^n (\bar{s}_j^a x_j^k) \prod_{j=1}^n (1/s_j^k x_j^k)\\ &\leqq \left(\frac{1}{n}\sum_{j=1} \bar{s}_j^a x_j^k\right)^n \prod_{j=1}^n (1/s_j^k x_j^k) = \left(\frac{n-1+\beta}{n}\right)^n \prod_{j=1}^n (1/s_j^k x_j^k)\\ &\leqq \exp(\beta - 1) \prod_{j=1}^n (1/s_j^k x_j^k).\end{aligned}$$

Thus, from (15),

$$\begin{aligned}P(\Omega^{k+1}) - P(\Omega^k) &\leqq (\beta - 1) - \sum_{j=1}^n \ln(s_j^k x_j^k)\\ &\leqq (\beta - 1) + \frac{\gamma^2}{2(1-\gamma)}.\end{aligned}$$

In case 2, let $\bar{s}^a = c^k - (A^k)^T\bar{y}^a > 0$ be the first n slacks at the new center $\bar{y}^a$. Then we have

$$\begin{aligned}
\bar{s}^a_{n+1} &= a_j^T(y^k - \bar{y}^a) + \beta\bar{r} \\
&= a_j^T(A^k(X^k)^2(A^k)^T)^{-1}(A^k(X^k)^2(A^k)^T)(y^k - \bar{y}^a) + \beta\bar{r} \\
&= a_j^T(A^k(X^k)^2(A^k)^T)^{-1}A^k(X^k)^2((A^k)^Ty^k - (A^k)^T\bar{y}^a) + \beta\bar{r} \\
&= a_j^T(A^k(X^k)^2(A^k)^T)^{-1}A^k(X^a)^2(-c^k + (A^k)^Ty^k + c^k - (A^k)^T\bar{y}^a) + \beta\bar{r} \\
&= a_j^T(A^k(X^k)^2(A^k)^T)^{-1}A^k(X^k)^2(\bar{s}^a - s^k) + \beta\bar{r} \\
&= a_j^T(A^k(X^k)^2(A^k)^T)^{-1}A^kX^k(X^k\bar{s}^a - X^ks^k) + \beta\bar{r} \\
&\leqq \|a_j^T(A^k(X^k)^2(A^k)^T)^{-1}A^kX^k\|\,\|X^k\bar{s}^a - X^ks^k\| + \beta\bar{r} \\
&= (\|X^k\bar{s}^a - X^ks^k\| + \beta)\bar{r} \\
&\leqq (\|X^k\bar{s}^a - e\| + \|X^ks^k - e\| + \beta)\bar{r} \\
&\leqq (\|X^k\bar{s}^a - e\| + \gamma + \beta)\bar{r}.
\end{aligned}$$

Thus,

$$\begin{aligned}
\frac{\exp P(\Omega^{k+1})}{\bar{r}\exp P(\Omega^k)} &\leqq \frac{\bar{s}^a_{n+1}}{\bar{r}}\prod_{j=1}^n \frac{\bar{s}^a_j}{s^k_j} \\
&= \frac{\bar{s}^a_{n+1}}{\bar{r}}\left(\prod_{j=1}^n \bar{s}^a_jx^k_j\right)\left(\prod_{j=1}^n 1/s^k_jx^k_j\right) \\
&\leqq (\|X^k\bar{s}^a - e\| + \gamma + \beta)\left(\prod_{j=1}^n \bar{s}^a_jx^k_j\right)\left(\prod_{j=1}^n 1/s^k_jx^k_j\right).
\end{aligned}$$

Note that we still have

$$e^TX^k\bar{s}^a = e^TX^kc^k = n.$$

From (15), (16), and Theorem 2,

$$\begin{aligned}
P(\Omega^{k+1}) - P(\Omega^k) &\leqq \ln(4\bar{r}) - (1.5 - \gamma - \beta) + \frac{\gamma^2}{2(1-\gamma)} \\
&\leqq \ln(\sqrt{2}) - \ln(1-\gamma) - (1.5 - \gamma - \beta) + \frac{\gamma^2}{2(1-\gamma)}.
\end{aligned}$$

Let $\beta = \frac{1}{2}$ in case 1 and $\beta = 1$ in case 2, and let γ be small enough but a constant. Then, the potential value is reduced by a constant δ. □

Now Theorem 3 can be modified as Theorem 4.

THEOREM 4. *In $O(qL)$ iterations and $O(qL)$ Newton's steps, the approximate center algorithm generates a feasible point in Ω, where q is the number of inequalities in the final system of $\{\Omega^k\}$.*

4. Further remarks. Theoretically, the complexity result of our algorithm is worse than the best result for linear programming. However, if the full linear system is known, we can also reduce the complexity of the algorithm to the best one using Corollary 1 of Ye [18]. More precisely, we use multiple cuts to shrink the polytope. For example, if we update $O(n)$ inequalities in Ω^+, Theorem 1 becomes

$$P(\Omega^+) \leqq P(\Omega) - O(n)(1-\beta).$$

Let $\beta = 1 - 1/\sqrt{n}$. Then,

$$P(\Omega^+) \leqq P(\Omega) - O(\sqrt{n}),$$

and it can be verified that y^k is still in the "quadratic convergence" region of Ω^+. This also explains why a factor of $\sqrt{n}$ is saved from Karmarkar's original potential reduction algorithm.

Like Todd [16] and Ye [20], we can analyze a primal potential function. At the kth iteration, we have a center pair s^k (y^k) and $x^k \in R^q$. Thus,

$$q \ln ((x^k)^T s^k) - \ln \left(\prod_{j=1}^{q} (x_j^k s_j^k) \right) = q \ln q.$$

Noting that $(x^k)^T s^k = (c^k)^T x^k$, we have

$$\phi(x^k) = q \ln ((c^k)^T x^k) - \ln \left(\prod_{j=1}^{q} (x_j^k) \right) = q \ln q + P(\Omega^k),$$

where $\phi(x^k)$ relates to the volume of an ellipsoid containing the feasible region. Therefore, $\phi(x^k)$ is reduced as $P(\Omega^k)$ is decreased, i.e., the containing ellipsoid shrinks like it does in the ellipsoid method.

At this moment, we only update or add inequalities to the working system of Ω^k. In fact, we can also eliminate inequalities from the working system using the criteria developed in Todd [16] and Ye [21]. By doing this, we delete some columns from the working matrix, i.e., the current basis-candidate-set, like the simplex method in the context of LP. Further research is needed to determine how deleting inequalities will affect the potential function that is used to measure the iterative progress, and new potential functions may have to be invented to characterize this process.

Additionally, the techniques developed in this paper can be applied to solve linear programming problems of the form

$$\begin{array}{lll} \text{LD} & \text{maximize} & b^T y \\ & \text{subject to} & y \in \Omega = \{y \in R^m \colon A^T y \leqq c, 0 \leqq y \leqq e\}. \end{array}$$

A dual-form of the algorithm can be described as follows. Assuming $\|b\| = 1$, we add one more inequality,

$$b^T y \geqq z,$$

to the inequality system. We start from y^0 and z^0, a lower bound for the optimal objective value z^*, where y^0 is the approximate center for a subset of the inequalities

$$\Omega^0 = \{y \in R^m \colon b^T y \geqq z^0, 0 \leqq y \leqq e\}.$$

If y^0 is feasible for LD we increase z^0; otherwise we add one violated inequality. Again, the potential value associated with the polytope

$$\Omega^k = \{y \in R^m \colon b^T y^k \geqq z^k, (A^k)^T y \leqq c^k\}$$

is reduced by a constant. In $O(qL)$ steps we can terminate the algorithm: either report that the problem is infeasible or generate a feasible solution with

$$b^T y^k - z^* < 2^{-L},$$

where q is the number of cuts added to the system. Note that in each step we generate a positive primal feasible solution which satisfies $A^k x = b$, where A^k is only a submatrix of the constraint matrix. It is our hope that many inequalities (or columns) can be ignored during the algorithm. As we mentioned before, Goffin, Haurie, and Vial [3] and Mitchell [9] have reported encouraging computational results in their primal-form approaches. Recently, Goldstein [4] also reported similar behavior of the algorithm in solving a min-max problem, where q is virtually independent of the total number of inequalities in the system.

The following is a *nonrigorous* probabilistic argument on why this behavior may be anticipated. From Theorem 2, when a hyperplane cuts through the analytic center (i.e., $\beta = 0$), the potential reduction in one iteration of the algorithm of § 2 is

$$\ln(\psi(\alpha)) = \ln(\|\alpha - e\|) + \sum_{j=1}^{n} \ln \alpha_j,$$

with $e^T\alpha = n$ and $\alpha > 0$. Let α_j be independently drawn from a probability distribution, say the uniform distribution [0, 1], and let

$$\alpha := \frac{n\alpha}{\sum_{j=1}^{n} \alpha_j}.$$

Then, the expected value of $\ln(\psi(\alpha))$ is at most $-O(n)$ (this is also confirmed by many simulation runs). If this reduction holds for each iteration, then after k iterations the total potential reduction is

$$\sum_{i=1}^{k} O(2m + i - 1) = O(2mk + k(k-1)/2).$$

Thus, the terminating condition in Theorem 3 indicates that we need

$$O(2mk + k(k-1)/2) = (2m + k)L$$

to stop the iterative process. Here, we see that k depends on L and m only.

REFERENCES

[1] G. B. Dantzig and P. Wolfe, *The decomposition algorithm for linear programming*, Econometrica, 29 (1961), pp. 767–778.

[2] J. L. Goffin and J. P. Vial, *Cutting planes and column generation techniques with the projective algorithm*, CORE Discussion Paper 8829, CORE, Louvain la Neuve, Belgium, 1988; J. Optim. Theory Appl., to appear.

[3] J. L. Goffin, A. Haurie, and J. P. Vial, *Decomposition and nondifferentiable optimization with the projective algorithm*, manuscript, Faculty of Management, McGill University, Montréal, Canada, 1989.

[4] A. A. Goldstein, private communication, 1990.

[5] B. Grünbaum, *Partitions of mass-distributions and of convex bodies by hyperplanes*, Pacific J. Math., 10 (1960), pp. 1257–1261.

[6] P. Huard and B. T. Liêu, *La méthode des centres dans un espace topologique*, Numer. Math., 8 (1966), pp. 56–67.

[7] N. Karmarkar, *A new polynomial-time algorithm for linear programming*, Combinatorica, 4 (1984), pp. 373–395.

[8] L. G. Khachiyan, *A polynomial algorithm for linear programming*, Dokl. Akad. Nauk USSR, 244 (1979), pp. 1093–1096; Soviet. Math. Dokl. 20 (1979), pp. 191–194. (English translation.)

[9] J. E. Mitchell, *Karmarkar's algorithm and combinatorial optimization problems*, Ph.D. thesis, Department of Operations Research and Industrial Engineering, Cornell University, Ithaca, NY, 1988.

[10] B. S. Mityagin, *Two inequalities for volumes of convex bodies*, Mat. Zametki, 1 (1969), pp. 99–106.

[11] J. Renegar, *A polynomial-time algorithm, based on Newton's method, for linear programming*, Math. Programming, 40 (1988), pp. 59–93.

[12] C. Roos and J.-Ph. Vial, *A polynomial methods of approximate centers for linear programming*, manuscript, Department of Mathematics and Computer Science, Delft University of Technology, Delft, the Netherlands, 1989.

[13] A. Schrijver, *Theory of Linear and Integer Programming*, John Wiley & Sons, New York, 1989.

[14] G. Sonnevend, *An analytic center for polyhedrons and new classes of global algorithms for linear (smooth convex) programming*, in Lecture Notes in Control and Information Sciences 84, Springer-Verlag, New York, 1985, pp. 866-876.

[15] S. P. Tarasov, L. G. Khachiyan, and I. I. Érlikh, *The method of inscribed ellipsoids*, Soviet Math. Dokl., 1 (1988), pp. 226-230.

[16] M. J. Todd, *Improved bounds and containing ellipsoids in Karmarkar's linear programming algorithm*, Math. Oper. Res., 13 (1988), pp. 650-659.

[17] P. M. Vaidya, *A new algorithm for minimizing convex functions over convex sets*, in Proc. 30th Annual IEEE Symposium on Foundations of Computing, 1989, pp. 332-337; Math. Programming, to appear.

[18] Y. Ye, *A combinatorial property of analytic centers of polytopes*, Working Paper, Department of Management Sciences, University of Iowa, Iowa City, IA, 1989.

[19] ———, *An $O(n^3L)$ potential reduction algorithm for linear programming*, Math. Programming, 50 (1991), pp. 239-258.

[20] ———, *Karmarkar's algorithm and the ellipsoid method*, Oper. Res. Lett., 4 (1987), pp. 177-182.

[21] ———, *A build-down scheme for linear programming*, Math. Programming, 46 (1990), pp. 61-72.

SIAM J. OPTIMIZATION
Vol. 2, No. 1, pp. 21–42, February 1992

GLOBAL CONVERGENCE PROPERTIES OF CONJUGATE GRADIENT METHODS FOR OPTIMIZATION*

JEAN CHARLES GILBERT[†] AND JORGE NOCEDAL[‡]

Abstract. This paper explores the convergence of nonlinear conjugate gradient methods without restarts, and with practical line searches. The analysis covers two classes of methods that are globally convergent on smooth, nonconvex functions. Some properties of the Fletcher–Reeves method play an important role in the first family, whereas the second family shares an important property with the Polak–Ribière method. Numerical experiments are presented.

Key words. conjugate gradient method, global convergence, unconstrained optimization, large-scale optimization

AMS(MOS) subject classifications. 65, 49

1. Introduction. The object of this paper is to study the convergence properties of several conjugate gradient methods for nonlinear optimization. We consider only the case where the methods are implemented without regular restarts, and ask under what conditions they are globally convergent for general smooth nonlinear functions. The analysis will allow us to highlight differences among various conjugate gradient methods, and will suggest new implementations.

Our problem is to minimize a function of n variables,

$$\min f(x), \tag{1.1}$$

where f is smooth and its gradient g is available. We consider iterations of the form

$$d_k = \begin{cases} -g_k & \text{for } k = 1, \\ -g_k + \beta_k d_{k-1} & \text{for } k \geq 2, \end{cases} \tag{1.2}$$

$$x_{k+1} = x_k + \alpha_k d_k, \tag{1.3}$$

where β_k is a scalar, and α_k is a steplength obtained by means of a one-dimensional search. We call this iteration a *conjugate gradient method* if β_k is such that (1.2)–(1.3) reduces to the linear conjugate gradient method in the case when f is a strictly convex quadratic and α_k is the exact one-dimensional minimizer. Some of the results of this paper, however, also apply to methods of the form (1.2)–(1.3) that do not reduce to the linear conjugate gradient method.

The best-known formulas for β_k are called the Fletcher–Reeves (FR), Polak–Ribière (PR), and Hestenes–Stiefel (HS) formulas, and are given by

$$\beta_k^{\text{FR}} = \|g_k\|^2 / \|g_{k-1}\|^2, \tag{1.4}$$

$$\beta_k^{\text{PR}} = \langle g_k, g_k - g_{k-1} \rangle / \|g_{k-1}\|^2, \tag{1.5}$$

* Received by the editors July 9, 1990; accepted for publication (in revised form) February 15, 1991.

[†] Institut National de Recherche d'Informatique et d'Automatique, Domaine de Voluceau, Rocquencourt, B.P. 105, 78153 Le Chesnay, France.

[‡] Department of Electrical Engineering and Computer Science, Northwestern University, Evanston, Illinois 60208. The work of this author was supported by United States Department of Energy grant DE-FG02-87ER25047-A001 and National Science Foundation grant CCR-8902096.

$$\beta_k^{\text{HS}} = \langle g_k, g_k - g_{k-1}\rangle / \langle d_{k-1}, g_k - g_{k-1}\rangle. \tag{1.6}$$

Here, $\langle\cdot,\cdot\rangle$ is the scalar product used to compute the gradient and $\|\cdot\|$ denotes its associated norm. The numerical performance of the Fletcher–Reeves [6] method is somewhat erratic: it is sometimes as efficient as the Polak–Ribière and Hestenes–Stiefel methods, but it is often much slower. Powell [18] gives an argument showing that, under some circumstances, the Fletcher–Reeves method with exact line searches will produce very small displacements, and will normally not recover unless a restart along the gradient direction is performed. In spite of these drawbacks, Zoutendijk [27] has shown that the method cannot fail. He proved that the Fletcher–Reeves method with exact line searches is globally convergent on general functions. Al-Baali [1] extended this result to inexact line searches.

The Hestenes–Stiefel and Polak–Ribière methods appear to perform very similarly in practice, and are to be preferred over the Fletcher–Reeves method. Nevertheless, in a remarkably laborious paper, Powell [19] was able to show that the Polak–Ribière method with exact line searches can cycle infinitely without approaching a solution point. The same result applies to the Hestenes–Stiefel method, since the two methods are identical when $\langle g_k, d_{k-1}\rangle = 0$, which holds when line searches are exact. Since the steplength of Powell's example would probably be accepted by any practical line search, it appears unlikely that a satisfactory global convergence result can be found for the Polak–Ribière and Hestenes–Stiefel methods. In contrast, Al-Baali's convergence result for the less efficient Fletcher–Reeves method is very satisfactory. This disconcerting state of affairs motivated the present study.

In this paper we will consider various choices of β_k and various line search strategies that result in globally convergent methods. In §2 we describe the approach used in our analysis, and summarize some of the previous work in the area. Section 3 establishes global convergence for the class of methods with $|\beta_k| \le \beta_k^{\text{FR}}$, and describes a modification of the Polak–Ribière formula. In §4 we consider methods that use only nonnegative values for β_k, and which are, in some sense, related to the Polak–Ribière method. In particular, we show that a suggestion of Powell [20] to set $\beta_k = \max\{\beta_k^{\text{PR}}, 0\}$ results in global convergence, even for inexact line searches. Further remarks on the convergence results are made in §5, and the results of some numerical experiments are presented in §6.

We note that this paper does not study the rate of convergence of conjugate gradient methods. For some results on this subject, see Crowder and Wolfe [5], Cohen [4], Powell [17], Baptist and Stoer [2], and Stoer [22].

2. Preliminaries.

Some important global convergence results for conjugate gradient methods have been given by Polak and Ribière [16], Zoutendijk [27], Powell [19], and Al-Baali [1]. In this section we will see that the underlying approach used for these analyses is essentially the same, and we will describe it in detail, since it is also the basis for the results presented in this paper. Before doing so, we describe our notation, state the assumptions we make about the objective function, and consider the line search strategy.

Notation and definitions. We denote the starting point by x_1, and define $s_k := x_{k+1} - x_k$ and $y_k := g_{k+1} - g_k$. We say that d_k is a *descent direction* if $\langle g_k, d_k\rangle < 0$. We will also make use of the angle θ_k between $-g_k$ and d_k:

$$\cos\theta_k := -\langle g_k, d_k\rangle / \|g_k\|\|d_k\|. \tag{2.1}$$

The Fletcher–Reeves, Polak–Ribière, and Hestenes–Stiefel methods will be abbreviated as FR, PR, and HS, respectively. For a derivation of these methods and a discussion of some of their properties, see Gill, Murray, and Wright [11] and Fletcher [7].

ASSUMPTIONS 2.1. (i) *The level set* $\mathcal{L} := \{x : f(x) \leq f(x_1)\}$ *is bounded.*

(ii) *In some neighborhood* $\mathcal{N}$ *of* $\mathcal{L}$*, the objective function* f *is continuously differentiable, and its gradient is Lipschitz continuous, i.e., there exists a constant* $L > 0$ *such that*

$$\|g(x) - g(\tilde{x})\| \leq L\|x - \tilde{x}\|, \tag{2.2}$$

for all $x, \tilde{x} \in \mathcal{N}$.

Note that these assumptions imply that there is a constant $\bar{\gamma}$ such that

$$\|g(x)\| \leq \bar{\gamma}, \quad \text{for all } x \in \mathcal{L}. \tag{2.3}$$

Let us now turn our attention to the line search. An efficient strategy, studied by Wolfe [25], consists in accepting a positive steplength α_k if it satisfies the two conditions:

$$f(x_k + \alpha_k d_k) \leq f(x_k) + \sigma_1 \alpha_k \langle g_k, d_k \rangle \tag{2.4}$$

$$\langle g(x_k + \alpha_k d_k), d_k \rangle \geq \sigma_2 \langle g_k, d_k \rangle, \tag{2.5}$$

where $0 < \sigma_1 < \sigma_2 < 1$. We will sometimes also refer to more ideal line search conditions. To this end let us define the following strategy: a positive steplength α_k is accepted if

$$f(x_k + \alpha_k d_k) \leq f(x_k + \hat{\alpha}_k d_k), \tag{2.6}$$

where $\hat{\alpha}_k$ is the *smallest* positive stationary point of the function $\xi_k(\alpha) := f(x_k + \alpha d_k)$. Assumptions 2.1 ensure that $\hat{\alpha}_k$ exists. Note that both the first local minimizer and the global minimizer of f along the search direction satisfy (2.6).

Any of these line search strategies is sufficient to establish the following very useful result.

THEOREM 2.1. *Suppose that Assumptions* 2.1 *hold, and consider any iteration of the form* (1.3)*, where* d_k *is a descent direction and* α_k *satisfies one of the following line search conditions:*

(i) *the Wolfe conditions* (2.4)–(2.5)*, or*

(ii) *the ideal line search condition* (2.6).

Then

$$\sum_{k \geq 1} \cos^2 \theta_k \, \|g_k\|^2 < \infty. \tag{2.7}$$

This result was essentially proved by Zoutendijk [27] and Wolfe [25], [26]. We shall call (2.7) the *Zoutendijk condition.*

We can now describe the basic ideas used for the convergence analysis. The first results, by Polak and Ribière [16] and Zoutendijk [27], assume exact line searches. The term *exact line search* can be ambiguous. Sometimes, it implies that a one-dimensional minimizer is found, but often it simply means that the orthogonality condition

$$\langle g_k, d_{k-1} \rangle = 0 \tag{2.8}$$

is satisfied. Throughout the paper we will indicate in detail the conditions required of the line search. Let us suppose that d_{k-1} is a descent direction and that the line search satisfies Zoutendijk's condition and condition (2.8). From (1.2) and (2.8) we have that

$$\cos\theta_k = \frac{\|g_k\|}{\|d_k\|}, \tag{2.9}$$

which shows that d_k is a descent direction. Substituting this relation in Zoutendijk's condition (2.7) we obtain

$$\sum_{k\geq 1} \frac{\|g_k\|^4}{\|d_k\|^2} < \infty. \tag{2.10}$$

If one can show that $\{\|d_k\|/\|g_k\|\}$ is bounded, which means that $\{\cos\theta_k\}$ is bounded away from zero, then (2.10) immediately gives

$$\lim_{k\to\infty} g_k = 0. \tag{2.11}$$

This is done by Polak and Ribière [16] for their method, assuming that f is *strongly convex*, i.e., $\langle g(x) - g(\tilde{x}), x - \tilde{x}\rangle \geq c\,\|x - \tilde{x}\|^2$, for some positive constant c and for all x and $\tilde{x}$ in $\mathcal{L}$.

For general functions, however, it is usually impossible to bound $\{\|d_k\|/\|g_k\|\}$ a priori, and only a weaker result than (2.11) can be obtained, namely,

$$\liminf_{k\to\infty} \|g_k\| = 0. \tag{2.12}$$

To obtain this result one proceeds by contradiction. Suppose that (2.12) does not hold, which means that the gradients remain bounded away from zero: there exists $\gamma > 0$ such that

$$\|g_k\| \geq \gamma \tag{2.13}$$

for all $k \geq 1$. Then (2.10) implies that

$$\sum_{k\geq 1} \frac{1}{\|d_k\|^2} < \infty. \tag{2.14}$$

We conclude that the iteration can fail only if $\|d_k\| \to \infty$ sufficiently rapidly. The method of proof used by Zoutendijk for the FR method consists in showing that, if (2.13) holds, then $\|d_k\|^2$ can grow at most linearly, i.e.,

$$\|d_k\|^2 \leq c\,k,$$

for some constant c. This contradicts (2.14), proving (2.12).

The analysis for *inexact* line searches that satisfy Zoutendijk's condition can proceed along the same lines if one can show that the iteration satisfies

$$\cos\theta_k \geq c\,\|g_k\|/\|d_k\|, \tag{2.15}$$

for some positive constant c. Then, this relation can be used instead of (2.9) to give (2.10), and the rest of the analysis is as in the case of exact line searches.

Al-Baali [1] shows that the FR method gives (2.15) if the steplength satisfies the *strong* Wolfe conditions:

$$f(x_k + \alpha_k d_k) \leq f(x_k) + \sigma_1 \alpha_k \langle g_k, d_k \rangle \tag{2.16}$$

$$|\langle g(x_k + \alpha_k d_k), d_k \rangle| \leq -\sigma_2 \langle g_k, d_k \rangle, \tag{2.17}$$

where $0 < \sigma_1 < \sigma_2 < 1$. In fact, it is necessary to require that $\sigma_2 < \frac{1}{2}$ for the result to hold. He thus shows that (2.12) holds for the FR method.

Al-Baali's result is also remarkable in another respect. By establishing (2.15), which by (2.1) is equivalent to

$$\langle g_k, d_k \rangle \leq -c \, \|g_k\|^2, \tag{2.18}$$

he proved that the FR method using the strong Wolfe conditions (with $\sigma_2 < \frac{1}{2}$) always generates descent directions. Prior to this result it was believed that it was necessary to enforce the descent condition while doing the line search.

In this paper we use the approach described above to establish the global convergence of various algorithms with inexact line searches. As we do so, we will repeatedly encounter (2.18), which appears to be a natural way of guaranteeing descent for conjugate gradient methods. We call (2.18) the *sufficient descent condition.* The first class of methods we consider, in §3, is related to the FR method. We show that any method of the form (1.2)–(1.3) is globally convergent if β_k satisfies $|\beta_k| \leq \beta_k^{\mathrm{FR}}$. The result readily suggests a new implementation of the PR method that preserves its efficiency and assures its convergence.

In §4, we study methods with $\beta_k \geq 0$ that are, in some sense, related to the PR method. A particular case is the following adaptation of the PR method, which consists in restricting β_k to positive values: we let

$$\beta_k = \max\{\beta_k^{\mathrm{PR}}, 0\}. \tag{2.19}$$

The motivation for this strategy arises from Powell's analysis of the PR method. Powell [19] assumes that the line search always finds the first stationary point, and shows that there is a twice continuously differentiable function and a starting point such that the sequence of gradients generated by the PR method stays bounded away from zero. Since Powell's example requires that some consecutive search directions become almost contrary, and since this can only be achieved (in the case of exact line searches) when $\beta_k < 0$, Powell [20] suggests modifying the PR method as in (2.19). In §4 we show that this choice of β_k does indeed result in global convergence, both for exact and inexact line searches. Moreover, we show that the analysis also applies to a family of methods with $\beta_k \geq 0$ that share a common property with the PR method—we call this Property (*).

3. Iterations constrained by the FR method.

In this section we will see that it is possible to obtain global convergence if the parameter β_k is appropriately bounded in magnitude. We consider a method of the form (1.2)–(1.3), where β_k is any scalar such that

$$|\beta_k| \leq \beta_k^{\mathrm{FR}}, \tag{3.1}$$

for all $k \geq 2$, and where the steplength satisfies the strong Wolfe conditions (2.16)–(2.17) with $\sigma_2 < \frac{1}{2}$. Note that Zoutendijk's result, Theorem 2.1, holds in this case, since the strong Wolfe conditions imply the Wolfe conditions (2.4)–(2.5). The next two results are based upon the work of Al-Baali [1] for the FR method, and are slightly stronger than those given by Touati-Ahmed and Storey [24].

LEMMA 3.1. *Suppose that Assumptions* 2.1 *hold. Consider any method of the form* (1.2)–(1.3), *where β_k satisfies* (3.1), *and where the steplength satisfies the Wolfe condition* (2.17) *with $0 < \sigma_2 < \frac{1}{2}$. Then, the method generates descent directions d_k satisfying*

$$-\frac{1}{1-\sigma_2} \leq \frac{\langle g_k, d_k\rangle}{\|g_k\|^2} \leq \frac{2\sigma_2 - 1}{1-\sigma_2}, \qquad k = 1, \cdots. \tag{3.2}$$

Proof. The proof is by induction. The result clearly holds for $k = 1$ since the middle term equals -1 and $0 \leq \sigma_2 < 1$. Assume that (3.2) holds for some $k \geq 1$. This implies that $\langle g_k, d_k\rangle < 0$, since

$$\frac{2\sigma_2 - 1}{1-\sigma_2} < 0, \tag{3.3}$$

by the condition $0 < \sigma_2 < \frac{1}{2}$. From (1.2) and (1.4) we have

$$\frac{\langle g_{k+1}, d_{k+1}\rangle}{\|g_{k+1}\|^2} = -1 + \beta_{k+1}\frac{\langle g_{k+1}, d_k\rangle}{\|g_{k+1}\|^2} = -1 + \frac{\beta_{k+1}}{\beta_{k+1}^{\mathrm{FR}}}\frac{\langle g_{k+1}, d_k\rangle}{\|g_k\|^2}. \tag{3.4}$$

Using the line search condition (2.17) we have

$$|\beta_{k+1}\langle g_{k+1}, d_k\rangle| \leq -\sigma_2|\beta_{k+1}|\langle g_k, d_k\rangle,$$

which, together with (3.4), gives

$$-1 + \sigma_2\frac{|\beta_{k+1}|}{\beta_{k+1}^{\mathrm{FR}}}\frac{\langle g_k, d_k\rangle}{\|g_k\|^2} \leq \frac{\langle g_{k+1}, d_{k+1}\rangle}{\|g_{k+1}\|^2} \leq -1 - \sigma_2\frac{|\beta_{k+1}|}{\beta_{k+1}^{\mathrm{FR}}}\frac{\langle g_k, d_k\rangle}{\|g_k\|^2}.$$

From the left-hand side of the induction hypothesis (3.2), we obtain

$$-1 - \frac{|\beta_{k+1}|}{\beta_{k+1}^{\mathrm{FR}}}\frac{\sigma_2}{1-\sigma_2} \leq \frac{\langle g_{k+1}, d_{k+1}\rangle}{\|g_{k+1}\|^2} \leq -1 + \frac{|\beta_{k+1}|}{\beta_{k+1}^{\mathrm{FR}}}\frac{\sigma_2}{1-\sigma_2}.$$

Using the bound (3.1), we conclude that (3.2) holds for $k + 1$. □

Lemma 3.1 achieves three objectives: (i) it shows that all search directions are descent directions, and the upper bound in (3.2) shows that the sufficient descent condition (2.18) holds; (ii) the bounds on $\langle g_k, d_k\rangle$ impose a limit on how fast $\|d_k\|$ can grow when the gradients are not small, as we will see in the next theorem; (iii) from (2.1) and (3.2) we see that there are positive constants c_1 and c_2 such that

$$c_1\frac{\|g_k\|}{\|d_k\|} \leq \cos\theta_k \leq c_2\frac{\|g_k\|}{\|d_k\|}. \tag{3.5}$$

Therefore, for the FR method or any method with $|\beta_k| \leq \beta_k^{\mathrm{FR}}$, we have that $\cos\theta_k$ is proportional to $\|g_k\|/\|d_k\|$. We will make good use of this fact later on.

THEOREM 3.2. *Suppose that Assumptions* 2.1 *hold. Consider any method of the form* (1.2)–(1.3), *where β_k satisfies* (3.1), *and where the steplength satisfies the strong Wolfe conditions* (2.16)–(2.17), *with $0 < \sigma_1 < \sigma_2 < \frac{1}{2}$. Then*

$$\liminf_{k\to\infty}\|g_k\| = 0.$$

Proof. From (2.17) and Lemma 3.1 we have

$$|\langle g_k, d_{k-1}\rangle| \leq -\sigma_2 \langle g_{k-1}, d_{k-1}\rangle \leq \frac{\sigma_2}{1-\sigma_2}\|g_{k-1}\|^2. \tag{3.6}$$

Thus from (1.2) and (3.1),

$$\begin{aligned}\|d_k\|^2 &\leq \|g_k\|^2 + 2|\beta_k|\,|\langle g_k, d_{k-1}\rangle| + \beta_k^2\|d_{k-1}\|^2\\ &\leq \|g_k\|^2 + \frac{2\sigma_2}{1-\sigma_2}|\beta_k|\,\|g_{k-1}\|^2 + \beta_k^2\|d_{k-1}\|^2\\ &\leq \left(\frac{1+\sigma_2}{1-\sigma_2}\right)\|g_k\|^2 + \beta_k^2\|d_{k-1}\|^2.\end{aligned}$$

Applying this relation repeatedly, defining $\hat{\sigma} := (1+\sigma_2)/(1-\sigma_2) \geq 1$, and using the condition $|\beta_k| \leq \beta_k^{\mathrm{FR}}$, we have

$$\begin{aligned}\|d_k\|^2 &\leq \hat{\sigma}\|g_k\|^2 + \beta_k^2(\hat{\sigma}\|g_{k-1}\|^2 + \beta_{k-1}^2\|d_{k-2}\|^2)\\ &\leq \hat{\sigma}\|g_k\|^4 \sum_{j=1}^{k} \|g_j\|^{-2}.\end{aligned}$$

Let us now assume that $\|g_k\| \geq \gamma > 0$ for all k. This implies, by (2.3), that

$$\|d_k\|^2 \leq \frac{\hat{\sigma}\bar{\gamma}^4}{\gamma^2}\, k. \tag{3.7}$$

We now follow the reasoning described in §2. From the left inequality in (3.5) and Zoutendijk's result (2.7), we obtain (2.10). If the gradients are bounded away from zero, (2.10) implies (2.14). We conclude the proof by noting that (3.7) and (2.14) are incompatible. □

This theorem suggests the following globally convergent modification of the PR method. It differs from that considered by Touati-Ahmed and Storey [24] in that it allows for negative values of β_k. For all $k \geq 2$ let

$$\beta_k = \begin{cases} -\beta_k^{\mathrm{FR}} & \text{if } \beta_k^{\mathrm{PR}} < -\beta_k^{\mathrm{FR}},\\ \beta_k^{\mathrm{PR}} & \text{if } |\beta_k^{\mathrm{PR}}| \leq \beta_k^{\mathrm{FR}},\\ \beta_k^{\mathrm{FR}} & \text{if } \beta_k^{\mathrm{PR}} > \beta_k^{\mathrm{FR}}.\end{cases} \tag{3.8}$$

This strategy avoids one of the main disadvantages of the FR method, as we will now discuss.

We have observed in numerical tests that the FR method with inexact line searches sometimes slows down away from the solution: the steps become very small and this behavior can continue for a very large number of iterations, unless the method is restarted. This behavior was observed earlier by Powell [18], who provides an explanation, under the assumption of exact line searches. It turns out that his argument can be extended to the case of inexact line searches, due to (3.5). The argument is as follows. Suppose that at iteration k an unfortunate search direction is generated, such that $\cos\theta_k \approx 0$, and that $x_{k+1} \approx x_k$. Thus $\|g_{k+1}\| \approx \|g_k\|$, and

$$\beta_{k+1}^{\mathrm{FR}} \approx 1. \tag{3.9}$$

Moreover, by (3.5),

$$\|g_{k+1}\| \approx \|g_k\| \ll \|d_k\|.$$

From this relation, (3.9), and (1.2), we see that $\|d_{k+1}\| \approx \|d_k\| \gg \|g_{k+1}\|$, which by (3.5) implies that $\cos\theta_{k+1} \approx 0$. The argument can therefore start all over again. In §6 we give a numerical example demonstrating this behavior.

The PR method would behave quite differently from the FR method in this situation. If $g_{k+1} \approx g_k$, then $\beta_{k+1}^{\mathrm{PR}} \approx 0$, so that by (1.2) and (3.5), $\cos\theta_{k+1} \gg \cos\theta_k$. Thus the PR method would recover from that situation. Let us now consider the behavior of method (3.8) in these circumstances. We have seen that $\beta_{k+1}^{\mathrm{FR}} \approx 1$, and $\beta_{k+1}^{\mathrm{PR}} \approx 0$, in this case. The method (3.8) will thus set $\beta_{k+1} = \beta_{k+1}^{\mathrm{PR}}$, as desired. It is reassuring that the modification (3.8), which falls back on the FR method to ensure global convergence, avoids the inefficiencies of this method.

The previous discussion highlights a property of the PR method that is not shared by the FR method: when the step is small, β_k^{PR} will be small. This property is essential for the analysis given in the next section, where a method that possesses it will be said to have Property (*).

It is natural to ask if the bound $|\beta_k| \leq \beta_k^{\mathrm{FR}}$ can be replaced by

$$|\beta_k| \leq c\,\beta_k^{\mathrm{FR}}, \tag{3.10}$$

where $c > 1$ is some suitable constant. We have not been able to establish global convergence in this case (although, by modifying Lemma 3.1, one can show that the descent property of the search directions can still be obtained provided $\sigma_2 < 1/(2c)$). In fact, one can prove the following negative result.

PROPOSITION 3.3. *Consider the method* (1.2)–(1.3), *with a line search that always chooses the first positive stationary point of* $\xi_k(\alpha) := f(x_k + \alpha d_k)$. *There exists a twice continuously differentiable objective function of three variables, a starting point, and a choice of* β_k *satisfying* (3.10) *for some constant* $c > 1$, *such that the sequence of gradients* $\{\|g_k\|\}$ *is bounded away from zero.*

Proof. The objective function is taken from the fourth example of Powell [19]. It is twice continuously differentiable. For this function, there is a starting point from which the PR method with a line search providing the first stationary point fails to converge, in the sense that $\|g_k\| \geq \gamma > 0$ for all k. Therefore, using (1.5) and (2.3), we have for all $k \geq 2$,

$$|\beta_k^{\mathrm{PR}}| \leq \frac{2\bar{\gamma}^2}{\gamma^2}.$$

Now, suppose that we computed (but did not use) β_k^{FR}. We would see that for all $k \geq 2$,

$$\beta_k^{\mathrm{FR}} \geq \frac{\gamma^2}{\bar{\gamma}^2}.$$

Combining the two inequalities we obtain

$$|\beta_k^{\mathrm{PR}}| \leq \frac{2\bar{\gamma}^4}{\gamma^4}\,\beta_k^{\mathrm{FR}}.$$

Therefore, if the constant c in (3.10) is chosen larger than $2\bar{\gamma}^4/\gamma^4$, the PR parameter β_k^{PR} in Powell's example would always satisfy (3.10). □

We end this section by making an observation about the restart criterion of Powell [18]. Even though this criterion was designed to ensure the convergence of Beale's method, we will apply it to the PR method, and see that it has some of the flavor of the modifications described in this section. Powell [18] suggests restarting if the following inequality is violated

$$|\langle g_k, g_{k-1}\rangle| \leq \nu \, \|g_{k-1}\|^2,$$

where ν is a small positive constant. (Powell actually uses g_k instead of g_{k-1} in the right-hand side, but one can argue for either choice.) From (1.4) and (1.5),

$$\beta_k^{\text{PR}} = \beta_k^{\text{FR}} - \frac{\langle g_k, g_{k-1}\rangle}{\|g_{k-1}\|^2}.$$

Applying the restart criterion to the PR method we see that a restart is *not* necessary as long as

$$\beta_k^{\text{FR}} - \nu \leq \beta_k^{\text{PR}} \leq \beta_k^{\text{FR}} + \nu.$$

Once more, β_k^{FR} appears as a measure of the adequacy of β_k^{PR}, but this measure is quite different from (3.1). In the next section we will view Powell's restart criterion from a somewhat different angle.

4. Methods related to the PR method with nonnegative β_k. We now turn our attention to methods with $\beta_k \geq 0$ for all k. In §2 we mentioned that a motivation for placing this restriction comes from the example of Powell, in which the PR method cycles without obtaining the solution. Another reason for keeping $\beta_k \geq 0$ is that it allows us to easily enforce the descent property of the algorithm, as we will now discuss.

Let us consider the iteration (1.2)–(1.3) with any $\beta_k \geq 0$. We will require the *sufficient descent condition*

$$\langle g_k, d_k\rangle \leq -\sigma_3 \|g_k\|^2, \tag{4.1}$$

for some $0 < \sigma_3 \leq 1$ and for all $k \geq 1$. In contrast to the FR method, the strong Wolfe conditions (2.16)–(2.17) no longer ensure (4.1). Note, from (1.2), that

$$\langle g_k, d_k\rangle = -\|g_k\|^2 + \beta_k \langle g_k, d_{k-1}\rangle. \tag{4.2}$$

Therefore, to obtain descent for an inexact line search algorithm, one needs to ensure that the last term is not too large. Suppose that we perform a line search along d_{k-1}, enforcing the Wolfe (or strong Wolfe) conditions, to obtain x_k. If $\langle g_k, d_{k-1}\rangle \leq 0$, the nonnegativity of β_k implies that the sufficient descent condition (4.1) holds. Moreover, if (4.1) is not satisfied, then $\langle g_k, d_{k-1}\rangle > 0$, which means that a one-dimensional minimizer has been bracketed. In this case it is easy to apply a line search algorithm, such as that given by Lemaréchal [12], Fletcher [7], or Moré and Thuente [15], to reduce $|\langle g_k, d_{k-1}\rangle|$ sufficiently and obtain (4.1). This will be discussed in detail in §6.

We now prove a global convergence result for methods that are related to the PR method, but that allow only nonnegative values of β_k. The idea of our analysis is simple, but is somewhat concealed in the proofs. We establish the results by contradiction, assuming that the gradients are bounded away from zero:

$$\text{for some } \gamma > 0, \qquad \|g_k\| \geq \gamma \quad \text{for all } k \geq 1. \tag{4.3}$$

Lemma 4.1 shows that in this case the direction of search changes slowly, asymptotically, and Lemma 4.2 proves that a certain fraction of the steps are not too small.

In Theorem 4.3 we show that these two results contradict the assumption that the iterates stay in the bounded level set $\mathcal{L}$. We conclude that a subsequence of the iterates converges to a stationary point.

For the results that follow, we do not specify a particular line search strategy. We only assume that the line search satisfies the following three properties:

(i) all iterates remain in the level set $\mathcal{L}$ defined in Assumptions 2.1:

$$\{x_k\} \subset \mathcal{L}; \tag{4.4}$$

(ii) the Zoutendijk condition (2.7) holds; and

(iii) the sufficient descent condition (4.1) holds.

We mentioned in §2 that the Wolfe line search, as well as the ideal line search (2.6), ensure Zoutendijk's condition and reduce f at each step, which implies (4.4). An exact line search satisfies the sufficient descent condition (4.1), because in this case $\langle g_k, d_k\rangle = -\|g_k\|^2$, and in §6 we describe an inexact line search procedure that satisfies the Wolfe conditions and (4.1) when $\beta_k \geq 0$. Therefore the results of this section apply to both ideal and practical line searches.

For the rest of the section, we assume that convergence does not occur in a finite number of steps, i.e., $g_k \neq 0$ for all k.

LEMMA 4.1. *Suppose that Assumptions* 2.1 *hold. Consider the method* (1.2)–(1.3)*, with $\beta_k \geq 0$, and with any line search satisfying both the Zoutendijk condition* (2.7) *and the sufficient descent condition* (4.1). *If* (4.3) *holds, then $d_k \neq 0$ and*

$$\sum_{k\geq 2} \|u_k - u_{k-1}\|^2 < \infty, \tag{4.5}$$

where $u_k := d_k/\|d_k\|$.

Proof. First, note that $d_k \neq 0$, for otherwise (4.1) would imply $g_k = 0$. Therefore, u_k is well defined. Now, let us define

$$r_k := \frac{-g_k}{\|d_k\|} \quad \text{and} \quad \delta_k := \frac{\beta_k\|d_{k-1}\|}{\|d_k\|}. \tag{4.6}$$

From (1.2), we have for $k \geq 2$:

$$u_k = r_k + \delta_k u_{k-1}. \tag{4.7}$$

Using the identity $\|u_k\| = \|u_{k-1}\|$ and (4.7), we have

$$\|r_k\| = \|u_k - \delta_k u_{k-1}\| = \|\delta_k u_k - u_{k-1}\| \tag{4.8}$$

(the last equality can be verified by squaring both sides). Using the condition $\delta_k \geq 0$, the triangle inequality, and (4.8), we obtain

$$\begin{aligned} \|u_k - u_{k-1}\| &\leq \|(1+\delta_k)u_k - (1+\delta_k)u_{k-1}\| \\ &\leq \|u_k - \delta_k u_{k-1}\| + \|\delta_k u_k - u_{k-1}\| \\ &= 2\|r_k\|. \end{aligned} \tag{4.9}$$

Now, by (2.1) and (4.1), we have

$$\cos\theta_k \geq \sigma_3 \frac{\|g_k\|}{\|d_k\|}.$$

This relation, Zoutendijk's condition (2.7), and (4.6) imply

$$\sum_{k\geq 2} \frac{\|g_k\|^4}{\|d_k\|^2} = \sum_{k\geq 2} \|r_k\|^2\|g_k\|^2 < \infty.$$

Using (4.3), we obtain

$$\sum_{k\geq 2} \|r_k\|^2 < \infty,$$

which together with (4.9) completes the proof. □

Of course, condition (4.5) does not imply the convergence of the sequence $\{u_k\}$, but shows that the search directions u_k change slowly, asymptotically.

Lemma 4.1 applies to any choice of $\beta_k \geq 0$. To proceed, we need to require, in addition, that β_k be small when the step $s_{k-1} = x_k - x_{k-1}$ is small. We saw in §3 that the PR method possesses this property and that it prevents the inefficient behavior of the FR method from occurring. We now state this property formally.

PROPERTY (∗). *Consider a method of the form* (1.2)–(1.3), *and suppose that*

$$0 < \gamma \leq \|g_k\| \leq \bar{\gamma}, \tag{4.10}$$

for all $k \geq 1$. Under this assumption we say that the method has Property (∗) *if there exist constants $b > 1$ and $\lambda > 0$ such that for all k:*

$$|\beta_k| \leq b, \tag{4.11}$$

and

$$\|s_{k-1}\| \leq \lambda \implies |\beta_k| \leq \frac{1}{2b}. \tag{4.12}$$

It is easy to see that under Assumptions 2.1 the PR and HS methods have Property (∗). For the PR method, using the constants γ and $\bar{\gamma}$ in (4.10), we can choose $b := 2\bar{\gamma}^2/\gamma^2$ and $\lambda := \gamma^2/(2L\bar{\gamma}b)$. Then we have, from (1.5) and (4.10),

$$|\beta_k^{\text{PR}}| \leq \frac{(\|g_k\| + \|g_{k-1}\|)\|g_k\|}{\|g_{k-1}\|^2} \leq \frac{2\bar{\gamma}^2}{\gamma^2} = b,$$

and when $\|s_{k-1}\| \leq \lambda$, we have from (2.2),

$$|\beta_k^{\text{PR}}| \leq \frac{\|y_{k-1}\|\|g_k\|}{\|g_{k-1}\|^2} \leq \frac{L\lambda\bar{\gamma}}{\gamma^2} = \frac{1}{2b}.$$

For the HS method, we assume that the descent condition (4.1) and the second Wolfe condition (2.5) are satisfied. Then

$$\begin{aligned}\langle d_{k-1}, y_{k-1}\rangle &= \langle d_{k-1}, g_k\rangle - \langle d_{k-1}, g_{k-1}\rangle \\ &\geq -(1-\sigma_2)\langle g_{k-1}, d_{k-1}\rangle \\ &\geq (1-\sigma_2)\sigma_3\|g_{k-1}\|^2 \\ &\geq (1-\sigma_2)\sigma_3\gamma^2.\end{aligned}$$

Using this in (1.6) we obtain

$$|\beta_k^{\text{HS}}| \leq \frac{2\bar{\gamma}^2}{(1-\sigma_2)\sigma_3\gamma^2} =: b.$$

Now define $\lambda := (1-\sigma_2)\sigma_3\gamma^2/(2L\bar{\gamma}b)$. Using (2.2) we see that if $\|s_{k-1}\| \leq \lambda$, then

$$|\beta_k^{\text{HS}}| \leq \frac{L\lambda\bar{\gamma}}{(1-\sigma_2)\sigma_3\gamma^2} = \frac{1}{2b}.$$

It is clear that many other choices of β_k give rise to algorithms with Property (*). For example, if β_k has Property (*), so do $|\beta_k|$ and $\beta_k^+ := \max\{\beta_k, 0\}$.

The next lemma shows that if the gradients are bounded away from zero, and if the method has Property (*), then a fraction of the steps cannot be too small. We let $\mathbf{N}^*$ denote the set of positive integers, and for $\lambda > 0$ we define

$$\mathcal{K}^\lambda := \{i \in \mathbf{N}^* : i \geq 2, \|s_{i-1}\| > \lambda\},$$

i.e., the set of integers corresponding to steps that are larger than λ. We will need to consider groups of Δ consecutive iterates, and for this purpose we define

$$\mathcal{K}_{k,\Delta}^\lambda := \{i \in \mathbf{N}^* : k \leq i \leq k+\Delta-1, \|s_{i-1}\| > \lambda\}.$$

Let $|\mathcal{K}_{k,\Delta}^\lambda|$ denote the number of elements of $\mathcal{K}_{k,\Delta}^\lambda$ and let $\lfloor\cdot\rfloor$ and $\lceil\cdot\rceil$ denote, respectively, the floor and ceiling operators.

LEMMA 4.2. *Suppose that Assumptions* 2.1 *hold. Consider the method* (1.2)–(1.3), *with any line search satisfying* (4.4), *the Zoutendijk condition* (2.7), *and the sufficient descent condition* (4.1), *and assume that the method has Property* (*). *Suppose also that* (4.3) *holds. Then there exists* $\lambda > 0$ *such that, for any* $\Delta \in \mathbf{N}^*$ *and any index* k_0, *there is a greater index* $k \geq k_0$ *such that*

$$|\mathcal{K}_{k,\Delta}^\lambda| > \frac{\Delta}{2}.$$

Proof. We proceed by contradiction. Suppose that

$$\left\{\begin{array}{l}\text{for any } \lambda > 0, \text{ there exists } \Delta \in \mathbf{N}^* \text{ and } k_0 \text{ such that}\\ \text{for any } k \geq k_0, \text{ we have } |\mathcal{K}_{k,\Delta}^\lambda| \leq \frac{\Delta}{2}.\end{array}\right. \tag{4.13}$$

Assumptions 2.1 and equations (4.4) and (4.3) imply that (4.10) holds. Since the method has Property (*), there exists $\lambda > 0$ and $b > 1$ such that (4.11) and (4.12) hold for all k. For this λ, let Δ and k_0 be given by (4.13).

For any given index $l \geq k_0 + 1$, we have

$$\begin{aligned}\|d_l\|^2 &\leq (\|g_l\| + |\beta_l|\,\|d_{l-1}\|)^2\\ &\leq 2\|g_l\|^2 + 2\beta_l^2\|d_{l-1}\|^2\\ &\leq 2\bar{\gamma}^2 + 2\beta_l^2\|d_{l-1}\|^2,\end{aligned}$$

where the second inequality follows from the fact that, for any scalars a and b, we have $2ab \leq a^2 + b^2$, and hence $(a+b)^2 \leq 2a^2 + 2b^2$. By induction, we obtain

$$\|d_l\|^2 \leq c\,(1 + 2\beta_l^2 + 2\beta_l^2 2\beta_{l-1}^2 + \cdots + 2\beta_l^2 2\beta_{l-1}^2 \cdots 2\beta_{k_0}^2), \tag{4.14}$$

where c depends on $\|d_{k_0-1}\|$, but not on the index l. Let us consider a typical term in (4.14):

$$2\beta_l^2 2\beta_{l-1}^2 \cdots 2\beta_k^2, \tag{4.15}$$

where

$$k_0 \le k \le l. \tag{4.16}$$

We now divide the $2(l-k+1)$ factors of (4.15) into groups of 2Δ elements, i.e., if $N := \lfloor (l-k+1)/\Delta \rfloor$, then (4.15) can be divided into N or $N+1$ groups, as follows:

$$(2\beta_{l_1}^2 \cdots 2\beta_{k_1}^2), \cdots, (2\beta_{l_N}^2 \cdots 2\beta_{k_N}^2), \tag{4.17}$$

and possibly

$$(2\beta_{l_{N+1}}^2 \cdots 2\beta_k^2), \tag{4.18}$$

where $l_i = l-(i-1)\Delta$, for $i = 1, \cdots, N+1$, and $k_i = l_{i+1}+1$, for $i = 1, \cdots, N$. Note from (4.16) that $k_i \ge k_0$, for $i = 1, \cdots, N$, so that we can apply (4.13) for $k = k_i$. We thus have

$$p_i := |\mathcal{K}_{k_i,\Delta}^\lambda| \le \frac{\Delta}{2}, \qquad i = 1, \cdots, N. \tag{4.19}$$

This means that in the range $[k_i, k_i + \Delta - 1]$ there are exactly p_i indices j such that $\|s_{j-1}\| > \lambda$, and thus there are $(\Delta - p_i)$ indices with $\|s_{j-1}\| \le \lambda$. Using this fact, (4.11), and (4.12), we examine a typical factor in (4.17),

$$\begin{aligned} 2\beta_{l_i}^2 \cdots 2\beta_{k_i}^2 &\le 2^\Delta b^{2p_i} \left(\frac{1}{2b}\right)^{2(\Delta - p_i)} \\ &= 2^{\Delta - 2\Delta + 2p_i} b^{2p_i - 2\Delta + 2p_i} \\ &= \left(2b^2\right)^{2p_i - \Delta} \\ &\le 1, \end{aligned}$$

since by (4.19), $2p_i - \Delta \le 0$ and $2b^2 > 1$. Therefore each of the factors in (4.17) is less than or equal to 1, and so is their product. For the last group of factors, given in (4.18), we simply use (4.11):

$$2\beta_{l_{N+1}}^2 \cdots 2\beta_k^2 < (2b^2)^\Delta.$$

We conclude that each term on the right-hand side of (4.14) is bounded by $(2b^2)^\Delta$, and as a result we have

$$\|d_l\|^2 \le c\,(l - k_0 + 2), \tag{4.20}$$

for a certain positive constant c independent of l. In other words, we have shown that $\|d_l\|^2$ grows at most linearly, and we now obtain a contradiction as described in §2. Recalling that (4.1) implies condition (2.15) and using the Zoutendijk condition (2.7), we obtain that

$$\gamma^4 \sum_{k\ge1} \frac{1}{\|d_k\|^2} \le \sum_{k \ge 1} \frac{\|g_k\|^4}{\|d_k\|^2} < \infty.$$

This contradicts (4.20), concluding the proof. □

THEOREM 4.3. *Suppose that Assumptions* 2.1 *hold. Consider the method* (1.2)–(1.3) *with the following three properties:*

(i) *$\beta_k \geq 0$ for all k;*
(ii) *the line search satisfies* (4.4)*, the Zoutendijk condition* (2.7)*, and the sufficient descent condition* (4.1)*;*
(iii) *Property* $(*)$ *holds.*

Then $\liminf \|g_k\| = 0$.

Proof. We proceed by contradiction, assuming (4.3). Therefore, the conditions of Lemmas 4.1 and 4.2 hold. Defining $u_i := d_i/\|d_i\|$, as before, we have for any two indices l, k, with $l \geq k$:

$$\begin{aligned} x_l - x_{k-1} &= \sum_{i=k}^{l} \|s_{i-1}\| u_{i-1} \\ &= \sum_{i=k}^{l} \|s_{i-1}\| u_{k-1} + \sum_{i=k}^{l} \|s_{i-1}\|(u_{i-1} - u_{k-1}). \end{aligned}$$

Taking norms,

$$\sum_{i=k}^{l} \|s_{i-1}\| \leq \|x_l - x_{k-1}\| + \sum_{i=k}^{l} \|s_{i-1}\| \|u_{i-1} - u_{k-1}\|.$$

By (4.4) and Assumptions 2.1 we have that the sequence $\{x_k\}$ is bounded, and thus there exists a positive constant B such that $\|x_k\| \leq B$, for all $k \geq 1$. Thus

$$\sum_{i=k}^{l} \|s_{i-1}\| \leq 2B + \sum_{i=k}^{l} \|s_{i-1}\| \|u_{i-1} - u_{k-1}\|. \tag{4.21}$$

Let $\lambda > 0$ be given by Lemma 4.2. Following the notation of this lemma, we define $\Delta := \lceil 8B/\lambda \rceil$. By Lemma 4.1, we can find an index k_0 such that

$$\sum_{i \geq k_0} \|u_i - u_{i-1}\|^2 \leq \frac{1}{4\Delta}. \tag{4.22}$$

With this Δ and k_0, Lemma 4.2 gives an index $k \geq k_0$ such that

$$|\mathcal{K}_{k,\Delta}^{\lambda}| > \frac{\Delta}{2}. \tag{4.23}$$

Next, for any index $i \in [k, k + \Delta - 1]$, we have, by the Cauchy–Schwarz inequality and (4.22),

$$\begin{aligned} \|u_{i-1} - u_{k-1}\| &\leq \sum_{j=k}^{i-1} \|u_j - u_{j-1}\| \\ &\leq (i-k)^{1/2} \left(\sum_{j=k}^{i-1} \|u_j - u_{j-1}\|^2 \right)^{1/2} \\ &\leq \Delta^{1/2} \left(\frac{1}{4\Delta} \right)^{1/2} = \frac{1}{2}. \end{aligned}$$

Using this relation and (4.23) in (4.21), with $l = k + \Delta - 1$, we have

$$2B \geq \frac{1}{2} \sum_{i=k}^{k+\Delta-1} \|s_{i-1}\| > \frac{\lambda}{2} |\mathcal{K}_{k,\Delta}^{\lambda}| > \frac{\lambda\Delta}{4}.$$

Thus $\Delta < 8B/\lambda$, which contradicts the definition of Δ. □

Since the PR and HS methods have Property (*), the previous theorem applies to them provided we restrict β_k to be nonnegative. This suggests, among other things, the following formulae:

$$\beta_k = \max\{\beta_k^{\text{PR}}, 0\}, \tag{4.24}$$

$$\beta_k = |\beta_k^{\text{PR}}|, \tag{4.25}$$

and the corresponding formulae for the HS method. Of particular interest are inexact line searches, such as the Wolfe search. We formally state the convergence result for (4.24)—a choice of β_k suggested by Powell [20].

COROLLARY 4.4. *Suppose that Assumptions* 2.1 *hold. Consider the method* (1.2)–(1.3) *with* $\beta_k = \max\{\beta_k^{\text{PR}}, 0\}$, *and with a line search satisfying the Wolfe conditions* (2.4)–(2.5) *and the sufficient descent condition* (4.1). *Then* $\liminf \|g_k\| = 0$.

We conclude this section by noting the relationship between (4.24), which can be viewed as an automatic restarting procedure, and Powell's restarting criterion. The latter states that a restart is not needed as long as

$$|\langle g_k, g_{k-1}\rangle| \leq \nu \|g_k\|^2, \tag{4.26}$$

where we now use g_k and not g_{k-1} in the right-hand side, and where ν is a small positive constant. By (1.5) the condition $\beta_k^{\text{PR}} \geq 0$ is equivalent to

$$\langle g_k, g_{k-1}\rangle \leq \|g_k\|^2.$$

Thus (4.24) can be viewed as a less restrictive restarting test than (4.26). It follows that the global convergence result of Corollary 4.4 also applies to the PR method with Powell's restart (4.26), provided $\nu \leq 1$.

5. Discussion.

In §3 we saw that global convergence is obtained for any β_k in the interval $\mathcal{I}_1 = [-\beta_k^{\text{FR}}, \beta_k^{\text{FR}}]$, and in §4 we proved global convergence for any β_k with Property (*) contained in the interval $\mathcal{I}_2 = [0, \infty)$. We now ask whether these results can be combined to obtain larger intervals of admissible β_k. In particular, since the PR method has Property (*), we ask whether global convergence is obtained by restricting β_k^{PR} to the larger interval $\mathcal{I}_1 \cup \mathcal{I}_2$, i.e., by letting

$$\beta_k = \begin{cases} \beta_k^{\text{PR}} & \text{if } \beta_k^{\text{PR}} \geq -\beta_k^{\text{FR}}, \\ -\beta_k^{\text{FR}} & \text{otherwise.} \end{cases}$$

Interestingly enough, global convergence cannot be guaranteed, and this is shown by the fourth example of Powell [19]. In this example, the sequence $\{\beta_k^{\text{PR}}/\beta_k^{\text{FR}}\}$ has exactly three accumulation points:

$$-\frac{1}{3}, 1, \text{ and } 10.$$

Therefore, there exists an index k_0 such that $\beta_k = \beta_k^{\text{PR}} \geq -\beta_k^{\text{FR}}$, for all $k \geq k_0$. Now the function can be modified and the starting point can be changed so that the PR method generates, from the new initial point $\tilde{x}_1$, a sequence $\{\tilde{x}_k\}$ with $\tilde{x}_k = x_{k+k_0-2}$, for $k \geq 2$. In this modified example, we have $\tilde{\beta}_k^{\text{PR}} \geq -\tilde{\beta}_k^{\text{FR}}$, for all $k \geq 2$, but the sequence of gradients is bounded away from zero.

There is another example in which intervals of admissible β_k cannot be combined. Any method of the form (1.2)–(1.3) with a line search giving $\langle g_k, d_{k-1}\rangle = 0$ for all k, and with $\beta_k \in \mathcal{I}_3 = [-1, 1]$, is globally convergent. This is easy to see, since in this case

$$\|d_k\|^2 \leq \|g_k\|^2 + \|d_{k-1}\|^2 \leq \cdots \leq \bar{\gamma}^2 k,$$

where $\bar{\gamma}$ is an upper bound on $\|g(x)\|$. Therefore $\|d_k\|^2$ grows at most linearly, and global convergence follows by the arguments given in §2. On the other hand, Corollary 4.4 shows that the PR method is convergent if restricted to $\mathcal{I}_2 = [0, \infty)$. However, the PR method may not converge if β_k^{PR} is restricted to $\mathcal{I}_3 \cup \mathcal{I}_2 = [-1, \infty)$. The argument is again based on the counterexample of Powell and on the fact that $\beta_k^{\text{PR}} \geq -\frac{1}{4}$ for all k (this is proved by means of the Cauchy–Schwarz inequality; see Powell [19]). Therefore, in this example $\beta_k^{\text{PR}} \in [-1, \infty)$, but convergence is not obtained.

Therefore we are not able to generalize the results of §§3 and 4, and instead look more closely at the conditions used in these sections. We ask under what conditions is $\beta_k^{\text{PR}} \geq 0$, or $\beta_k^{\text{PR}} \geq -\beta_k^{\text{FR}}$. For strictly convex quadratic functions and exact line searches, the PR method coincides with the FR method. Since β_k^{FR} is always positive, so is β_k^{PR}. Let us now consider strongly convex functions. It turns out that in this case β_k^{PR} can be negative, and in fact can be less than $-\beta_k^{\text{FR}}$.

PROPOSITION 5.1. *There exists a C^∞ strongly convex function of two variables and a starting point x_1 for which the* PR *method with exact line searches gives $\beta_3^{\text{PR}} < -\beta_3^{\text{FR}} < 0$.*

Proof. Let us introduce the following strictly convex quadratic function $\tilde{f}$ of two variables $x = (x_{(1)}, x_{(2)})$:

$$\tilde{f}(x) := x_{(1)}^2 + \frac{1}{2}x_{(2)}^2,$$

with gradient and Hessian (the Euclidean scalar product is assumed)

$$\nabla\tilde{f}(x) = \begin{pmatrix} 2x_{(1)} \\ x_{(2)} \end{pmatrix}, \quad \nabla^2\tilde{f}(x) = \begin{pmatrix} 2 & 0 \\ 0 & 1 \end{pmatrix}.$$

Starting from the point $x_1 = (-3, 3)$, the PR method with exact line searches gives

$$\nabla\tilde{f}(x_1) = \begin{pmatrix} -6 \\ 3 \end{pmatrix}, \quad \tilde{\alpha}_1 = \frac{5}{9}, \quad \text{and} \quad \tilde{x}_2 = \frac{1}{3}\begin{pmatrix} 1 \\ 4 \end{pmatrix}.$$

Next, it finds

$$\nabla\tilde{f}(\tilde{x}_2) = \frac{2}{3}\begin{pmatrix} 1 \\ 2 \end{pmatrix}, \quad \tilde{\beta}_2^{\text{PR}} = \frac{4}{81}, \quad \tilde{d}_2 = -\frac{10}{27}\begin{pmatrix} 1 \\ 4 \end{pmatrix}, \quad \text{and} \quad \tilde{\alpha}_2 = \frac{9}{10}.$$

The third point is the solution point $x_* = (0, 0)$.

We now perturb the function $\tilde{f}$ inside the ball $B(0, 1) := \{x : x_{(1)}^2 + x_{(2)}^2 < 1\}$, defining

$$f(x) := \tilde{f}(x) + \epsilon\psi(x),$$

where the function ψ will be such that

$$\psi(x) = 0 \quad \forall x \notin B(0,1), \tag{5.1}$$

and ϵ will be a small positive number. As the line joining x_1 and $\tilde{x}_2$ does not intersect the closure of $B(0,1)$, we see that the PR method on this new function f, starting from the same point x_1, will give $x_2 = \tilde{x}_2$ and $d_2 = \tilde{d}_2$. We now show how to choose the function ψ and the number $\epsilon > 0$ so that f is strongly convex and $\beta_3^{\rm PR}$ is negative.

We take for ψ a function of the form

$$\psi(x) := \eta(x)\ell(x),$$

where ℓ is the linear function

$$\ell(x) := 4x_{(1)} - x_{(2)},$$

and η is a C^∞ function satisfying

$$\eta(x) = \begin{cases} 1 & \text{if } x \in B(0,\frac{1}{2}) \\ 0 & \text{if } x \notin B(0,1). \end{cases}$$

Clearly, ψ satisfies (5.1), and has bounded second-order derivatives. Therefore, by choosing ϵ sufficiently small, say $0 < \epsilon < \epsilon_1$, the Hessian of f will be uniformly positive definite and f will be a C^∞ strongly convex function.

Now, when the function f is determined in this manner, there is a unique minimum of f from x_2 in the direction d_2. As

$$\nabla f(0) = \nabla \tilde{f}(0) + \epsilon \nabla \psi(0) = \epsilon \begin{pmatrix} 4 \\ -1 \end{pmatrix}$$

is orthogonal to $d_2 = \tilde{d}_2$, the one-dimensional minimum is still obtained at $x_3 = (0,0)$ (but this is no longer the solution point). Therefore,

$$\beta_3^{\rm PR} + \beta_3^{\rm FR} = \frac{2|\nabla f(0)|^2 - \langle \nabla f(0), \nabla \tilde{f}(x_2)\rangle}{|\nabla \tilde{f}(x_2)|^2} = \frac{34\epsilon^2 - 4\epsilon/3}{20/9}.$$

We see that $\beta_3^{\rm PR} < -\beta_3^{\rm FR} < 0$, if $0 < \epsilon < \epsilon_2 := 2/51$. By taking $\epsilon \in \ (0, \min(\epsilon_1, \epsilon_2))$, we obtain the desired result. □

This proposition shows that the convergence result given by Polak and Ribière [16], which was obtained for strongly convex functions and exact line searches, is not a consequence of Theorem 4.3, since the latter requires $\beta_k \geq 0$. Nor is it a consequence of Theorem 3.2, because Proposition 5.1 shows that $\beta_k^{\rm PR}$ can lie outside the interval $[-\beta_k^{\rm FR}, \beta_k^{\rm FR}]$.

6. Numerical experiments. We have tested several of the algorithms suggested by the convergence analysis of this paper, on the collection of large test problems given in Table 1.

The starting points used are those given in the references. For the problems of Moré, Garbow, and Hillstrom [14], we set the parameter factor equal to 1; for test problems 8, 9 and 10, starting point 3 from the reference was used. We verified that, in each run, all the methods converged to the same solution point; otherwise the problem was not included in the test set. The problems are not numbered consecutively because they belong to a larger test set. Since conjugate gradient methods are mainly useful for large problems, our test problems have at least 100 variables.

The following are the methods tested; they differ only in the choice of β_k and, possibly, in the line search.

TABLE 1
List of test functions.

Problem	Name	Reference	n
2	Calculus of variations 2	Gill and Murray [9]	100, 200
3	Calculus of variations 3	Gill and Murray [9]	100, 200
6	Generalized Rosenbrock	Moré et al. [14]	100, 500
8	Penalty 1	Gill and Murray [9]	100, 1000
9	Penalty 2	Gill and Murray [9]	100
10	Penalty 3	Gill and Murray [9]	100, 1000
28	Extended Powell singular	Moré et al. [14]	100, 1000
31	Brown almost linear	Moré et al. [14]	100, 200
38	Tridiagonal 1	Buckley and LeNir [3]	100, 1000
39	Linear minimal surface	Toint [23]	121, 961
40	Boundary-value problem	Toint [23]	100
41	Broyden tridiagonal nonlinear	Toint [23]	100
42	Extended ENGVL1	Toint [23]	1000, 10000
43	Extended Freudenstein and Roth	Toint [23]	100, 1000
45	Wrong extended Wood	Toint [23]	100
46(1)	Matrix square root (ns=1)	Liu and Nocedal [13]	100
46(2)	Matrix square root (ns=2)	Liu and Nocedal [13]	100
47	Sparse matrix square root	Liu and Nocedal [13]	100, 1000
48	Extended Rosenbrock	Moré et al. [14]	1000, 10000
49	Extended Powell	Moré et al. [14]	100, 1000
50	Tridiagonal 2	Toint [23]	100, 1000
51	Trigonometric	Moré et al. [14]	100, 1000
52	Penalty 1 (2nd version)	Moré et al. [14]	1000, 10000
53	INRIA u1ts0.4 (u0=0.95)	Gilbert and Lemaréchal [8]	403
54	INRIA u1cr1.2	Gilbert and Lemaréchal [8]	455
55	INRIA u1cr1.3	Gilbert and Lemaréchal [8]	1559

1. FR: The Fletcher–Reeves method.
2. PR-FR: The Polak–Ribière method constrained by the FR method, as in (3.8).
3. PR: The Polak–Ribière method.
4. PR$^+$: The Polak–Ribière method allowing only positive values of β_k^{PR}, as in (4.24).

For the line search we used the algorithm of Moré and Thuente [15]. This algorithm finds a point satisfying the strong Wolfe conditions (2.16)–(2.17). We used the values $\sigma_1 = 10^{-4}$ and $\sigma_2 = 0.1$, which, by Theorem 3.2, ensure that methods FR and PR–FR are globally convergent. The line search for the PR and PR$^+$ methods was performed as follows. We first found a point satisfying the strong Wolfe conditions, using the values of σ_1 and σ_2 mentioned above. If at this point the directional derivative of f is negative, we know that the sufficient descent condition (4.1) holds for the PR$^+$ method, and we terminate the line search (this was discussed at the beginning of §4). On the other hand, if the directional derivative is positive, the algorithm of Moré and Thuente has bracketed a one-dimensional minimizer, and if the line search iteration is continued it will give, in the limit, a point x_k with $\langle g_k, d_{k-1}\rangle = 0$. By continuity and (4.2) it is clear that the line search will find a point satisfying the sufficient descent condition (4.1) in a finite number of iterations. In the numerical tests we set $\sigma_3 = 10^{-2}$ in (4.1). This line search can fail to produce a descent direction for

TABLE 2
Smaller problems.

		FR	PR–FR		PR	PR$^+$	
P	N	it/f-g	it/f-g	mod	it/f-g	it/f-g	mod
2	100	405/827	405/820	351	400/812	400/812	0
3	100	1313/2627	1313/2627	1313	1299/2599	1299/2599	0
6	100	*	261/547	95	256/529	254/525	1
8	100	10/36	15/49	12	9/39	12/47	2
9	100	7/20	8/22	6	8/25	7/20	2
10	100	116/236	93/191	91	118/244	119/244	1
28	100	1426/2855	1291/2584	1289	120/280	168/382	3
31	100	2/3	2/3	1	1/4	1/4	0
38	100	70/142	70/142	47	71/144	71/144	0
39	121	*	59/122	4	59/122	59/122	0
40	100	175/351	175/351	175	132/266	132/266	0
41	100	29/60	24/50	1	24/50	24/50	0
42	1000	10/27	9/25	8	10/34	9/30	2
43	100	16/41	14/39	13	16/44	13/37	1
45	100	*	74/166	66	37/90	45/109	3
46(1)	100	617/1238	253/510	248	257/518	257/518	0
46(2)	100	886/1776	251/506	243	251/506	251/506	0
47	100	151/306	59/122	50	60/124	60/124	0
48	1000	79/185	71/172	66	26/73	23/70	3
49	100	1426/2855	1291/2584	1289	117/281	168/382	3
50	100	72/146	72/146	52	72/146	72/146	0
51	100	202/409	42/94	12	45/103	45/103	0
52	1000	3/10	3/10	2	4/12	4/12	2

the PR method if it terminates at a point with negative directional derivative, and if $\beta_k < 0$ (see the discussion in §4). We used it, nevertheless, because we know of no line search algorithm that is guaranteed to satisfy the strong Wolfe conditions and also provide the descent property for the PR method. Fortunately, in our tests the line search strategy described above always succeeded for the PR method.

Our numerical experience with conjugate gradient methods indicates that it is advantageous to perform a reasonably accurate line search. Therefore, in addition to setting σ_2 to the small number 0.1, we ensured that the line search evaluated the function at least twice. The choice of the initial trial value for the line search is also important. For the first iteration we set it to $1/\|g_1\|$, and for subsequent iterations we used the formula recommended by Shanno and Phua [21], which is based on quadratic interpolation.

The tests were performed on a SPARCstation 1, using FORTRAN in double precision. All runs were stopped when

$$\|g(x_k)\|_\infty < 10^{-5}(1 + |f(x_k)|),$$

except for the INRIA problems for which the runs were stopped when the value of the function had reached a given threshold ($f_{\text{stop}} = 10^{-12}$ for ults0.4, $f_{\text{stop}} = -0.8876\ 10^{-2}$ for ulcrl.2 and $f_{\text{stop}} = -0.10625\ 10^{-1}$ for ulcrl.3). The results in Tables 2 and 3 are given in the form: (number of iterations)/(number of function evaluations). The number given under the column "mod" for method PR–FR denotes

TABLE 3
Larger problems.

		FR	FR–PR		PR	PR$^+$	
P	N	it/f-g	it/f-g	mod	it/f-g	it/f-g	mod
2	200	703/1424	701/1420	596	701/1420	701/1420	0
3	200	2808/5617	2808/5617	2808	2631/5263	2631/5263	0
6	500	*	1107/2231	433	1068/2151	1067/2149	1
8	1000	12/39	9/34	7	6/28	10/42	2
10	1000	138/281	145/299	142	165/338	165/338	0
28	1000	533/1102	1369/2741	1366	212/473	97/229	3
31	200	2/4	2/4	1	1/5	1/5	0
38	1000	264/531	263/529	217	262/527	262/527	0
39	961	*	143/290	5	142/287	142/287	0
42	10000	6/26	6/26	5	7/28	6/26	1
43	1000	10/27	15/38	15	10/33	9/29	2
47	1000	422/849	114/233	92	113/231	113/231	0
48	10000	61/143	130/283	123	24/73	19/62	4
49	1000	568/1175	1369/2741	1366	212/473	97/229	3
50	1000	274/551	273/549	245	274/551	274/551	0
51	1000	231/467	40/91	5	40/92	40/92	0
52	10000	4/15	4/15	4	3/13	3/13	1
53	403	**	233/494	130	237/508	237/508	0
54	455	**	44/91	7	44/87	44/87	0
55	1559	23/47	23/49	15	23/47	23/47	0

TABLE 4
Relative performance, in terms of function evaluations.

FR	FR-PR	PR	PR$^+$
> 4.07	1.55	1.02	1.00

the number of iterations for which $|\beta_k^{\mathrm{PR}}| > \beta_k^{\mathrm{FR}}$. For method PR$^+$, "mod" denotes the number of iterations for which $\beta_k^{\mathrm{PR}} < 0$. If the limit of 9999 function evaluations was exceeded the run was stopped; this is indicated by "*." The sign "**" means that the run stopped because the line search procedure described above failed to find a steplength. This occurred when the stopping criterion was very demanding.

It is interesting to note that β_k^{PR} was constrained in most of the iterations of the method PR–FR, but was only rarely modified in the PR$^+$ method. Many of the problems were run again for a larger number of variables. The results are given in Table 3.

In these runs the methods were implemented without restarting. We also performed tests in which the methods were restarted along the steepest descent direction every n iterations. (Since n is large, very few restarts were performed.) The FR method improved substantially, but this method was still the least efficient of the four. The other three methods performed similarly with and without restarts, and we will not present the results here.

In Table 4 we summarize the results of Tables 2 and 3 by giving the relative number of function evaluations required by the four methods. We have normalized the numbers so that PR$^+$ corresponds to 1. The symbol $>$ means that FR requires more function evaluations than the number given, since for some runs the method was stopped prematurely; also, problems 53 and 54, in which FR failed, were not taken into account.

The FR method is clearly the least efficient, requiring a very large number of function evaluations in some problems. The performance of methods PR–FR, PR, and PR^+ appears to be comparable, but we would not like to draw any firm conclusions from our experiments. PR–FR appears to be preferable to FR, but we have no explanation for its poor performance on some problems. A close examination of the runs provided no new insights about the behavior of the methods. The global convergence analysis of this paper has not suggested a method that is clearly superior to PR. For that it may be necessary to study the convergence rate or other measures of efficiency of the methods. We leave this for a future study.

We conclude by giving an example that illustrates the inefficient behavior of the FR method, as predicted in §3. For problem 45 with $n = 100$, we observed that for hundreds of iterations $\cos\theta_k$ stays fairly constant, and is of order 10^{-2}, while the steps $\|x_k - x_{k-1}\|$ are of order 10^{-2} to 10^{-3}. This causes the algorithm to require a very large number of iterations to approach the solution. A restart along the steepest descent direction terminates this cycle of bad search directions and tiny steps. A similar behavior was observed in several other problems.

Acknowledgment. We are grateful to Michael Powell for showing us how to shorten the proof of Lemma 4.1, and for several other helpful comments.

REFERENCES

[1] M. Al-Baali, *Descent property and global convergence of the Fletcher–Reeves method with inexact line search*, IMA J. Numer. Anal., 5 (1985), pp. 121–124.

[2] P. Baptist and J. Stoer, *On the relation between quadratic termination and convergence properties of minimization algorithms, Part* II, *Applications*, Numer. Math., 28 (1977), pp. 367–392.

[3] A. Buckley and A. Lenir, *QN-like variable storage conjugate gradients*, Math. Programming, 27 (1983), pp. 155–175.

[4] A. Cohen, *Rate of convergence of several conjugate gradient algorithms*, SIAM J. Numer. Anal., 9 (1972), pp. 248–259.

[5] H. P. Crowder and P. Wolfe, *Linear convergence of the conjugate gradient method*, IBM J. Res. Develop., 16 (1972), pp. 431–433.

[6] R. Fletcher and C. M. Reeves, *Function minimization by conjugate gradients*, Comput. J., 7 (1964), pp. 149–154.

[7] R. Fletcher, *Practical Methods of Optimization*, John Wiley, New York.

[8] J. Ch. Gilbert and C. Lemaréchal, *Some numerical experiments with variable-storage quasi-Newton algorithms*, Math. Programming B, 45 (1989), pp. 407–436.

[9] P. E. Gill and W. Murray, *The numerical solution of a problem in the calculus of variations*, in Recent Mathematical Developments in Control, D.J. Bell, ed., Academic Press, New York, 1973, pp. 97–122.

[10] ———, *Conjugate-gradient methods for large-scale nonlinear optimization*, Technical report SOL 79-15, Department of Operations Research, Stanford University, Stanford, CA, 1979.

[11] P. E. Gill, W. Murray and M. H. Wright, *Practical Optimization*, Academic Press, New York, 1981.

[12] C. Lemaréchal, *A view of line searches*, in Optimization and Optimal Control, A. Auslander, W. Oettli, and J. Stoer, eds.,Lecture Notes in Control and Information Science 30, Springer-Verlag, Berlin, 1981, pp. 59–78.

[13] D. C. Liu and J. Nocedal, *Test results of two limited memory methods for large scale optimization*, Report NAM 04, Dept. of Electrical Engineering and Computer Science, Northwestern University, Evanston, IL, 1988.

[14] J. J. Moré, B. S. Garbow, and K. E. Hillstrom, *Testing unconstrained optimization software*, ACM Trans. Math. Software, 7 (1981), pp. 17-41.

[15] J. J. Moré and D. J. Thuente, *On line search algorithms with guaranteed sufficient decrease*, Mathematics and Computer Science Division Preprint MCS-P153-0590, Argonne National Laboratory, Argonne, IL, 1990.

[16] E. Polak and G. Ribière, *Note sur la convergence de méthodes de directions conjuguées*, Revue Française d'Informatique et de Recherche Opérationnelle, 16 (1969), pp. 35–43.

[17] M. J. D. Powell, *Some convergence properties of the conjugate gradient method*, Math. Programming, 11 (1976), pp. 42–49.

[18] ———, *Restart procedures for the conjugate gradient method*, Math. Programming, 12 (1977), pp. 241–254.

[19] ———, *Nonconvex minimization calculations and the conjugate gradient method*, in Lecture Notes in Mathematics 1066, Springer-Verlag, Berlin, 1984, pp. 122–141.

[20] ———, *Convergence properties of algorithms for nonlinear optimization*, SIAM Rev., 28 (1986), pp. 487–500.

[21] D. F. Shanno and K. H. Phua, *Remark on algorithm* 500 : *Minimization of unconstrained multivariate functions*, ACM Trans. Math. Software, 6 (1980), pp. 618–622.

[22] J. Stoer, *On the relation between quadratic termination and convergence properties of minimization algorithms, Part* I, *Theory*, Numer. Math., 28 (1977), pp. 343–366.

[23] Ph. L. Toint, *Test problems for partially separable optimization and results for the routine PSPMIN*, Report Nr 83/4, Department of Mathematics, Facultés Universitaires de Namur, Namur, Belgium, 1983.

[24] D. Touati-Ahmed and C. Storey, *Efficient hybrid conjugate gradient techniques*, J. Optim. Theory Appl., 64 (1990), pp. 379–397.

[25] P. Wolfe, *Convergence conditions for ascent methods*, SIAM Rev., 11 (1969), pp. 226–235.

[26] ———, *Convergence conditions for ascent methods* II: *some corrections*, SIAM Rev., 13 (1971), pp. 185–188.

[27] G. Zoutendijk, *Nonlinear programming, computational methods*, in Integer and Nonlinear Programming, J. Abadie, ed., North-Holland, Amsterdam, 1970, pp. 37–86.

SIAM J. OPTIMIZATION
Vol. 2, No. 1, pp. 43–54, February 1992

ERROR BOUND AND CONVERGENCE ANALYSIS OF MATRIX SPLITTING ALGORITHMS FOR THE AFFINE VARIATIONAL INEQUALITY PROBLEM*

ZHI-QUAN LUO† AND PAUL TSENG‡

Abstract. Consider the affine variational inequality problem. It is shown that the distance to the solution set from a feasible point near the solution set can be bounded by the norm of a natural residual at that point. This bound is then used to prove linear convergence of a matrix splitting algorithm for solving the symmetric case of the problem. This latter result improves upon a recent result of Luo and Tseng that further assumes the problem to be monotone.

Key words. affine variational inequality, linear complementarity, error bound, matrix splitting, linear convergence

AMS(MOS) subject classifications. 49, 90

1. Introduction. Let M be an $n \times n$ matrix and let q be a vector in $\Re^n$, the n-dimensional Euclidean space. Let X be a polyhedral set in $\Re^n$. We consider the following *affine* variational inequality problem associated with M, q, and X:

$$\text{find an } x^* \in X \text{ satisfying } \langle x - x^*, Mx^* + q \rangle \geq 0 \quad \forall x \in X. \tag{1.1}$$

The problem (1.1) is well known in optimization and contains as special cases linear (and quadratic) programming, bimatrix games, etc. (see Cottle and Dantzig [CoD68]). When X is the nonnegative orthant in $\Re^n$, it is called the linear complementarity problem (LCP). We will not attempt to survey the literature on this problem, which is vast. Expository articles on the subject include [CoD68], [Eve71], [CGL80], and [Mur88]. For a discussion of variational inequality problems in general, see [Aus76], [BeT89], [CGL80], and [KiS80].

Let X^* denote the set of solutions of the affine variational inequality problem (1.1), which we assume from here on to be nonempty. It is well known (and not difficult to see from the convexity of X) that X^* is precisely the set of fixed points of the nonlinear mapping $x \mapsto [x - Mx - q]^+$, where $[\cdot]^+$ denotes the orthogonal projection onto X, i.e., $[x]^+ = \arg\min_{z \in X} \|x - z\|$ and $\|\cdot\|$ denotes the usual Euclidean norm in $\Re^n$. In other words, we have

$$X^* = \{\, x^* \in \Re^n \mid x^* = [x^* - Mx^* - q]^+ \,\}. \tag{1.2}$$

(Our notation for the projection operator is nonstandard but has the advantage of simplicity.) Although in general X^* is not convex, it can be shown that X^* is the union of a finite collection of polyhedral sets (see (3.10)).

An important topic in the study of variational inequalities and complementarity problems concerns *error bounds* for estimating the closeness of a point to X^* (see [Pan87], [MaD88], [MaS86]). Such error bounds can serve as termination criteria for

* Received by the editors July 5, 1990; accepted for publication (in revised form) February 27, 1991. This work was supported by U.S. Army Research Office contract DAAL03-86-K-0171 (Center for Intelligent Control Systems), National Science Foundation grant NSF–DDM–8903385, and by a grant from the Science and Engineering Research Board of McMaster University.

† Room 225, Communications Research Laboratory, Department of Electrical and Computer Engineering, McMaster University, Hamilton, Ontario, L8S 4K1, Canada (luozq@sscvax.cis.mcmaster.ca).

‡ Department of Mathematics, GN-50, University of Washington, Seattle, Washington 98195 (tseng@math.washington.edu).

iterative algorithms and can be used to estimate the amount of error allowable in an inexact computation of the iterates (see [Pan86b]). Recently the authors [LuT90] showed that one such bound, based on the norm of the natural residual function

$$\|x - [x - Mx - q]^+\| \tag{1.3}$$

is also useful for analyzing the *rate* of convergence of iterative algorithms for solving (1.1). In particular, they showed that, for the problem of minimizing a certain convex essentially smooth function over a polyhedral set, a bound analogous to the above can be used as the basis for proving the linear convergence of a number of well-known iterative algorithms (applied to solve this problem).

The contribution of this paper is twofold: (i) we show that the error bound (1.3) holds locally for the affine variational inequality problem (1.1) for general M, thus extending a result of [LuT90, §2] for the case where M is symmetric positive semidefinite, (ii) we show, by using the above error bound, that if M is symmetric, then any matrix splitting algorithm using regular Q-splitting, applied to solve (1.1), is linearly convergent. (Here, by linear convergence, we mean linear convergence in the root sense of [OrR70].) This latter result extends the one in [LuT90, §5], which proved linear convergence for the same algorithm under the additional assumption that M is positive semidefinite. It also improves upon the results of Pang [Pan84, §4], [Pan86a, §2], which showed convergence (respectively, weak convergence) for a special case of the algorithm, i.e., one that solves LCP, under the additional assumption that M is nondegenerate (respectively, strictly copositive). Matrix splitting algorithms using regular Q-splitting represent an important class of algorithms for solving affine variational inequality problems and LCPs (see [LiP87]), so the resolution o^f their convergence (and·their rate of convergence) is of great interest. (See §3 for a more detailed discussion of the subject.)

This paper proceeds as follows. In §2, we prove that an error bound based on (1.3) holds for all points in X near X^*. In §3, we consider the special case of (1.1) where M is symmetric and we use the bound of §2 to prove the linear convergence of matrix splitting algorithms using regular Q-splitting, applied to solve this problem. Finally, in §4, we give our conclusion and discuss possible extensions.

We adopt the following notations throughout. For any $x \in \Re^n$ and $y \in \Re^n$, we denote by $\langle x, y\rangle$ the Euclidean inner product of x with y. For any $x \in \Re^n$, we denote by $\|x\|$ the usual Euclidean norm of x, i.e., $\|x\| = \sqrt{\langle x, x\rangle}$. For any two subsets C_1, C_2 of $\Re^n$, we denote by $d(C_1, C_2)$ the usual Euclidean distance between the sets C_1 and C_2, that is,

$$d(C_1, C_2) = \inf_{x \in C_1, y \in C_2} \|x - y\|.$$

For any $k \times l$ matrix A, we denote by A^T the transpose of A, by $\|A\|$ the matrix norm of A induced by the vector norm $\|\cdot\|$ (i.e., $\|A\| = \max_{\|x\|=1} \|Ax\|$), by A_i the ith row of A and, for any subset $I \subseteq \{1, \cdots, k\}$, by A_I the submatrix of A obtained by removing all rows $i \notin I$ of A. Analogously, for any vector $x \in \Re^k$, we denote by x_i the ith coordinate of x and, for any subset $I \subseteq \{1, \cdots, k\}$, by x_I the vector with components x_i, $i \in I$ (with the x_i's arranged in the same order as in x).

2. A local error bound. In this section we show that $d(x, X^*)$ can be bounded from above by the norm of $x - [x - Mx - q]^+$, the natural residual at x, whenever the latter quantity is small. Our proof, like the proof of Theorem 2.3 in [LuT90], exploits heavily the affine structure of the problem.

Since X is a polyhedral set, we can for convenience express it as

$$X = \{ \, x \in \Re^n \mid Ax \geq b \, \},$$

for some $m \times n$ matrix A and some $b \in \Re^m$. Then, for any $x \in X$, the vector $[x - Mx - q]^+$ is simply the unique vector z which, together with some multiplier vector $\lambda \in \Re^m$, satisfies the Kuhn–Tucker conditions

$$(2.1) \qquad z - x + Mx + q - A^T\lambda = 0, \quad Az \geq b, \quad \lambda \geq 0,$$

$$(2.2) \qquad A_i z = b_i \quad \forall i \in I(x), \quad \lambda_i = 0 \quad \forall i \notin I(x),$$

where we denote

$$I(x) = \{ \, i \in \{1, \cdots, n\} \mid A_i[x - Mx - q]^+ = b_i \, \}.$$

We say that an $I \subseteq \{1, \cdots, m\}$ is *active* at a vector $x \in X$ if $z = [x - Mx - q]^+$, together with some $\lambda \in \Re^m$, satisfies (2.1) and

$$(2.3) \qquad A_i z = b_i \quad \forall i \in I, \quad \lambda_i = 0 \quad \forall i \notin I.$$

(Clearly, $I(x)$ is active at x for all $x \in X$.)

The following lemma, due originally to Hoffman [Hof52] (also see [Rob73], [MaS87]), will be used extensively in the analysis to follow.

LEMMA 2.1. *Let B be a $k \times l$ matrix, let C be an $h \times l$ matrix, and let d be a vector in $\Re^h$. There exists a scalar $\tau > 0$ depending on B and C only such that, for any $\bar{x}$ satisfying $C\bar{x} \geq d$ and any $e \in \Re^k$ such that the linear system $By = e$, $Cy \geq d$ is consistent, there is a point $\bar{y}$ satisfying $B\bar{y} = e$, $C\bar{y} \geq d$ with $\|\bar{x} - \bar{y}\| \leq \tau\|B\bar{x} - e\|$.*

We next have the following lemma, which roughly says that if $x \in X$ is sufficiently close to X^*, then those constraint indices that are active at x are also active at some element of X^*.

LEMMA 2.2. *There exists a scalar $\epsilon > 0$ such that, for any $x \in X$ with $\|x - [x - Mx - q]^+\| \leq \epsilon$, $I(x)$ is active at some $x^* \in X^*$.*

Proof. We argue by contradiction. If the claim does not hold, then there would exist an $I \subseteq \{1, \cdots, m\}$ and a sequence of vectors $\{x^1, x^2, \cdots\}$ in X satisfying $I(x^r) = I$ for all r and $x^r - z^r \to 0$, where we let $z^r = [x^r - Mx^r - q]^+$ for all r, and yet there is no $x^* \in X^*$ for which I is active at x^*.

For each r, consider the following linear system in x, z, and λ:

$$x - z - Mx + A^T\lambda = q, \quad Az \geq b, \quad \lambda \geq 0,$$

$$A_i z = b_i \quad \forall i \in I, \quad \lambda_i = 0 \quad \forall i \notin I,$$

$$x - z = x^r - z^r.$$

The above system is consistent since, by $I(x^r) = I$ and (2.1)–(2.2), (x^r, z^r) together with some $\lambda^r \in \Re^m$ is a solution of it. Then, by Lemma 2.1, it has a solution $(\hat{x}^r, \hat{z}^r, \hat{\lambda}^r)$ whose size is bounded by some constant (depending on A and M only) times the size of the right-hand side. Since the right-hand side of the above system is clearly bounded as $r \to \infty$, we have that $\{(\hat{x}^r, \hat{z}^r, \hat{\lambda}^r)\}$ is bounded. Moreover, every one of its cluster points, say $(x^\infty, z^\infty, \lambda^\infty)$, satisfies (cf. $x^r - z^r \to 0$)

$$x^\infty - z^\infty - Mx^\infty + A^T\lambda^\infty = q, \quad Az^\infty \geq b, \quad \lambda^\infty \geq 0,$$

$$A_i z^\infty = b_i \quad \forall i \in I, \quad \lambda_i^\infty = 0 \quad \forall i \notin I,$$

$$x^\infty - z^\infty = 0.$$

This shows that $x^\infty = [x^\infty - Mx^\infty - q]^+$ (cf. (2.1), (2.2)) and that I is active at x^∞ (cf. (2.1), (2.3)), a contradiction of our earlier hypothesis on I. □

By using Lemma 2.2, we can now establish the main result of this section.

THEOREM 2.3. *There exist scalars $\epsilon > 0$ and $\tau > 0$ such that*

$$d(x, X^*) \leq \tau \|x - [x - Mx - q]^+\|$$

for all $x \in X$ with $\|x - [x - Mx - q]^+\| \leq \epsilon$.

Proof. Let ϵ be the scalar given in Lemma 2.2. Consider any $x \in X$ satisfying the hypothesis of Lemma 2.2, and let $z = [x - Mx - q]^+$. Then, by (2.1) and (2.2), there exists some $\lambda \in \Re^m$ satisfying, together with x and z,

$$x - z - Mx + A^T\lambda = q, \quad Az \geq b, \quad \lambda \geq 0,$$

$$A_i z = b_i \quad \forall i \in I(x), \quad \lambda_i = 0 \quad \forall i \notin I(x).$$

By Lemma 2.2, there exists an $x^* \in X^*$ such that $I(x)$ is active at x^*, so the following linear system in (x^*, z^*, λ^*)

$$x^* - z^* - Mx^* + A^T\lambda^* = q, \quad Az^* \geq b, \quad \lambda^* \geq 0,$$

$$A_i z^* = b_i \quad \forall i \in I(x), \quad \lambda_i^* = 0 \quad \forall i \notin I(x), \quad x^* - z^* = 0,$$

is consistent. Moreover, every solution (x^*, z^*, λ^*) of this linear system satisfies $x^* = [x^* - Mx^* - q]^+$ (cf. (2.1), (2.2)) so, by (1.2), $x^* \in X^*$. Upon comparing the above two systems, we see that, by Lemma 2.1, there exists a solution (x^*, z^*, λ^*) to the second system such that

$$\|(x^*, z^*, \lambda^*) - (x, z, \lambda)\| \leq \tau \|x - z\|,$$

where τ is some scalar constant depending on A and M only. Hence

$$\|x^* - x\| \leq \tau \|x - z\|.$$

Since $x^* \in X^*$, so $d(x, X^*) \leq \|x^* - x\|$; this then completes the proof. □

Error bounds for estimating the distance from a point to the solution set, similar to that given in Theorem 2.3, have been fairly well studied. In fact, the same bound had been demonstrated by Pang [Pan87] and by Mathias and Pang [MaP90] to hold globally on X for the special cases of an LCP where M is, respectively, positive *definite* and a P-matrix. The bound has also been demonstrated by the authors [LuT90] to hold locally on X for the special case where M is symmetric and positive semidefinite. (This bound also extends to strongly monotone variational inequality problems [Pan87] and to problems of minimizing a certain convex, essentially smooth, function over a polyhedral set [LuT90].)

Alternative bounds have also been proposed by Mangasarian and Shiau [MaS86] for the special case of an LCP where M is positive semidefinite and for strongly convex programs [MaD88]. These alternative error bounds have the advantage that they hold globally *everywhere* (even for points outside X), whereas the bound of Theorem 2.3 holds only locally on X. Might the latter bound hold globally also? For general matrices M, the answer unfortunately is "no," as shown by an example of a nonsymmetric LCP furnished in [MaS86] (see Example 2.10 therein). What if M is symmetric? The answer is still "no," as shown by the following modification of Example 2.10 in [MaS86].

Example 2.1. Let

$$M = \begin{pmatrix} 0 & 1 \\ 1 & 1 \end{pmatrix}, \quad q = \begin{pmatrix} -1 \\ -2 \end{pmatrix}, \quad X = [0, \infty)^2.$$

It is easily checked that $X^* = \{ (1,1), (0,2) \}$. Let $x(\theta) = (\theta, 1)$, where $\theta \in [0, \infty)$. Then, as $\theta \to \infty$, we have $d(x(\theta), X^*) \to \infty$ but $\|x(\theta) - [x(\theta) - Mx(\theta) - q]^+\|$ remains bounded.

Subsequent to the writing of this paper, we learned that Theorem 2.3 can also be deduced from a result of Robinson [Rob81] on a locally upper Lipschitzian property of polyhedral multifunctions. More precisely, let $R : \Re^n \to \Re^n$ be the natural residual function given by

$$R(x) = x - [x - Mx - q]^+.$$

Then, the inverse of R is a polyhedral multifunction and thus, by Robinson's result [Rob81, Prop. 1], is locally upper Lipschitzian at the origin, that is, there exist scalars $\epsilon > 0$ and $\tau > 0$ such that

$$R^{-1}(z) \subseteq R^{-1}(0) + \tau \|z\| \mathcal{B},$$

for all z with $\|z\| \leq \epsilon$, where $\mathcal{B}$ denotes the unit Euclidean ball in $\Re^n$. This statement is entirely equivalent to Theorem 2.3.

3. Linear convergence of matrix splitting algorithm for the symmetric case.

In this section we further assume that M is symmetric, in which case the variational inequality problem (1.1) may be formulated as a quadratic program of the form

$$\begin{aligned} &\text{minimize } f(x) \\ &\text{subject to } x \in X, \end{aligned} \tag{3.1}$$

where f is the quadratic function in $\Re^n$ given by

$$f(x) = \tfrac{1}{2}\langle x, Mx \rangle + \langle q, x \rangle. \tag{3.2}$$

It is easily seen that the set of stationary points for (3.1) is precisely X^* (cf. (1.2)), which, by assumption, is nonempty. Note, however, that f may not be bounded from below on X.

Let (B, C) be a *regular splitting* of M (see, e.g., [OrR70], [Kel65], [LiP87]), i.e.,

$$M = B + C, \qquad B - C \text{ is positive definite.} \tag{3.3}$$

Consider the following well-known iterative algorithm for solving (3.1), based on the splitting (B, C).

MATRIX SPLITTING ALGORITHM. At the rth iteration we are given an $x^r \in X$ (with $x^0 \in X$ chosen arbitrarily), and we compute a new iterate x^{r+1} in X satisfying

$$x^{r+1} = [x^{r+1} - Bx^{r+1} - Cx^r - q + h^r]^+, \tag{3.4}$$

where h^r is some n-vector.

The problem of finding an x^{r+1} satisfying (3.4) may be viewed as an affine variational inequality problem, whereby x^{r+1} is the vector in X that satisfies the variational inequality

$$\langle Bx^{r+1} + Cx^r + q - h^r, z - x^{r+1} \rangle \geq 0 \quad \forall z \in X. \tag{3.5}$$

In general, such an x^{r+1} need not exist, in which case the above algorithm would break down. To ensure that this does not happen, we will, following [LiP87], assume that

$$(B, C) \quad \text{is a } Q\text{-splitting} \tag{3.6}$$

or, equivalently, an x satisfying

$$x = [x - Bx - Cx^r - q]^+ \tag{3.7}$$

exists for all r. (For example, (B, C) is a Q-splitting if B is positive definite (see [BeT89], [KiS80]).)

The vector x^{r+1} may be viewed as an inexact solution of (3.7) with h^r as the associated "error" vector (so $h^r = 0$ corresponds to an exact solution of (3.7)). The idea of introducing an error vector in this manner is adopted from Mangasarian [Man90]. Let γ denote the smallest eigenvalue of the symmetric part of $B - C$ (which by hypothesis is positive) and let ϵ be a fixed scalar in $(0, \gamma/2]$. We will consider the following restriction on h^r governing how fast h^r tends to zero:

$$\|h^r\| \leq (\gamma/2 - \epsilon)\|x^r - x^{r+1}\| \quad \forall r. \tag{3.8}$$

The above restriction on h^r provides a *finite* termination criterion for any iterative method used to solve (3.7). To illustrate, fix r and suppose that we have a sequence of points converging to a solution of (3.7). (If B is positive definite, then such a sequence can be generated, for example, by applying the projection iteration $y := [y - \alpha(By + Cx^r + q)]^+$, with α a suitably chosen positive stepsize.) Suppose that the limit is not x^r (otherwise x^r is already in X^*) and let $F : \Re^n \mapsto \Re^n$ be the continuous function given by $F(y) = [y - By - Cx^r - q]^+$. Then, for all points y sufficiently far along in this sequence we have

$$\|(I - B)(y - F(y))\| \leq (\gamma/2 - \epsilon)\|x^r - F(y)\|.$$

(This is because the limit, say $\bar{y}$, is not equal to x^r and satisfies $\bar{y} = F(\bar{y})$.) Take any such point y and set

$$h^r = (I - B)(y - F(y)), \qquad x^{r+1} = F(y).$$

Then, h^r and x^{r+1} satisfy (3.4) and (3.8).

The above matrix splitting algorithm was first proposed by Pang [Pan82], based on the works of Hildreth [Hil57], Cryer [Cry71], Mangasarian [Man77], and others. (Actually, Pang considered the somewhat simpler case of an LCP with no error vector, i.e., X is the nonnegative orthant in $\Re^n$ and $h^r = 0$ for all r.) This algorithm has been studied extensively (see [LiP87], [LuT90], [LuT91], [Man77], [Man90], [Pan82], [Pan84], [Pan86a], and references therein), but, owing to the possible unboundedness of the set of stationary points, its convergence was very difficult to establish and was typically shown under restrictive assumptions on the problem (such as that the stationary point is unique). It was shown only recently that, if M is positive semidefinite (in addition to being symmetric) and f given by (3.2) is bounded from below on X, then the iterates generated by this algorithm converge to a stationary point [LuT91] with a rate of convergence that is at least linear [LuT90, §5]. In this section we show that the same linear convergence result holds for *any* symmetric M, and thus we resolve the issue of convergence (and rate of convergence) for this algorithm on symmetric problems. The convergence of this algorithm for the special case of a symmetric LCP has been studied by Pang (see [Pan84, §4] and [Pan86a, §2]). However, Pang did not analyze the rate of convergence of the algorithm and his convergence results require restrictive assumptions on the problem, such as that the set of stationary points be finite.

The line of our analysis follows that outlined in [LuT90] (also see [LuT92] for a similar analysis) and is based on using the error bound of Theorem 2.3 to show that, asymptotically, the objective function value, evaluated at the new iterate x^{r+1} and at some stationary point, differ by only an order of $\|x^{r+1} - x^r\|^2$ (see (3.19)). This then enables us to show that the objective function values converge at least linearly, from which one can deduce that the iterates converge at least linearly. (This is the main motivation for considering the symmetric case, so that an objective function exists and can be used to monitor the progress of the algorithm. The algorithm itself is well defined whether M is symmetric or not.) On the other hand, because f is not convex and the set of stationary points X^* is not necessarily convex or even connected, a new analysis, different from that in [LuT90], is needed to show the above relation.

We begin our analysis by giving, in the lemma below, a characterization of the connected components of X^* and the behaviour of f over these connected components.

LEMMA 3.1. *Suppose that M is symmetric. Let $C_1, C_2, \cdots, C_t$ denote the connected components of X^*, where t is some positive integer. Then,*

$$X^* = \bigcup_{i=1}^{t} C_i,$$

and the following hold:

(a) *Each C_i is the union of a finite collection of polyhedral sets.*

(b) *The C_i's are properly separated from one another, that is, $d(C_i, C_j) > 0$ for all $i \neq j$.*

(c) *f given by* (3.2) *is constant on each C_i.*

Proof. Since X is a polyhedral set, we can express it as

$$X = \{\ x \in \Re^n \mid Ax \geq b\ \}$$

for some $m \times n$ matrix A and some $b \in \Re^m$. For each $I \subseteq \{1, 2, \cdots, m\}$, let

$$X_I = \{\ x \mid Ax \geq b,\ A_I x = b_I,\ Mx + q = A^T\lambda \\ \text{for some } \lambda \in [0,\infty)^m \text{ with } \lambda_i = 0\ \ \forall i \notin I\ \}. \tag{3.9}$$

Then, each X_I simply comprises those elements of X^* at which I is active (see (2.1) and (2.2)), so it readily follows that

$$X^* = \bigcup_{I \subseteq \{1,\cdots,m\}} X_I. \tag{3.10}$$

Moreover, each X_I, if nonempty, is a polyhedral set. We claim that f is *constant* on each nonempty X_I. To see this, fix any $I \subseteq \{1,\cdots,m\}$ for which X_I is nonempty. Let x and y be any two elements of X_I (possibly equal). Since $x \in X_I$ and $y \in X_I$, we have from (3.9) that $A_I(x - y) = 0$ and there exists some $\lambda \in [0,\infty)^m$ with $My + q = (A_I)^T \lambda_I$. Then we have from (3.2) that

$$\begin{aligned} f(x) - f(y) &= \langle My + q, x - y \rangle + \tfrac{1}{2}\langle x - y, M(x - y)\rangle \\ &= \langle (A_I)^T \lambda_I, x - y \rangle + \tfrac{1}{2}\langle x - y, M(x - y)\rangle \\ &= \langle \lambda_I, A_I(x - y)\rangle + \tfrac{1}{2}\langle x - y, M(x - y)\rangle \\ &= \tfrac{1}{2}\langle x - y, M(x - y)\rangle. \end{aligned}$$

By symmetry, we also have

$$f(y) - f(x) = \tfrac{1}{2}\langle x - y, M(x - y)\rangle,$$

and thus $f(x) = f(y)$. Since the above choice of x and y was arbitrary, then $f(y) = f(x)$ for all $x \in X_I, y \in X_I$.

Since each X_I is connected, it follows from (3.10) that each C_i is the union of a finite collection of nonempty X_I's. Since the nonempty X_I's are polyhedral and the C_i's are, by definition, mutually disjoint, this then proves parts (a) and (b). Since f is constant on each X_I, this also proves part (c). □

(Lemma 3.1(c) is quite remarkable, since the gradient of f does not need to be constant on each C_i, as can be seen from an example.)

By using Theorem 2.3 and Lemma 3.1, we can now prove the main result of this section. (The first third of our proof follows closely that of Theorem 5.1 in [LuT90].)

THEOREM 3.2. *Suppose that M is symmetric and that f given by* (3.2) *is bounded from below on X. Let $\{x^r\}$ be iterates generated by the matrix splitting algorithm* (3.3), (3.4), (3.6), (3.8). *Then $\{x^r\}$ converges at least linearly* (*in the root sense*) *to an element of X^*.*

Proof. First we claim that

$$f(x^{r+1}) - f(x^r) \leq -\epsilon \|x^{r+1} - x^r\|^2 \quad \forall r. \tag{3.11}$$

To see this, fix any r. Since the variational inequality (3.5) holds, then, by plugging in x^r for z in (3.5), we obtain

$$\langle Bx^{r+1} + Cx^r + q - h^r, x^{r+1} - x^r \rangle \leq 0.$$

Also, from $M = B + C$ (cf. (3.3)) and the definition of f (cf. (3.2)), we have that

$$f(x^{r+1}) - f(x^r) = \langle Bx^{r+1} + Cx^r + q, x^{r+1} - x^r \rangle + \langle x^{r+1} - x^r, (C - B)(x^{r+1} - x^r)\rangle / 2.$$

Combining the above two relations then gives

$$\begin{aligned} f(x^{r+1}) - f(x^r) &\leq \langle h^r, x^{r+1} - x^r \rangle + \langle x^{r+1} - x^r, (C - B)(x^{r+1} - x^r) \rangle/2 \\ &\leq \|h^r\| \|x^{r+1} - x^r\| - \gamma \|x^{r+1} - x^r\|^2/2 \\ &\leq -\epsilon \|x^{r+1} - x^r\|^2, \end{aligned}$$

where the last inequality follows from (3.8). Thus, (3.11) holds.

Next we claim that there exists a scalar constant $\kappa_1 > 0$ for which

$$\|x^r - [x^r - Mx^r - q]^+\| \leq \kappa_1 \|x^{r+1} - x^r\| \quad \forall r. \tag{3.12}$$

To see this, fix any r. From (3.4) we have that

$$\begin{aligned} \|x^r - [x^r - Mx^r - q]^+\| &= \|x^r - [x^r - Mx^r - q]^+ - x^{r+1} \\ &\qquad + [x^{r+1} - Bx^{r+1} - Cx^r - q + h^r]^+\| \\ &\leq \|x^r - x^{r+1}\| + \|[x^r - Mx^r - q]^+ \\ &\qquad - [x^{r+1} - Bx^{r+1} - Cx^r - q + h^r]^+\| \\ &\leq 2\|x^r - x^{r+1}\| + \|Mx^r - Bx^{r+1} - Cx^r + h^r\| \\ &\leq 2\|x^r - x^{r+1}\| + \|B(x^r - x^{r+1})\| + \|h^r\| \\ &\leq (2 + \|B\| + \gamma/2)\|x^r - x^{r+1}\|, \end{aligned}$$

where the second inequality follows from the nonexpansive property of the projection operator $[\cdot]^+$, the third inequality follows from $M = B + C$, and the last inequality follows from (3.8). This shows that (3.12) holds with $\kappa_1 = 2 + \|B\| + \gamma/2$.

Since f is bounded from below on X, (3.11) implies

$$\|x^{r+1} - x^r\| \to 0. \tag{3.13}$$

Then we have from (3.12) that $\|x^r - [x^r - Mx^r - q]^+\| \to 0$, so, by Theorem 2.3 (and using (3.12)), there exist a scalar constant $\kappa_2 > 0$ and an index r_1 such that

$$d(x^r, X^*) \leq \kappa_2 \|x^{r+1} - x^r\| \quad \forall r \geq r_1.$$

For each r, let y^r be any element of X^* attaining $\|y^r - x^r\| = d(x^r, X^*)$. Then the above relation implies

$$\|y^r - x^r\| \leq \kappa_2 \|x^{r+1} - x^r\| \quad \forall r \geq r_1, \tag{3.14}$$

which, when combined with (3.13), yields

$$y^r - x^r \to 0. \tag{3.15}$$

Let $C_1, C_2, \cdots, C_t$ denote the connected components of X^*, where t is some positive integer. By Lemma 3.1 (b), the C_i's are properly separated from one another. Since $y^r \in X^*$ for all r and, by (3.13) and (3.15), $y^r - y^{r+1} \to 0$, this implies that the sequence $\{y^r\}$ eventually settles down at one of the C_i's. In other words, there exists a $k \in \{1, \cdots, t\}$ and a scalar $r_2 \geq r_1$ such that

$$y^r \in C_k \quad \forall r \geq r_2.$$

By Lemma 3.1 (c), f is constant on C_k. Let us denote this constant by f^∞. Then the above relation implies

$$f(y^r) = f^\infty \qquad \forall r \geq r_2. \tag{3.16}$$

For any $r \geq r_2$ we have from $y^r \in X^*$ and $x^r \in X$ that $\langle My^r + q, x^r - y^r\rangle \geq 0$ and from the Mean Value Theorem (also using (3.2)) that $f(y^r) - f(x^r) = \langle M\psi^r + q, y^r - x^r\rangle$, for some n-vector ψ^r lying on the line segment joining y^r with x^r. Upon summing these two relations and using (3.16), we obtain

$$\begin{aligned} f^\infty - f(x^r) &\leq \langle M\psi^r - My^r, y^r - x^r\rangle \\ &\leq \|M\|\|\psi^r - y^r\|\|y^r - x^r\| \\ &\leq \|M\|\|y^r - x^r\|^2. \end{aligned}$$

This, together with (3.15), yields

$$\liminf_{r\to\infty} f(x^r) \geq f^\infty. \tag{3.17}$$

We now show that $f(x^r) \to f^\infty$ and estimate the speed at which this convergence takes place. Fix any $r \geq r_2$. Since $r \geq r_1$ (cf. $r_2 \geq r_1$) so that (3.14) holds, this implies

$$\begin{aligned} \langle My^r + q, x^{r+1} - y^r\rangle &\leq \langle My^r + q, x^{r+1} - y^r\rangle + \langle Bx^{r+1} + Cx^r + q - h^r, y^r - x^{r+1}\rangle \\ &= \langle B(x^{r+1} - x^r) + M(x^r - y^r) - h^r, y^r - x^{r+1}\rangle \\ &\leq \left(\|B\|\|x^{r+1} - x^r\| + \|M\|\|x^r - y^r\| + \|h^r\|\right)\|y^r - x^{r+1}\| \\ &\leq \left(\|B\|\|x^{r+1} - x^r\| + \|M\|\kappa_2\|x^{r+1} - x^r\| + \|h^r\|\right) \\ &\qquad \times(\kappa_2 + 1)\|x^{r+1} - x^r\| \\ &\leq \left(\|B\| + \|M\|\kappa_2 + \gamma/2\right)(\kappa_2 + 1)\|x^{r+1} - x^r\|^2, \end{aligned} \tag{3.18}$$

where the first inequality follows from (3.5) with z set to y^r, the equality follows from $C = M - B$ (cf. (3.3)), the third inequality follows from (3.14), and the last inequality follows from (3.8). For convenience, let κ_3 denote the scalar constant on the right-hand side of (3.18). Then we obtain from (3.16) that

$$\begin{aligned} f(x^{r+1}) - f^\infty &= f(x^{r+1}) - f(y^r) \\ &= \langle My^r + q, x^{r+1} - y^r\rangle + \tfrac{1}{2}\langle x^{r+1} - y^r, M(x^{r+1} - y^r)\rangle \\ &\leq \kappa_3\|x^{r+1} - x^r\|^2 + \tfrac{1}{2}\|M\|\|x^{r+1} - y^r\|^2 \\ &\leq \left(\kappa_3 + \tfrac{1}{2}\|M\|(\kappa_2 + 1)^2\right)\|x^{r+1} - x^r\|^2, \end{aligned} \tag{3.19}$$

where the second equality follows from (3.2), the first inequality follows from (3.18), and the last inequality follows from (3.14).

Let κ_4 denote the scalar constant on the right-hand side of (3.19). Then (3.11) and (3.19) yield

$$\begin{aligned} f(x^{r+1}) - f^\infty &\leq \kappa_4\|x^{r+1} - x^r\|^2 \\ &\leq \frac{\kappa_4}{\epsilon}(f(x^r) - f(x^{r+1})) \quad \forall r \geq r_2. \end{aligned}$$

Upon rearranging terms, we find that

$$\left(1 + \frac{\kappa_4}{\epsilon}\right)(f(x^{r+1}) - f^\infty) \leq \frac{\kappa_4}{\epsilon}(f(x^r) - f^\infty) \quad \forall r \geq r_2.$$

On the other hand, we have from (3.17) and the fact that $f(x^r)$ is monotonically decreasing with r (cf. (3.11)) that $f(x^r) \geq f^\infty$ for all r, so the above relation implies that $\{f(x^r)\}$ converges at least linearly (in the root sense) to f^∞. By (3.11), $\{x^r\}$ also converges at least linearly (in the root sense). Since $d(x^r, X^*) \to 0$ (cf. (3.15)), the point to which $\{x^r\}$ converges is an element of X^*. □

Note that we can allow the matrix splitting (B, C) to vary from iteration to iteration, provided that the eigenvalues of the symmetric part of $B - C$ are bounded away from zero and that $\|B\|$ is bounded.

Also note that because f is not convex, the point to which the iterates converge need not be an optimal solution of (3.1). (Finding such an optimal solution is certainly desirable.) On the other hand, it is easily seen from Lemma 3.1(c) and the fact that the f value of the iterates are monotonically decreasing that *local* convergence to an optimal solution holds. In other words, if the initial iterate (namely, x^0) is sufficiently close to the optimal solution set of (3.1), then the point to which the iterates converge is an optimal solution of (3.1).

4. Concluding remarks. In this paper, we have shown that a certain error bound holds locally for the affine variational inequality problem. By using this bound, we are able to prove the linear convergence of matrix splitting algorithms using regular Q-splitting for the symmetric case of the problem.

There are a number of open questions raised by our work. The first question concerns whether the error bound studied here holds globally. Example 2.1 shows that it does not hold globally even when M is symmetric. But what if M, in addition, is positive semidefinite? A "yes" answer to this question would allow us to show *global* linear convergence for the matrix splitting algorithm of §3 on symmetric monotone problems. Also, our convergence result (Theorem 3.2) asserts convergence only when f given by (3.2) is bounded from below on X. If this were not the case, could something meaningful about convergence still be said? Another question concerns whether other error bounds, such as those proposed in [MaD88] and [MaS86], can be used to analyze the convergence of an iterative algorithm, as is done here. Also, can the analysis of §3 be extended to the nonsymmetric case by finding an appropriate "objective function" to work with? Or to the simpler case of a nonsymmetric LCP? (It is well known that any LCP can be converted to a quadratic program. However, except under certain conditions (see [CPV89]), the set of solutions for the former does not need to coincide with the set of stationary points for the latter.)

It would also be worthwhile to find other problem classes for which the error bound studied here holds. Then, we can be hopeful of proving linear convergence results for these other problems.

Acknowledgment. We thank Professor O. L. Mangasarian for pointing out to us the references [MaD88], [MaS87] on error bounds and the examples therein. We thank Professor J.-S. Pang for bringing to our attention the paper [Rob81].

REFERENCES

[Aus76] A. Auslender, *Optimisation: Méthodes Numériques*, Masson, Paris, New York, Barcelona, Milan, 1976.

[BeT89] D. P. Bertsekas and J. N. Tsitsiklis, *Parallel and Distributed Computation: Numerical Methods*, Prentice–Hall, Englewood Cliffs, NJ, 1989.

[CoD68] R. W. Cottle and G. B. Dantzig, *Complementary pivot theory of mathematical programming*, Linear Algebra Appl., 1 (1968), pp. 103–125.

[CGL80] R. W. Cottle, F. Giannessi, and J.-L. Lions, *Variational Inequalities and Complementarity Problems: Theory and Applications*, John Wiley, New York, NY, 1980.

[CPV89] R. W. Cottle, J.-S. Pang, and V. Venkateswaran, *Sufficient matrices and the linear complementarity problem*, Linear Algebra Appl., 114/115 (1989), pp. 231–249.

[Cry71] C. W. Cryer, *The solution of a quadratic programming problem using systematic overrelaxation*, SIAM J. Control Optim., 9 (1971), pp. 385–392.

[Eve71] B. C. Eaves, *The linear complementarity problem*, Management Sci., 17 (1971), pp. 612–635.

[Hil57] C. Hildreth, *A quadratic programming procedure*, Naval Res. Logist. Quart., 4 (1957), pp. 79–85; see also *Erratum*, Naval Res. Logist. Quart., 4 (1957), p. 361.

[Hof52] A. J. Hoffman, *On approximate solutions of systems of linear inequalities*, J. Res. Nat. Bur. Standards, 49 (1952), pp. 263–265.

[Kel65] H. B. Keller, *On the solution of singular and semidefinite linear systems by iteration*, SIAM J. Numer. Anal., 2 (1965), pp. 281–290.

[KiS80] D. Kinderlehrer and G. Stampacchia, *An Introduction to Variational Inequalities and Applications*, Academic Press, New York, 1980.

[LiP87] Y. Y. Lin and J.-S. Pang, *Iterative methods for large convex quadratic programs: A survey*, SIAM J. Control Optim., 25 (1987), pp. 383–411.

[LuT90] Z.-Q. Luo and P. Tseng, *On the linear convergence of descent methods for convex essentially smooth minimization*, Laboratory for Information and Decision Systems Report No. P–1979, Massachusetts Institute of Technology, Cambridge, MA (June 1990); SIAM J. Control Optim., 30 (1992), to appear.

[LuT91] ———, *On the convergence of a matrix–splitting algorithm for the symmetric monotone linear complementarity problem*, SIAM J. Control Optim., 29 (1991), pp. 1037–1060.

[LuT92] ———, *On the convergence of the coordinate descent method for convex differentiable minimization*, J. Optim. Theory Appl., 72 (1992), pp. 7–35.

[Man77] O. L. Mangasarian, *Solution of symmetric linear complementarity problems by iterative methods*, J. Optim. Theory Appl., 22 (1977), pp. 465–485.

[Man90] ———, *Convergence of iterates of an inexact matrix splitting algorithm for the symmetric monotone linear complementarity problem*, SIAM J. Optimization, 1 (1991), pp. 114–122.

[MaD88] O. L. Mangasarian and R. De Leone, *Error bounds for strongly convex programs and (super)linearly convergent iterative schemes for the least 2-norm solution of linear programs*, Appl. Math. Optim., 17 (1988), pp. 1–14.

[MaS86] O. L. Mangasarian and T.-H. Shiau, *Error bounds for monotone linear complementarity problems*, Math. Programming, 36 (1986), pp. 81–89.

[MaS87] ———, *Lipschitz continuity of solutions of linear inequalities, programs and complementarity problems*, SIAM J. Control Optim., 25 (1987), pp. 583–595.

[MaP90] R. Mathias and J.-S. Pang, *Error bounds for the linear complementarity problem with a P-matrix*, Linear Algebra Appl., 132 (1990), pp. 123–136.

[Mur88] K. G. Murty, *Linear Complementarity, Linear and Nonlinear Programming*, Helderman-Verlag, Berlin, 1988.

[OrR70] J. M. Ortega and W. C. Rheinboldt, *Iterative Solution of Nonlinear Equations in Several Variables*, Academic Press, New York, 1970.

[Pan82] J.-S. Pang, *On the convergence of a basic iterative method for the implicit complementarity problem*, J. Optim. Theory Appl., 37 (1982), pp. 149–162.

[Pan84] ———, *Necessary and sufficient conditions for the convergence of iterative methods for the linear complementarity problem*, J. Optim. Theory Appl., 42 (1984), pp. 1–17.

[Pan86a] ———, *More results on the convergence of iterative methods for the symmetric linear complementarity problem*, J. Optim. Theory Appl., 49 (1986), pp. 107–134.

[Pan86b] ———, *Inexact Newton methods for the nonlinear complementarity problem*, Math. Programming, 36 (1986), pp. 54–71.

[Pan87] ———, *A posteriori error bounds for the linearly-constrained variational inequality problem*, Math. Oper. Res., 12 (1987), pp. 474–484.

[Rob73] S. M. Robinson, *Bounds for error in the solution set of a perturbed linear program*, Linear Algebra Appl., 6 (1973), pp. 69–81.

[Rob81] ———, *Some continuity properties of polyhedral multifunctions*, Math. Programming Stud., 14 (1981), pp. 206–214.

SIAM J. OPTIMIZATION
Vol. 2, No. 1, pp. 55–70, February 1992

A LARGE-STEP ANALYTIC CENTER METHOD FOR A CLASS OF SMOOTH CONVEX PROGRAMMING PROBLEMS*

D. DEN HERTOG[†], C. ROOS[†], AND T. TERLAKY[‡]

Abstract. In this paper, a large-step analytic center method for smooth convex programming is proposed. The method is a natural implementation of the classical method of centers. It is assumed that the objective and constraint functions fulfil the so-called Relative Lipschitz Condition, with Lipschitz constant $M > 0$. A great advantage of the method, over the existing path-following methods, is that the steps can be made long by performing linesearches.

In this method linesearches are performed along the Newton direction with respect to a strictly convex potential function when located far away from the central path. When sufficiently close to this path a lower bound for the optimal value is updated. It is proven that the number of iterations required by the algorithm to converge to an ϵ-optimal solution is $O((1 + M^2)\sqrt{n}|\ln \epsilon|)$ or $O((1 + M^2)n|\ln \epsilon|)$, depending on the updating scheme for the lower bound.

Key words. convex programming, analytic center method, Newton method

AMS(MOS) subject classification. 90C25

1. Introduction.

Since Karmarkar [6] presented his projective method for the solution of the linear programming problem in 1984, many other variants have been developed by researchers. Among them are the large-step path-following methods such as those proposed by Roos and Vial [9]; Gonzaga [3]; and Den Hertog, Roos, and Terlaky [1]; and the potential reduction methods such as those proposed by Ye [12], Freund [2], and Gonzaga [4]. The advantages of these methods are that they do not use projective transformations as the projective methods do, and that they do not need to follow the so-called central path closely, contrary to the small-step path-following methods.

In Jarre [5] and Mehrotra and Sun [8] small-step path-following algorithms are proposed for smooth convex programming problems. Again, the great disadvantage of these methods is that they are based on very small stepsizes to remain in the vicinity of the central trajectory. This characteristic makes these methods unattractive for practical use. To accelerate his method, Jarre proposed a (higher-order) extrapolation scheme.

In this paper we propose a large-step path-following method for smooth convex programming problems, which fulfil the so-called Relative Lipschitz Condition. Jarre [5] also uses this condition. Our method is a generalization of the method for linear programming presented in [1] and is also based on Jarre's paper.

In our method we do a linesearch along the Newton direction with respect to a certain strictly convex potential function. If we are close to the current analytic center we update the lower bound somehow, whereafter we do linesearches aiming at getting close to the analytic center associated with the new lower bound. We prove that after a linesearch the potential value reduces with at least a certain constant. Using this result, we prove that the number of iterations required by the algorithm to converge

* Received by the editors April 11, 1990; accepted for publication (in revised form) January 29, 1991.

[†] Faculty of Technical Mathematics and Computer Science, Delft University of Technology, P.O. Box 5031, 2600 GA Delft, the Netherlands.

[‡] Eötvös University, Department of Operations Research, 1088 Budapest, Muzeum Krt. 6-8, Hungary. This work was performed while the author was visiting Delft University of Technology, P.O. Box 5031, 2600 GA Delft, the Netherlands.

to an ϵ-optimal solution is bounded by a polynomial in $|\ln \epsilon|$, the dimension of the problem, and the Lipschitz constant.

We note that Kojima, Mizuno, and Yoshise [7] already proposed a primal–dual potential reduction algorithm for linear complementarity problems. To our knowledge our algorithm is the first large-step algorithm for (a class of) smooth convex programming. Our algorithm can also be viewed as a natural implementation of the classical method of centers. In a coming report we will deal with a natural implementation of the logarithmic barrier function method.

This paper is organized as follows. In §2 we will do some preliminary work. In §3 we describe our algorithm. Then in §4 we prove some lemmas needed for the convergence analysis in §5.

2. Preliminaries. We consider the primal formulation of the smooth convex programming problem,

$$\text{(CP)} \qquad \max \{f_0(y) \ : \ y \in \mathcal{F}\},$$

where $\mathcal{F}$ denotes the feasible region, which is given by

$$\mathcal{F} := \{y \in \mathbb{R}^m \ : f_i(y) \leq 0, \ \ 1 \leq i \leq n\};$$

the functions $-f_0(y)$ and $f_i(y)$, $1 \leq i \leq n$, are convex functions with continuous first- and second-order derivatives in $\mathcal{F}$. We assume an additional smoothness condition, namely, that the Hessian matrix of $f_i(y)$, $0 \leq i \leq n$, fulfils the so-called Relative Lipschitz Condition, which will be specified below. Moreover, we suppose that the interior of the feasible region $\mathcal{F}$, denoted as $\mathcal{F}'$, is nonempty and bounded. This assumption is not essential.

Wolfe's [11] formulation of the dual problem associated with this primal problem is

$$\text{(D)} \qquad \begin{array}{l} \min \ f_0(y) - \sum_{i=1}^n u_i f_i(y), \\ \sum_{i=1}^n u_i \nabla f_i(y) = \nabla f_0(y), \\ u_i \geq 0. \end{array}$$

Note that there is no symmetry between the primal and dual problem, as in linear programming, because the dual problem (D) contains both y and u variables. Moreover, the dual problem is not necessarily convex!

However, it is a well-known result that if y is a feasible solution of (CP) and $(\overline{y}, u)$ is a feasible solution of the dual problem (D), then

$$f_0(y) \leq f_0(\overline{y}) - \sum_{i=1}^{n} u_i f_i(\overline{y}).$$

Due to the assumption that $\mathcal{F}'$ is nonempty, the Slater condition is satisfied, and hence (D) has a minimum solution and the extremum values are equal.

We associate the following potential function with (CP):

$$\phi(y,z) = -q \ln(f_0(y) - z) - \sum_{i=1}^{n} \ln(-f_i(y)),$$

where z is a lower bound for the optimal value z^*, and q is a positive integer value, which will be discussed below. For $q = n$ this potential function is exactly the same as the one used by Jarre [5].

It can be proved that $\phi(y,z)$ is strictly convex on its domain $\mathcal{F}$ (see Jarre [5, p. 8]). It also takes infinite values on the boundary of the feasible set. Hence this potential function achieves the minimal value in its domain (for fixed z) at a unique point, which is denoted by $y(z)$. The necessary and sufficient Karush–Kuhn–Tucker conditions for these minima are:

$$\begin{array}{ll} & f_i(y) \leq 0, \qquad 1 \leq i \leq n, \\ (1) & \sum_{i=1}^n u_i \nabla f_i(y) = \nabla f_0(y), \qquad u \geq 0, \\ & -f_i(y)u_i = \frac{f_0(y)-z}{q}, \qquad 1 \leq i \leq n. \end{array}$$

Using this it can easily be verified that $y(z)$ lies on the so-called central path of the problem, which is the set of analytic centers for $\mathcal{F} \cap \{y : f_0(y) \geq \mu\}$, where μ varies from $-\infty$ to z^*.

We can rewrite $\phi(y,z)$ as

$$\phi(y,z) = -\sum_{i=1}^{n+q} \ln(-f_i(y)),$$

where $-f_i(y) = f_0(y) - z$ for $n+1 \leq i \leq n+q$. The first- and second-order derivatives of $\phi(y,z)$ are given by

$$g(y,z) := \nabla\phi(y,z) = \sum_{i=1}^{n+q} \frac{\nabla f_i(y)}{-f_i(y)}$$

and

$$H(y,z) := \nabla^2\phi(y,z) = \sum_{i=1}^{n+q} \left[\frac{\nabla^2 f_i(y)}{-f_i(y)} + \frac{\nabla f_i(y)\nabla f_i(y)^T}{f_i(y)^2}\right].$$

If no confusion is possible we will write, for shortness' sake, g and H instead of $g(y,z)$ and $H(y,z)$.

In the sequel to this paper we will also use the quadratic approximation $q_y(x,z)$ for $\phi(y,z)$ when x is near the point y, defined as

$$q_y(x,z) := \phi(y,z) + g^T(x-y) + \tfrac{1}{2}(x-y)^T H(x-y).$$

We will use the H-norm $\|.\|_H$ to measure closeness of points, and especially closeness to the central trajectory. The definition of this norm is as follows:

$$\|x\|_H = \sqrt{x^T H x}.$$

Because H is positive definite, $\|.\|_H$ defines a norm.

Having introduced this notation we are able to formulate the Relative Lipschitz Condition:

$$\exists M > 0 : \ \forall v \in \mathbb{R}^m \quad \forall y, y+h \in \mathcal{F}' :$$

$$|v^T(\nabla^2 f_i(y+h) - \nabla^2 f_i(y))v| \leq M\|h\|_H v^T \nabla^2 f_i(y) v, \tag{2}$$

for all $1 \leq i \leq n+1$.

This condition is also used by Jarre [5]. In general, the condition might be hard to check for a given problem.

3. The algorithm. In our algorithm we do not need to stay close to the central path, as in Jarre [5]. If we are far away from the central path we do a linesearch along the Newton direction with respect to $\phi(y,z)$. The Newton direction $p(y,z)$ associated with $\phi(y,z)$ at y is given by

$$p(y,z) = -H(y,z)^{-1}g(y,z) = -H^{-1}g.$$

If no confusion is possible we will write, for shortness' sake, p instead of $p(y,z)$. This process is repeated until we are sufficiently close to the central path. More precisely, we stop doing linesearches if $\|p\|_H \leq \tau$, where τ is a certain tolerance. This proximity criterion is also used by Jarre [5]. In the algorithm we will use $\tau = \frac{1}{8(1+2M)}$, which will appear to be appropriate later on. (Note that $\|p\|_H = 0$ if and only if $y = y(z)$.) If the proximity criterion is satisfied, we update the lower bound z as follows: $\overline{z} := z + \theta(f_0(y) - z)$, for some $0 < \theta < 1$, and the whole process is repeated again and again until some stopping criterion is satisfied. Note that $\overline{z}$ is really a lower bound, because $\overline{z} < f_0(y) \leq z^*$.

We can now describe the algorithm.

Algorithm
Input:
θ is the updating factor, $0 < \theta < 1$;
$\tau = \frac{1}{8(1+2M)}$ is the proximity tolerance;
t is an accuracy parameter, $t \in \mathbb{N}$;
y^0 is a given interior feasible point and $z^0 < f_0(y^0)$ is a lower bound for the optimal value, such that $\|p(y^0, z^0)\|_{H(y^0,z^0)} \leq \tau$ and $z^* - z^0 \leq \frac{1}{\epsilon}$.

```
begin
  $y := y^0$; $z := z^0$;
  while $f_0(y) - z > e^{-t}$ do
  begin (outer step)
    while $\|p\|_H > \tau$ do
    begin (inner step)
      $\bar{\alpha} := \arg\min_{\alpha>0} \{\phi(y + \alpha p, z) : y + \alpha p \in \mathcal{F}'\}$
      $y := y + \bar{\alpha}p$
    end (end inner step)
    $z := z + \theta(f_0(y) - z)$;
  end (end outer step)
end.
```

For finding the initial point that satisfies the input assumptions of the algorithm we refer the reader to Jarre [5] and Mehrotra and Sun [8]. Later on, the "centering assumption" will be alleviated.

4. Preliminary lemmas. In §5 we will prove the complexity result on the Algorithm. In this section we deal with some lemmas that will be needed to obtain an upper bound for the total number of outer and inner iterations. The lemmas are built up as follows:

- Lemma 4.1 gives an upper bound for the error in the quadratic approximation if the functions $-f_0(y)$ and $f_i(y)$, $1 \leq i \leq n$, are linear or quadratic;
- Lemma 4.2 states the same as Lemma 4.1, but now $-f_0(y)$ and $f_i(y)$, $1 \leq i \leq n$, are general convex functions;

- Lemma 4.3 states that if the proximity criterion holds, then y lies close to the exact center $y(z)$ (with respect to the H-norm);
- Lemma 4.4 states that if we do a linesearch along the Newton direction, then a sufficient decrease in the potential value can be guaranteed;
- Lemma 4.5 gives an upper bound for the difference in potential value of the current iterate and the exact center;
- Lemma 4.6 states that if the lower bound is updated, then the potential value increases with a constant;
- Lemma 4.7 gives a relation between the objective value in the exact center and the current point;
- Lemma 4.8 gives an upper bound for the gap between the optimal value and the lower bound z;
- Lemma 4.9 states that the gap $f_0(y(z)) - z$ decreases monotonically if z increases.

The following lemma improves Lemma 2.1 of Jarre [5].

LEMMA 4.1. *If $-f_0(y)$ and $f_i(y)$, $1 \leq i \leq n$, are linear or quadratic functions with positive semidefinite Hessian matrix, and if $y \in \mathcal{F}'$ and $\|d\|_H < 1$, then $y+d \in \mathcal{F}'$ and*

$$|\phi(y+d,z) - q_y(y+d,z)| < \frac{\|d\|_H^3}{3(1-\|d\|_H)}.$$

Proof. We expand $\phi(y+d,z)$ in a Taylor series about y:

$$\phi(y+d,z) = q_y(y+d,z) + \sum_{i=3}^{\infty} t_i, \tag{3}$$

where t_i is the ith-order Taylor term in the expansion. Note that ϕ only takes finite values in $\mathcal{F}'$. Hence, if $\sum_{i=3}^{\infty} t_i$ can be shown to converge for d such that $\|d\|_H < 1$, then it follows that $y + d \in \mathcal{F}'$. It can be proved that

$$|t_i| \leq \frac{1}{i}\|d\|_H^i. \tag{4}$$

The proof of this inequality is quite technical. Therefore, it is omitted here (see Appendix). From (4), and using the fact that $\|d\|_H < 1$, we derive that

$$\sum_{i=3}^{\infty} |t_i| \leq \sum_{i=3}^{\infty} \frac{\|d\|_H^i}{i} \leq \frac{\|d\|_H^3}{3(1-\|d\|_H)}.$$

Substituting this into (3) yields

$$|\phi(y+d,z) - q_y(y+d,z)| \leq \sum_{i=3}^{\infty} |t_i| \leq \frac{\|d\|_H^3}{3(1-\|d\|_H)}.$$

Thus the lemma has been proved. □

LEMMA 4.2. *If the functions $f_i(y)$ satisfy the Relative Lipschitz Condition with Lipschitz constant $M > 0$, and if*

$$y \in \mathcal{F}' \quad \textit{and} \quad \|d\|_H < \min\left\{\frac{1}{2}, \frac{1}{2M^{1/3}}\right\},$$

then $y + d \in \mathcal{F}'$ *and*

$$|\phi(y+d,z) - q_y(y+d,z)| < \frac{\|d\|_H^3}{3(1-\|d\|_H)}(1+2M).$$

Proof. Using Lemma 4.1, one can use the same reasoning as in the proof of Lemma 2.10 of Jarre [5] to obtain the result of the lemma. □

The next lemma simplifies Lemma 2.16 of Jarre [5].

LEMMA 4.3. *If* $\|p\|_H \leq \frac{1}{8(1+2M)}$, *then*

$$\|y - y(z)\|_H \leq \frac{5}{2}\|p\|_H.$$

Proof. Let h be arbitrary, such that $\|h\|_H = \frac{3}{2}\|p\|_H$. We consider the values on the ellipsoid $\{y + p + h \ : \ \|h\|_H = \frac{3}{2}\|p\|_H\}$. We have

$$\|h+p\|_H \leq \|h\|_H + \|p\|_H = \frac{5}{2}\|p\|_H < \frac{1}{3(1+2M)}. \tag{5}$$

With the help of Lemma 4.2, and using the fact that $y + p = \arg\min_x q_y(x,z)$, we obtain

$$\begin{aligned}
\phi(y+p+h,z) &> q_y(y+p+h,z) - \frac{1}{3(1-\frac{1}{3})}\|p+h\|_H^3(1+2M) \\
&\geq q_y(y+p,z) + \frac{1}{2}\|h\|_H^2 - \frac{1}{2}\|p+h\|_H^3(1+2M) \\
&\geq q_y(y+p,z) + \frac{9}{8}\|p\|_H^2 - \frac{125}{16}\|p\|_H^3(1+2M) \\
&\geq q_y(y+p,z) + 9\|p\|_H^3(1+2M) - \frac{125}{16}\|p\|_H^3(1+2M) \\
&> q_y(y+p,z) + \|p\|_H^3(1+2M).
\end{aligned}$$

Using Lemma 4.2 once more, we also obtain that

$$\phi(y+p,z) < q_y(y+p,z) + \frac{1}{2}\|p\|_H^3(1+2M).$$

Hence $\phi(y+p+h,z) > \phi(y+p,z)$. Thus in the center, $y+p$, of the ellipsoid the potential value is less than the value on its boundary. Therefore, by the strict convexity of ϕ, the minimum of ϕ is in the interior of the ellipsoid, which means that $\|y - y(z)\|_H \leq \|p+h\|_H$. Now using (5), the lemma follows. □

LEMMA 4.4. *If* $\|p\|_H \geq \frac{1}{8(1+2M)}$, *then the decrease* $\Delta\phi$ *in the potential function after a linesearch along the Newton direction* p *satisfies*

$$\Delta\phi \geq \frac{1}{140(1+2M)^2}.$$

Proof. Let λ be a steplength such that

$$\|\lambda p\|_H \leq \min\left\{\frac{1}{2}, \frac{1}{2M^{1/3}}\right\}. \tag{6}$$

Then, as a consequence of Lemma 4.2, we have

$$\phi(y+\lambda p, z) \le q_y(y+\lambda p, z) + \frac{\lambda^3\|p\|_H^3}{3(1-\lambda\|p\|_H)}(1+2M).$$

Now using the definition of q_y, we obtain

$$\begin{aligned}\phi(y,z)-\phi(y+\lambda p,z) &\ge -\lambda g^T p - \frac{1}{2}\lambda^2 p^T H p - \frac{\lambda^3\|p\|_H^3}{3(1-\lambda\|p\|_H)}(1+2M)\\ &= \lambda\|p\|_H^2 - \frac{1}{2}\lambda^2\|p\|_H^2 - \frac{\lambda^3\|p\|_H^3}{3(1-\lambda\|p\|_H)}(1+2M).\end{aligned}$$

Replacing λ by the value

$$\frac{1}{9(1+2M)\|p\|_H},$$

which satisfies (6), yields the lemma. □

LEMMA 4.5. *If* $\|p\|_H \le \frac{1}{8(1+2M)}$, *then*

$$\phi(y,z)-\phi(y(z),z) \le 4\|p\|_H^2. \tag{7}$$

Proof. Let d be defined as $y(z)-y$. Using Lemmas 4.2 and 4.3 we get

$$\begin{aligned}\phi(y(z),z) &\ge q_y(y+d,z) - \frac{\|d\|_H^3}{3(1-\|d\|_H)}(1+2M)\\ &= \phi(y,z) - p^T H d + \frac{1}{2}d^T H d - \frac{\|d\|_H^3}{3(1-\|d\|_H)}(1+2M).\end{aligned}$$

Using the Cauchy–Schwarz inequality we may write

$$-p^T H d \ge -\|p\|_H\|d\|_H.$$

Also using Lemma 4.3, we obtain

$$\begin{aligned}\phi(y,z)-\phi(y(z),z) &\le \|p\|_H\|d\|_H - \frac{1}{2}\|d\|_H^2 + \frac{\|d\|_H^3}{3(1-\|d\|_H)}(1+2M)\\ &\le \|p\|_H\|d\|_H + \frac{\|d\|_H^3}{24(1-\|d\|_H)\|p\|_H}\\ &\le \frac{5}{2}\|p\|_H^2 + \frac{\frac{125}{8}}{24(1-\frac{5}{16})}\|p\|_H^2\\ &\le 4\|p\|_H^2.\end{aligned}$$

□

LEMMA 4.6. *Let* $\overline{z}$ *be the new lower bound, i.e.,* $\overline{z} = z + \theta(f_0(y) - z)$, *where* $0 < \theta < 1$, *then*

$$\phi(y,\overline{z}) - \phi(y,z) = -q\ln(1-\theta).$$

Proof. The proof is simple and straightforward. We have

$$f_0(y) - \bar{z} = f_0(y) - z - \theta(f_0(y) - z) = (1-\theta)(f_0(y) - z).$$

Hence

$$\phi(y, \bar{z}) - \phi(y, z) = -q \ln \frac{f_0(y) - \bar{z}}{f_0(y) - z} = -q \ln(1-\theta). \qquad \square$$

LEMMA 4.7. *If* $\|y - y(z)\|_H \leq \beta$, *then*

$$f_0(y(z)) - z \leq \left(1 + \frac{\beta}{\sqrt{q}}\right)(f_0(y) - z).$$

Proof. The lemma is trivial if $f_0(y(z)) \leq f_0(y)$. So let us assume that $f_0(y(z)) > f_0(y)$. By definition we have

$$\begin{aligned}
\beta^2 &\geq \|y - y(z)\|_H^2 \\
&= (y - y(z))^T \left[\sum_{i=1}^{n+q} \left(\frac{\nabla f_i(y) \nabla f_i(y)^T}{f_i(y)^2} + \frac{\nabla^2 f_i(y)}{-f_i(y)}\right)\right](y - y(z)) \\
&\geq (y - y(z))^T q \frac{\nabla f_0(y) \nabla f_0(y)^T}{(f_0(y) - z)^2}(y - y(z)) \\
&\geq q \frac{(f_0(y) - f_0(y(z)))^2}{(f_0(y) - z)^2},
\end{aligned}$$

where the last inequality follows from the convexity of $-f_0(y)$ and the assumption that $f_0(y(z)) > f_0(y)$. Consequently,

$$f_0(y(z)) - f_0(y) \leq \frac{\beta}{\sqrt{q}}(f_0(y) - z).$$

This means that

$$f_0(y(z)) - z \leq \left(1 + \frac{\beta}{\sqrt{q}}\right)(f_0(y) - z). \qquad \square$$

LEMMA 4.8. *If* $\|y - y(z)\|_H \leq \beta$, *then*

$$z^* - z \leq \left(1 + \frac{n}{q}\right)\left(1 + \frac{\beta}{\sqrt{q}}\right)(f_0(y) - z).$$

Proof. The exact center $y(z)$ minimizes the potential function for z. The necessary and sufficient conditions for these minima are (1). From these conditions we derive that $(u(z), y(z))$ is dual-feasible. Moreover, using $z^* \leq f_0(y(z)) - \sum_{i=1}^n u_i(z) f_i(y(z))$, it follows that

$$z^* - f_0(y(z)) \leq -\sum_{i=1}^n u_i(z) f_i(y(z)) = \frac{n}{q}(f_0(y(z)) - z).$$

Consequently,

$$(z^* - z) - (f_0(y(z)) - z) \leq \frac{n}{q}(f_0(y(z)) - z).$$

This means that

$$z^* - z \leq \left(1 + \frac{n}{q}\right)(f_0(y(z)) - z) \leq \left(1 + \frac{n}{q}\right)\left(1 + \frac{\beta}{\sqrt{q}}\right)(f_0(y) - z),$$

where the last inequality follows from Lemma 4.7. This proves the lemma. □

The next lemma generalizes an inequality of Vaidya [10] for the LP-case to the present convex case.

LEMMA 4.9. *The gap $f_0(y(z)) - z$ decreases monotonically if $z < z^*$ increases.*

Proof. We have that $u(z)$ and $y(z)$ satisfy the Karush–Kuhn–Tucker conditions (1). Taking the derivative with respect to z of the last two equations in (1), we obtain

$$\sum_{i=1}^{n} u_i' \nabla f_i(y) + \sum_{i=1}^{n} u_i H_i y' = H_0 y', \tag{8}$$

$$-u_i' f_i(y) - u_i \nabla f_i(y)^T y' = \frac{\nabla f_0(y)^T y' - 1}{q}, \qquad i = 1, \cdots, n, \tag{9}$$

where the prime denotes the derivative with respect to z and H_i denotes the Hessian matrix of $f_i(y)$. The Jacobian of this system of equations is clearly nonsingular for $z < z^*$, and hence, as a consequence of the implicit function theorem, we may conclude that u' and y' exist for $z < z^*$. Multiplying (9) with u_i and using (1), we get

$$\frac{f_0(y) - z}{q} u_i' - u_i^2 \nabla f_i(y)^T y' = \frac{\nabla f_0(y)^T y' - 1}{q} u_i.$$

Multiplying this equation with $\nabla f_i(y)$, summing over i, and using (8) and (1) results in

$$-\frac{f_0(y) - z}{q}\left(\sum_{i=1}^{n} u_i H_i y' - H_0 y'\right) - \sum_{i=1}^{n} u_i^2 \nabla f_i(y)^T y' \nabla f_i(y) = \frac{\nabla f_0(y)^T y' - 1}{q} \nabla f_0(y).$$

Now, taking the inner product with y', we obtain

$$\begin{aligned}\frac{\nabla f_0(y)^T y' - 1}{q} \nabla f_0(y)^T y' = & -\frac{f_0(y) - z}{q} (y')^T \left(\sum_{i=1}^{n} u_i H_i - H_0\right) y' \\ & - \sum_{i=1}^{n} u_i^2 (\nabla f_i(y)^T y')^2 \leq 0.\end{aligned}$$

We conclude that $0 \leq \nabla f_0(y)^T y' \leq 1$, which means that the derivative of $f_0(y(z)) - z$, which is equal to $\nabla f_0(y)^T y' - 1$, is not positive. This proves the lemma. □

5. Convergence analysis. Based on the lemmas in the previous section, we will give upper bounds for the total number of outer iterations and inner iterations.

THEOREM 5.1. *Let $\beta \leq \frac{5}{16(1+2M)}$; then after at most*

$$K = \frac{(1 + \frac{n}{q})(1 + \frac{\beta}{\sqrt{q}})}{\theta} O(|\ln \epsilon|)$$

outer iterations, the algorithm finds an ϵ-optimal solution for (CP).

Proof. Let z^k be the lower bound in the kth outer iteration. Then we have

$$\begin{aligned}\frac{z^*-z^k}{z^*-z^{k-1}} &= \frac{z^*-(z^{k-1}+\theta(f_0(y^{k-1})-z^{k-1}))}{z^*-z^{k-1}}\\ &= 1-\theta\frac{f_0(y^{k-1})-z^{k-1}}{z^*-z^{k-1}}\\ &\le 1-\frac{\theta}{(1+\frac{n}{q})(1+\frac{\beta}{\sqrt{q}})},\end{aligned}$$

where y^{k-1} is the iterate at the end of the $(k-1)$th outer iteration. The last inequality follows from Lemma 4.8. Hence after K outer iterations we have

$$\begin{aligned}z^*-f_0(y^K) &\le z^*-z^{K+1}\\ &\le \left(1-\frac{\theta}{(1+\frac{n}{q})(1+\frac{\beta}{\sqrt{q}})}\right)(z^*-z^K)\\ &\le \left(1-\frac{\theta}{(1+\frac{n}{q})(1+\frac{\beta}{\sqrt{q}})}\right)^K(z^*-z^0).\end{aligned}$$

This means that $z^*-f_0(y^K)\le\epsilon$ certainly holds if

$$\left(1-\frac{\theta}{(1+\frac{n}{q})(1+\frac{\beta}{\sqrt{q}})}\right)^K(z^*-z^0)\le\epsilon.$$

Taking logarithms, this inequality reduces to

$$-K\ln\left(1-\frac{\theta}{(1+\frac{n}{q})(1+\frac{\beta}{\sqrt{q}})}\right)\ge|\ln\epsilon|+\ln(z^*-z^0).$$

Since $-\ln(1-v)>v$, this will certainly hold if

$$K>\frac{(1+\frac{n}{q})(1+\frac{\beta}{\sqrt{q}})}{\theta}(|\ln\epsilon|+\ln(z^*-z^0)).$$

Now using the assumption on z^0, i.e. $z^*-z^0\le\frac{1}{\epsilon}$, the theorem follows. □

From Lemma 4.8 it follows that it suffices to take

$$t=\ln\frac{(1+\frac{n}{q})(1+\frac{\beta}{\sqrt{q}})-1}{\epsilon},$$

i.e., for such t the algorithm ends up with a solution y such that $z^*-f_0(y)\le\epsilon$.

Now we give an upper bound for the total number of inner iterations during an arbitrary outer iteration.

THEOREM 5.2. *The total number P of inner iterations during an arbitrary outer iteration satisfies*

$$P\delta\le 1+\frac{\theta\beta q}{\sqrt{q}+\beta}+\frac{\theta^2 q}{1-\theta},$$

where δ is the guaranteed decrease in each inner iteration, and $\beta \leq \frac{5}{16(1+2M)}$.

Proof. We denote the used lower bound in an arbitrary outer iteration by $\tilde{z}$, while the lower bound in the previous outer iteration is denoted by $\overline{z}$. The iterate at the beginning of the outer iteration is denoted by y. Hence y is centered with respect to $y(\overline{z})$ and $\tilde{z} = \overline{z} + \theta(f_0(y) - \overline{z})$. Because of Lemma 4.4 and definition of $y(\tilde{z})$ we have

$$P\delta \leq \phi(y, \tilde{z}) - \phi(y(\tilde{z}), \tilde{z}). \tag{10}$$

Let us call the right-hand side of (10) $\Phi(y, \tilde{z})$. According to the mean value theorem there is a $\hat{z} \in (\overline{z}, \tilde{z})$ such that

$$\Phi(y, \tilde{z}) = \Phi(y, \overline{z}) + \left.\frac{d\,\Phi(y,z)}{d\,z}\right|_{z=\hat{z}} (\tilde{z} - \overline{z}). \tag{11}$$

Let us now look at $\frac{d\,\Phi(y,z)}{d\,z}$. We have

$$\frac{d\,\phi(y,z)}{d\,z} = \frac{q}{f_0(y) - z}$$

and

$$\begin{aligned}
\frac{d\,\phi(y(z),z)}{d\,z} &= -q\frac{\nabla f_0(y(z))^T y' - 1}{f_0(y(z)) - z} + \sum_{i=1}^{n} \frac{\nabla f_i(y(z))^T y'}{-f_i(y(z))} \\
&= -q\frac{\nabla f_0(y(z))^T y' - 1}{f_0(y(z)) - z} + \frac{q}{f_0(y(z)) - z}\sum_{i=1}^{n} u_i(z) \nabla f_i(y(z))^T y' \\
&= \frac{q}{f_0(y(z)) - z},
\end{aligned}$$

where the two last equations follow from (1). So

$$\left.\frac{d\,\Phi(y,z)}{d\,z}\right|_{z=\hat{z}} = q\left.\left(\frac{1}{f_0(y) - z} - \frac{1}{f_0(y(z)) - z}\right)\right|_{z=\hat{z}} \leq q\left(\frac{1}{f_0(y) - \tilde{z}} - \frac{1}{f_0(y(\overline{z})) - \overline{z}}\right),$$

where the last inequality follows from the fact that $\tilde{z} > \hat{z}$ and from Lemma 4.9. Substituting this into (11) gives

$$\begin{aligned}
\Phi(y, \tilde{z}) &\leq \Phi(y, \overline{z}) + q\left(\frac{1}{f_0(y) - \tilde{z}} - \frac{1}{f_0(y(\overline{z})) - \overline{z}}\right)(\tilde{z} - \overline{z}) \\
&= \Phi(y, \overline{z}) + q\theta\left(\frac{1}{1-\theta} - \frac{f_0(y) - \overline{z}}{f_0(y(\overline{z})) - \overline{z}}\right) \\
&\leq 1 + q\theta\left(\frac{1}{1-\theta} - \frac{1}{1 + \frac{\beta}{\sqrt{q}}}\right) \qquad (12) \\
&= 1 + q\theta\left(\frac{\theta}{1-\theta} + \frac{\beta}{\sqrt{q} + \beta}\right),
\end{aligned}$$

where inequality (12) follows because $\Phi(y, \overline{z}) \leq 1$ according to Lemma 4.5, and $f_0(y(\overline{z})) - \overline{z} \leq (1 + (\beta/\sqrt{q}))(f_0(y) - \overline{z})$ according to Lemma 4.7. □

From Theorem 5.1 we know that the total number of outer iterations is at most

$$\frac{(1 + \frac{n}{q})(1 + \frac{\beta}{\sqrt{q}})}{\theta} O(|\ln \epsilon|).$$

Hence the total number of inner iterations during the whole process is given by

$$\frac{1}{\delta}\left(1+\frac{n}{q}\right)\left(1+\frac{\beta}{\sqrt{q}}\right)\left(\frac{1}{\theta}+\frac{\beta q}{\sqrt{q}+\beta}+\frac{\theta q}{1-\theta}\right)O(|\ln \epsilon|). \tag{13}$$

This makes clear that if we take $q = n$, then

- if we take $\theta = \nu/\sqrt{n}$, $\nu > 0$ and independent of n, M, and ϵ, then the algorithm has an $O((1+M)^2\sqrt{n}|\ln \epsilon|)$ iteration bound.
- if we take $0 < \theta < 1$, independent of n, M, and ϵ, then the algorithm has an $O((1+M)^2 n|\ln \epsilon|)$ iteration bound.

The first case corresponds to a small reduction factor θ. In this case we can return to the vicinity of the central trajectory in $O((1+M)^2)$ steps, while the lower bound must be updated $O(\sqrt{n}|\ln \epsilon|)$ times. In the path-following algorithm of Jarre [5], the same iteration bound is obtained for

$$\theta = \frac{1}{200\sqrt{n}(1+M^2)}.$$

The second case corresponds to a large reduction factor θ. In this case we can return to the vicinity of the central trajectory in $O((1+M)^2 n)$ linesearches, while the lower bound must be updated $O(|\ln \epsilon|)$ times.

Remark 1. We note that the upper bound for the number of iterations is not better than Jarre's. However, while Jarre's bound is more or less exact, our bound can be very pessimistic, because of the linesearches involved in the inner iterations. This can also be one of the reasons for the fact that a large reduction factor gives a worse bound than a small reduction factor, while one would expect the contrary.

Remark 2. The "centering assumption" $\|p(y^0,z^0)\|_{H(y^0,z^0)} \leq \tau$ can be alleviated to

$$\phi(y^0,z^0) - \phi(y(z^0),z^0) \leq O(\sqrt{n}|\ln \epsilon|)$$

for the first case, and to

$$\phi(y^0,z^0) - \phi(y(z^0),z^0) \leq O(n|\ln \epsilon|)$$

for the second case. This follows easily from Lemma 4.4.

Remark 3. Note that the updating factor θ is independent from M, contrary to Jarre's [5] method.

Remark 4. For linear programming problems, i.e., $M = 0$, we can find an exact solution if we take $\epsilon = 2^{-L}$, where L denotes the input length of the problem. In this case our results reduce to an $O(\sqrt{n}L)$ iteration bound if we take $\theta = \nu/\sqrt{n}$, $\nu > 0$ and independent of n, M, and ϵ; and to an $O(nL)$ iteration bound if we take $0 < \theta < 1$, independent of n, M, and ϵ. These results are also obtained by Den Hertog, Roos, and Terlaky [1]; Gonzaga [3]; and Roos and Vial [9].

Appendix: Proof of the inequalities (4). Since each function f_i is assumed to be linear or quadratic, the kth-order term in its Taylor expansion has the following form:

$$t_k = \frac{1}{k!}\sum_{j=1}^{n+q}\sum_{i=0}^{\lfloor k/2 \rfloor} a_{k,i}\frac{(\nabla f_j(y)^T d)^{k-2i}(d^T\nabla^2 f_j(y)d)^i}{(-f_j(y))^{k-i}}, \qquad k \geq 1, \tag{14}$$

where $a_{k,i}$ has to be determined yet. For shortness' sake we use the following notations:

$$\chi_j := \frac{\nabla f_j(y)^T d}{-f_j(y)},$$
$$\psi_j := \frac{d^T \nabla^2 f_j(y) d}{-f_j(y)},$$
$$D^2 := \|d\|_H^2.$$

Using these notations (14) becomes

$$t_k = \frac{1}{k!} \sum_{j=1}^{n+q} \sum_{i=0}^{\lfloor k/2 \rfloor} a_{k,i} \chi_j^{k-2i} \psi_j^i. \tag{15}$$

We also have

$$D^2 = \sum_{j=1}^{n+q} (\chi_j^2 + \psi_j).$$

Now we derive a recursive formula for $a_{k,i}$. Using the chain rule for taking derivatives, we obtain from (14) an expression for t_{k+1}:

$$t_{k+1} = \frac{1}{(k+1)!} \sum_{j=1}^{n+q} \left[\sum_{i=0}^{\lfloor k/2 \rfloor} (k-i) a_{k,i} \chi_j^{k-2i+1} \psi_j^i + \sum_{i=0}^{\lfloor k/2 \rfloor} (k-2i) a_{k,i} \chi_j^{k-2i-1} \psi_j^{i+1} \right].$$

This can be rewritten as

$$t_{k+1} = \frac{1}{(k+1)!} \sum_{j=1}^{n+q} \left[\sum_{i=0}^{\lfloor k/2 \rfloor} (k-i) a_{k,i} \chi_j^{k-2i+1} \psi_j^i \right. \\ \left. + \sum_{i=1}^{\lfloor k/2 \rfloor + 1} (k-2i+2) a_{k,i-1} \chi_j^{k-2i+1} \psi_j^i \right].$$

From this the following recursive formula can be derived:

$$\begin{aligned} &a_{1,0} = 1, \quad a_{1,i} = 0, \quad i \neq 0, \\ &a_{k+1,i} = (k-i) a_{k,i} + (k-2i+2) a_{k,i-1}, \qquad k \geq 1. \end{aligned} \tag{16}$$

From this recursive scheme we derive an explicit formula for $a_{k,i}$:

$$a_{k,i} = \begin{cases} \frac{k!(k-i-1)!}{i!(k-2i)!2^i} & \text{if } 0 \leq i \leq \lfloor \frac{k}{2} \rfloor, \\ 0 & \text{otherwise.} \end{cases} \tag{17}$$

We prove this formula by using induction. For $k = 1, 2$ our formula is certainly correct, as follows by inspection. From (16) it readily follows that $a_{k,0} = (k-1)!$, $k \geq 1$. This is in accordance with (17), since

$$(k-1)! = \frac{k!(k-1)!}{0!k!2^0}.$$

Now suppose that the formula is correct for some value of k, $k \geq 1$. Then, using (16) and (17) one has, for $i \geq 1$,

$$\begin{aligned} a_{k+1,i} &= (k-i)a_{k,i} + (k-2i+2)a_{k,i-1} \\ &= (k-i)\frac{k!(k-i-1)!}{i!(k-2i)!2^i} + (k-2i+2)\frac{k!(k-i)!}{(i-1)!(k-2i+2)!2^{i-1}} \\ &= \frac{(k+1)!(k-i)!}{i!(k-2i+1)!2^i}. \end{aligned}$$

This proves that formula (17) is correct indeed.

We proceed by deriving an upper bound for t_k. To this end we consider the following optimization problem:

$$\max \left(\frac{1}{k!} \sum_{j=1}^{n+q} \sum_{i=0}^{\lfloor k/2 \rfloor} a_{k,i} \chi_j^{k-2i} \psi_j^i \; : \; \sum_{j=1}^{n+q} (\chi_j^2 + \psi_j) = D^2, \; \chi_j \geq 0, \; \psi_j \geq 0, \; 1 \leq j \leq n \right),$$

where the maximization is done over χ_j and ψ_j. The nonnegativity of ψ_j is an obvious consequence of its definition; the nonnegativity of χ_j can be assumed, since the constraints are sign independent as far as χ_j is concerned, whereas positive values will give a larger objective than negative values. The Kuhn–Tucker optimality conditions for this problem are given by

$$\sum_{i=0}^{\lfloor k/2 \rfloor - 1} (k-2i) a_{k,i} \chi_j^{k-2i-1} \psi_j^i + \left(\left\lfloor \frac{k+1}{2} \right\rfloor - \left\lfloor \frac{k}{2} \right\rfloor \right) a_{k,\lfloor k/2 \rfloor} \psi^{\lfloor k/2 \rfloor} \leq 2\alpha \chi_j, \tag{18}$$

$$\sum_{i=1}^{\lfloor k/2 \rfloor} i a_{k,i} \chi_j^{k-2i} \psi_j^{i-1} \leq \alpha, \tag{19}$$

$$\begin{aligned} \chi_j \Bigg(& \sum_{i=0}^{\lfloor k/2 \rfloor - 1} (k-2i) a_{k,i} \chi_j^{k-2i-1} \psi_j^i \\ & + \left(\left\lfloor \frac{k+1}{2} \right\rfloor - \left\lfloor \frac{k}{2} \right\rfloor \right) a_{k,\lfloor k/2 \rfloor} \psi^{\lfloor k/2 \rfloor} - 2\alpha \chi_j \Bigg) = 0, \end{aligned} \tag{20}$$

$$\psi_j \left(\sum_{i=1}^{\lfloor k/2 \rfloor} i a_{k,i} \chi_j^{k-2i} \psi_j^{i-1} - \alpha \right) = 0, \tag{21}$$

where α is the multiplier.

From these conditions we will derive that either ψ_j or χ_j must be zero for each j. Assume that neither ψ_j nor χ_j is zero. From this we shall derive a contradiction. By multiplying (20) by ψ_j and (21) by $2\chi_j^2$ and subtracting, we derive

$$\begin{aligned} & \sum_{i=0}^{\lfloor k/2 \rfloor - 1} \left[(k-2i) a_{k,i} - 2(i+1) a_{k,i+1} \right] \chi_j^{k-2i} \psi_j^{i+1} \\ & + \left(\left\lfloor \frac{k+1}{2} \right\rfloor - \left\lfloor \frac{k}{2} \right\rfloor \right) a_{k,\lfloor k/2 \rfloor} \chi_j \psi^{\lfloor k/2 \rfloor + 1} = 0. \end{aligned}$$

It is easy to see that $\lfloor\frac{k+1}{2}\rfloor - \lfloor\frac{k}{2}\rfloor$ is 1 if k is odd, and 0 if k is even. Furthermore, let

$$\alpha_{k,i} := (k-2i)a_{k,i} - 2(i+1)a_{k,i+1}.$$

We easily obtain that $\alpha_{k,0} = 0$ and $\alpha_{k,i} > 0$, for $i \geq 1$, by using the general formula (17) for $a_{k,i}$:

$$\begin{aligned}\alpha_{k,i} &= (k-2i)\frac{k!(k-i-1)!}{i!(k-2i)!2^i} - 2(i+1)\frac{k!(k-i-2)!}{(i+1)!(k-2i-2)!2^{i+1}}\\ &= \frac{k!(k-i-2)!}{i!(k-2i-2)!2^i}\left[\frac{k-i-1}{k-2i-1} - 1\right]\\ &> 0.\end{aligned}$$

Hence, we have obtained a contradiction. This means that either χ_j or ψ_j is zero in the maximum. Consequently, the objective function either consists of "pure" χ_j terms or "pure" ψ_j terms.

Now if k is even, it easily follows that the largest objective value is obtained if $\psi_j = 0$, since, using the general formula (17), we have

$$a_{k,0} = 2^{(k/2)-1}a_{k,k/2},$$

which means that the coefficient of the pure χ_j term is greater than the coefficient of the pure ψ_j term. Now assume that $\psi_j > 0$ if k is odd. Then (19) holds with equality. By multiplying (19) by $2\chi_j$ and subtracting from (18), we derive

$$(22)\qquad \sum_{i=0}^{\lfloor k/2\rfloor-1}\left[(k-2i)a_{k,i} - 2(i+1)a_{k,i+1}\right]\chi_j^{k-2i-1}\psi_j^i + a_{k,\lfloor k/2\rfloor}\psi^{\lfloor k/2\rfloor} = 0.$$

Again, we have obtained a contradiction, since we assumed that $\psi_j > 0$. Consequently, both for k even and k odd, the maximum is reached if $\psi_j = 0$. So an upper bound for the maximum of the function is

$$\frac{1}{k!}\sum_{j=1}^{n+q}(k-1)!\chi_j^k = \frac{1}{k}\sum_{j=1}^{n+q}\chi_j^k \leq \frac{1}{k}\left(\sum_{j=1}^{n+q}\chi_j^2\right)^{k/2} = \frac{1}{k}(D^2)^{k/2} = \frac{1}{k}D^k.$$

Hence the proof of (4) is complete.

REFERENCES

[1] D. Den Hertog, C. Roos, and T. Terlaky (1991), *A potential reduction variant of Renegar's short-step path-following method for linear programming*, Linear Algebra Appl., 152, pp. 43–68.

[2] R. M. Freund (1988), *Polynomial-time algorithms for linear programming based only on primal scaling and projected gradients of a potential function*, Working Paper OR 182–88, Sloan School of Management, Massachusetts Institute of Technology, Cambridge, MA.

[3] C. C. Gonzaga (1989), *Large-steps path-following methods for linear programming, Part* I: *Barrier function method*, Report ES–210/89, Department of Systems Engineering and Computer Sciences, COPPE–Federal University of Rio de Janeiro, Rio de Janeiro, Brasil; also in SIAM J. Optimization, 1 (1991), pp. 268–279.

[4] ——— (1989), *Large-step path-following methods for linear programming, Part* II: *Potential reduction method*, Report ES–211/89, Department of Systems Engineering and Computer Sciences, COPPE–Federal University of Rio de Janeiro, Rio de Janeiro, Brasil; also in SIAM J. Optimization, 1 (1991), pp. 280–292.

[5] F. JARRE (1989), *The method of analytic centers for smooth convex programs*, Ph.D. thesis, Institut für Angewandte Mathematik und Statistik, Universität Würzburg, Würzburg, Germany.

[6] N. KARMARKAR (1984), *A new polynomial-time algorithm for linear programming*, Combinatorica, 4, pp. 373–395.

[7] M. KOJIMA, S. MIZUNO, AND A. YOSHISE (1991), *An $O(\sqrt{n}L)$ iteration potential reduction algorithm for linear complementarity problems*, Math. Programming, 50, pp. 331–342.

[8] S. MEHROTRA AND J. SUN (1988), *An interior point algorithm for solving smooth convex programs based on Newton's method*, Tech. Report 88–08, Department of Industrial Engineering and Management Science, Northwestern University, Evaston, IL.

[9] C. ROOS AND J.-PH. VIAL (1990), *Long steps with the logarithmic penalty barrier function in linear programming*, in Economic Decision-Making: Games, Economics and Optimization, dedicated to Jacques H. Drèze; J. Gabszevwicz, J.-F. Richard, and L. Wolsey, eds., Elsevier Science Publisher B.V., Amsterdam, pp. 433–441.

[10] P. M. VAIDYA (1990), *An algorithm for linear programming which requires $O(((m+n)n^2 + (m+n)^{1.5}n)L)$ arithmetic operations*, Math. Programming, 47, pp. 175–201.

[11] PH. WOLFE (1961), *A duality theorem for nonlinear programming*, Quart. Appl. Math., 19, pp. 239–244.

[12] Y. YE (1991), *An $O(n^3L)$ potential reduction algorithm for linear programming*, Math. Programming, 50, pp. 239–258.

SIAM J. OPTIMIZATION
Vol. 2, No. 1, pp. 71–87, February 1992

A COMPLEXITY REDUCTION FOR THE LONG-STEP PATH-FOLLOWING ALGORITHM FOR LINEAR PROGRAMMING*

D. DEN HERTOG†, C. ROOS†, AND J.-PH. VIAL‡

Abstract. A modification of previously published long-step path-following algorithms for the solution of the linear programming problem is presented. This modification uses the simple Goldstein–Armijo rule. A $\sqrt{n}$ reduction in the complexity bound is obtained, while a linesearch may still be done. Depending on the updating scheme for the barrier parameter, the resulting complexity bounds are $O(n^3L)$ or $O(n^{3.5}L)$.

Key words. linear programming, interior point method, logarithmic barrier function, polynomial algorithm, Goldstein–Armijo rule

AMS(MOS) subject classification. 90C05

1. Introduction. The original $O(n^{3.5}L)$ complexity bound of short-step path-following methods was reduced to $O(n^3L)$ by Vaidya [15]; Gonzaga [6]; Kojima, Mizuno, and Yoshise [9]; and Monteiro and Adler [11]. This reduction was achieved by using Karmarkar's [8] partial updating scheme. Their partial updating analysis is based on steps of a fixed, short length, which fits into short-step methods in a natural way. In Mizuno and Todd [10] a partial updating analysis for an "adaptive-step" path-following algorithm is given.

In Roos and Vial [14] a long-step path-following algorithm is proposed, which is in fact a natural implementation of the classical logarithmic barrier function approach. The number of reductions of the barrier parameter is $\mathrm{O}(L)$. Each reduction is followed by a series of inner steps, aiming at getting close to the analytic center associated with the current value of the penalty parameter. It was proved that at most $\mathrm{O}(nL)$ inner steps are needed. This means that the total complexity is $O(n^4L)$.

This result was also obtained independently by Gonzaga [7] in a more general approach. He also showed that if the barrier parameter is reduced by a factor $1-(\nu/\sqrt{n})$, $\nu > 0$, then at most $O(\sqrt{n}L)$ reductions and at most $O(1)$ inner steps are needed. So, the total complexity of this variant is $O(n^{3.5}L)$.

In this paper we show that, using a Goldstein–Armijo rule to safeguard the linesearches of the barrier function, a $\sqrt{n}$ reduction in the complexity bounds can be obtained for both versions. As mentioned above, the partial updating analysis in [15], [6], [9], and [11] is based on steps of a short, fixed length, and so it cannot be used in long-step algorithms. The Goldstein–Armijo rule was introduced in the complexity analysis for Karmarkar's [8] projective algorithm by Anstreicher [1]. Anstreicher and Bosch [2] used the rule to improve the complexity bound for Ye [16] and Freund's [4] affine potential reduction algorithm.

Some new aspects are used in the analysis. We will use quadratic convergence in the neighbourhood of the central path to prove some properties of nearly centered points. This also enables us to improve Gonzaga's [7] results. Also, the reduction in

* Received by the editors July 16, 1990; accepted for publication (in revised form) March 27, 1991. This paper was first presented at the Second Asilomar Workshop on Progress in Mathematical Programming, Yale University, New Haven, Connecticut 06520.

† Department of Mathematics and Computer Science, Delft University of Technology, P.O. Box 5031, 2600 GA Delft, Netherlands.

‡ Department of Commercial and Industrial Economics, University of Geneva, 2 rue de Candolle, CH-1211 Geneva 4, Switzerland. This author's research was partially supported by the Fonds National Suisse de la Recherche Scientifique, under research grant # 1 467 0 86.

the barrier function value after an inner step is proved in a more natural way by using a Taylor expansion.

The paper is organized as follows. In §2 we prove some properties of (nearly) centered points. Then, in §3 we describe our algorithm and in §4 we prove that the algorithm reduces the complexity bound by a factor $\sqrt{n}$.

Notation. As far as notations are concerned, e shall denote the vector of all ones. Given an n-dimensional vector x we denote by X the $n \times n$ diagonal matrix whose diagonal entries are the coordinates x_j of x; x^T is the transpose of the vector x and the same notation holds for matrices. Finally $\|x\|$ denotes the l_2 norm.

2. Properties near the central path. We consider the linear programming problem:

$$\text{(P)} \qquad \min\left\{c^T x : Ax = b, x \geq 0\right\}.$$

Here A is an $m \times n$ matrix and b and c are m- and n-dimensional vectors, respectively; the n-dimensional vector x is the variable in which the minimization is done. The dual formulation for (P) is:

$$\text{(D)} \qquad \max\left\{b^T y : A^T y + s = c, s \geq 0\right\}.$$

Without loss of generality, we assume that all the coefficients are integers. We shall denote by L the length of the input data of (P).

We make the standard assumption that the feasible set of (P) is bounded and has a nonempty relative interior. In order to simplify the analysis we shall also assume that A has full rank, though this assumption is not essential.

We consider the primal logarithmic barrier

$$f(x, \mu) := \frac{c^T x}{\mu} - \sum_{j=1}^{n} \ln x_j, \tag{1}$$

where μ is a positive parameter. The first- and second-order derivatives of f are

$$\nabla f(x, \mu) = \frac{c}{\mu} - X^{-1} e,$$
$$\nabla^2 f(x, \mu) = X^{-2}.$$

Consequently, f is strictly convex on the relative interior of the feasible set. It also takes infinite values on the boundary of the feasible set. Thus it achieves a minimum value at a unique point. The necessary and sufficient first-order optimality conditions for this point are:

$$\begin{aligned} A^T y + s &= c, \qquad & s &\geq 0, \\ Ax &= b, \qquad & x &\geq 0, \\ Xs &= \mu e, \end{aligned} \tag{2}$$

where y and s are m- and n-dimensional vectors, respectively. It is well known that the necessary and sufficient first-order optimality conditions for the minimum of the dual logarithmic barrier function are also (2).

Let us denote the unique solution of this system by $(x(\mu), y(\mu), s(\mu))$. The primal and dual central path is defined as the solution set $x(\mu)$ and $y(\mu)$, respectively, for

$\mu > 0$. It is well known that the duality gap in $(x(\mu), y(\mu), s(\mu))$ satisfies $x(\mu)^T s(\mu) = n\mu$. Hence, if $\mu \to 0$, then $x(\mu)$ and $y(\mu)$ will converge to optimal primal and dual solutions, respectively.

The following lemma states that the primal objective decreases along the primal path and the dual objective increases along the dual path. These results also follow from Fiacco and McCormick [3]. We will give another simple proof.

LEMMA 2.1. *The objective $c^T x(\mu)$ of the primal problem* (P) *is monotonically decreasing and the objective $b^T y(\mu)$ of the dual problem* (D) *is monotonically increasing if μ decreases.*

Proof. Using the fact that $x(\mu)$ and $y(\mu)$ satisfy (2) and taking derivatives with respect to μ we obtain

$$\begin{aligned} A^T y' + s' &= 0, \\ Ax' &= 0, \\ Xs' + Sx' &= e, \end{aligned} \tag{3}$$

where the prime denotes the derivative with respect to μ. Now, using the relations of (2) and (3), we find

$$\begin{aligned} c^T x' &= (x')^T(s + A^T y) = (x')^T s = e^T(Sx') = (Xs' + Sx')^T Sx' \\ &= \mu(x')^T s' + (x')^T S^2 x' = (x')^T S^2 x' \geq 0, \end{aligned}$$

where the last equality follows because $(x')^T s' = -(Ax')^T y' = 0$. This proves the first part of the lemma.

To prove the second part of the lemma, we multiply the last equality of (3) by AS^{-1}:

$$AS^{-1}Xs' + Ax' = AS^{-1}e,$$

which reduces to $AX^2 s' = b$. Taking the inner product with y' results in

$$b^T y' = (y')^T AX^2 s' = (A^T y')^T X^2 s' = -(s')^T X^2 s' \leq 0.$$

This proves the second part of the lemma. □

Roos and Vial [13] introduced the following measure of the distance of an interior feasible point to the central path:

$$\delta(x, \mu) := \min_{y,s} \left\{ \left\| \frac{Xs}{\mu} - e \right\| : A^T y + s = c \right\}. \tag{4}$$

The unique solution of the minimization problem in the definition of $\delta(x, \mu)$ is denoted by $(y(x, \mu), s(x, \mu))$. It can easily be verified that

$$x = x(\mu) \Longleftrightarrow \delta(x, \mu) = 0 \Longrightarrow s(x, \mu) = s(\mu).$$

The next lemma states that there is a close relationship between this measure and the projected Newton direction $p(x, \mu)$, which is obtained from (cf., e.g., [5])

$$\begin{pmatrix} X^{-2} & A^T \\ A & 0 \end{pmatrix} \begin{pmatrix} -p \\ \frac{y}{\mu} \end{pmatrix} = \begin{pmatrix} \frac{c}{\mu} - X^{-1}e \\ 0 \end{pmatrix}. \tag{5}$$

LEMMA 2.2. *For given x and μ, $\delta(x,\mu) = \|X^{-1}p(x,\mu)\|$.*

Proof. From (5) we can derive that $p(x,\mu) = Xq$, where

$$q = e - \frac{Xs}{\mu} \tag{6}$$

with

$$s = c - A^T y \tag{7}$$

and

$$y = (AX^2A^T)^{-1}AX(Xc - \mu e). \tag{8}$$

It can easily be verified that $s = s(x,\mu)$. Thus the lemma is proved. □

We note that a closed-form solution for $p(x,\mu)$ is given by

$$p(x,\mu) = -XP_{AX}\left(\frac{Xc}{\mu} - e\right), \tag{9}$$

where P_{AX} denotes the orthogonal projection on the null space of the matrix AX. Consequently, the projected Newton direction and the scaled projected gradient direction associated with f coincide. In the following we will write p instead of $p(x,\mu)$.

Now we will prove some fundamental lemmas for nearly centered points.

LEMMA 2.3. *If $\delta := \delta(x,\mu) \le 1$, then $y := y(x,\mu)$ is dual-feasible. Moreover,*

$$\mu(n - \delta\sqrt{n}) \le c^T x - b^T y \le \mu(n + \delta\sqrt{n}).$$

Proof. By the definition of $s(x,\mu)$ we have

$$\left\|\frac{Xs(x,\mu)}{\mu} - e\right\| \le 1.$$

This implies $s(x,\mu) \ge 0$, so $y(x,\mu)$ is dual-feasible. Moreover,

$$\left|\frac{x^T s(x,\mu)}{\mu} - n\right| = \left|e^T\left(\frac{Xs(x,\mu)}{\mu} - e\right)\right| \le \|e\| \left\|\frac{Xs(x,\mu)}{\mu} - e\right\| = \delta\sqrt{n}.$$

Consequently, since $x^T s(x,\mu) = c^T x - b^T y$,

$$\mu(n - \delta\sqrt{n}) \le c^T x - b^T y \le \mu(n + \delta\sqrt{n}).$$ □

LEMMA 2.4. *If $\delta(x,\mu) < 1$, then $x^* = x + p$ is a strictly feasible point for* (P). *Moreover,*

$$\delta(x^*,\mu) \le \delta(x,\mu)^2.$$

Proof. In the proof we make use of the vector t defined by

$$t = \frac{Xs(x,\mu)}{\mu}.$$

Note that

$$x^* = x + p = x + X(e - t) = 2x - Xt.$$

From $\delta(x, \mu) < 1$ we deduce that $\|t - e\| < 1$. Hence

$$2e - t > 0.$$

As a consequence one has, since $x > 0$,

$$x^* = 2x - Xt = X(2e - t) > 0.$$

So x^* is strictly feasible, because $Ax^* = Ax + Ap = b$.

The definition of $s(x^*, \mu)$ implies the following:

$$\delta(x^*, \mu) = \left\| \frac{X^* s(x^*, \mu)}{\mu} - e \right\| \leq \left\| \frac{X^* s(x, \mu)}{\mu} - e \right\| = \|X^* X^{-1} t - e\|.$$

Using that $x^* = 2x - Xt$ we find

$$X^* X^{-1} t - e \;= (2X - XT) X^{-1} t - e = 2t - Tt - e = (E - T)(t - e).$$

Hence

$$\delta(x^*, \mu) \leq \max_i |1 - t_i| \|t - e\| \leq \delta(x, \mu)^2. \qquad \square$$

LEMMA 2.5. *If* $\delta := \delta(x, \mu) < 1$, *then*

$$f(x, \mu) - f(x(\mu), \mu) \leq \frac{\delta^2}{1 - \delta^2}.$$

Proof. The barrier function f is convex in x, whence

$$f(x + p, \mu) \geq f(x, \mu) + p^T \nabla f(x, \mu).$$

Now using (9) and $AXX^{-1}p = Ap = 0$,

$$\begin{aligned} p^T \nabla f(x, \mu) &= (X^{-1}p)^T X \nabla f(x, \mu) \\ &= (X^{-1}p)^T P_{AX} (X \nabla f(x, \mu)) \\ &= -(X^{-1}p)^T X^{-1} p \\ &= -\delta^2, \end{aligned} \tag{10}$$

where the last equality follows from Lemma 2.2. Substitution gives

$$f(x + p, \mu) \geq f(x, \mu) - \delta^2,$$

or equivalently,

$$f(x, \mu) - f(x + p, \mu) \leq \delta^2. \tag{11}$$

Now let $x^0 := x$ and let $x^0, x^1, x^2, \cdots$ denote the sequence of points obtained by repeating Newton steps, starting at x^0. By Lemma 2.4 we have

$$\delta(x^i, \mu) \leq \delta(x^0, \mu)^{2^i} = \delta^{2^i}. \tag{12}$$

So, using (11), we may write

$$\begin{aligned} f(x,\mu) - f(x(\mu),\mu) &= \sum_{i=0}^{\infty} \left(f(x^i,\mu) - f(x^{i+1},\mu)\right) \\ &\leq \sum_{i=0}^{\infty} \delta(x^i,\mu)^2 \\ &\leq \sum_{i=0}^{\infty} \delta^{2^{i+1}} \\ &\leq \frac{\delta^2}{1-\delta^2}. \end{aligned}$$

□

LEMMA 2.6. *If* $\delta := \delta(x,\mu) < 1$, *then*

$$|c^Tx - c^Tx(\mu)| \leq \frac{\delta(1+\delta)}{1-\delta}\mu\sqrt{n}.$$

Proof. From (10) we have $p^T\nabla f(x,\mu) = -\delta^2$. On the other hand,

$$\begin{aligned} p^T\nabla f(x,\mu) &= p^T\left(\frac{c}{\mu} - X^{-1}e\right) \\ &= \frac{c^Tp}{\mu} - e^TX^{-1}p. \end{aligned}$$

So we have

$$\frac{c^Tp}{\mu} = -\delta^2 + e^TX^{-1}p$$

or

$$c^Tp = \mu(-\delta^2 + e^TX^{-1}p).$$

Using the Cauchy–Schwarz inequality, we obtain

$$|e^TX^{-1}p| \leq \|X^{-1}p\|\|e\| = \delta\sqrt{n},$$

where the last equality follows from Lemma 2.2. From this we deduce that

$$(13) \qquad |c^Tp| \leq \mu(\delta^2 + \delta\sqrt{n}) = \delta\left(1 + \frac{\delta}{\sqrt{n}}\right)\mu\sqrt{n} \leq \delta(1+\delta)\,\mu\sqrt{n}.$$

Again, let $x^0 := x$ and let $x^0, x^1, x^2, \cdots$ denote the sequence of points obtained by repeating Newton steps, starting at x^0. Then we have

$$\begin{aligned} |c^Tx - c^Tx(\mu)| &= \left|\sum_{i=0}^{\infty} \left(c^Tx^i - c^Tx^{i+1}\right)\right| \\ &\leq \sum_{i=0}^{\infty} \left|c^Tp(x^i,\mu)\right| \end{aligned}$$

$$\begin{aligned}
&\leq \sum_{i=0}^{\infty} \delta(x^i, \mu)(1 + \delta(x^i, \mu))\mu\sqrt{n} \\
&\leq \sum_{i=0}^{\infty} \delta^{2^i}(1 + \delta^{2^i})\mu\sqrt{n} \\
&\leq (1 + \delta)\mu\sqrt{n}\sum_{i=0}^{\infty} \delta^{2^i} \\
&\leq \frac{\delta(1 + \delta)}{1 - \delta}\mu\sqrt{n},
\end{aligned}$$

where the second inequality follows from (13) and the third inequality from (12). □

In [7] results similar to those in Lemmas 2.5 and 2.6 have been obtained in a different way for more centered x, namely, $\delta(x, \mu) \leq 0.1$. Our results hold for $\delta(x, \mu) < 1$. Moreover, for $\delta(x, \mu) \leq 0.1$, our bounds are tighter.

3. The revised long-step algorithm. Long-step barrier methods work as follows: fix μ, do linesearches along Newton directions until the iterate is in the neighbourhood of the current center, then reduce the barrier parameter, and repeat this process. Hence, at each iteration of these methods, one has to solve the linear system (5). Essentially this means that the $(m+n) \times (m+n)$ coefficient matrix of this system, denoted M,

$$\begin{pmatrix} X^{-2} & A^T \\ A & 0 \end{pmatrix}$$

has to be inverted. Hence, assuming that $m = O(n)$, at each iteration $O(n^3)$ arithmetic operations are needed. The matrices in two successive iterations differ only due to changes in X. Now consider the hypothetical case when only one entry of x changes. Then the new coefficient matrix M' differs from M only by a rank-one matrix. This makes it clear that we can write

$$M' = M + uv^T,$$

where u and v are suitable vectors. With the help of the Sherman–Morrison formula,

$$(M + uv^T)^{-1} = M^{-1} - \frac{M^{-1}uv^TM^{-1}}{1 + v^TM^{-1}u},$$

the inverse of M' can be calculated from the inverse of M in only $O(n^2)$ arithmetic operations. If we require an exact solution of the system of equations we will generally need to make n such rank-one modifications. Therefore, $O(n^3)$ arithmetic operations will be needed at each iteration.

However, assume that instead of solving system (5) we solve

$$\begin{pmatrix} \tilde{X}^{-2} & A^T \\ A & 0 \end{pmatrix}\begin{pmatrix} -\tilde{p} \\ \frac{\tilde{y}}{\mu} \end{pmatrix} = \begin{pmatrix} \frac{c}{\mu} - X^{-1}e \\ 0 \end{pmatrix}, \quad \tilde{y} \in \mathbb{R}^m, \tag{14}$$

where $\tilde{X}$ is a working matrix closely related to X. Actually, the diagonal term $\tilde{x}_j$ of $\tilde{X}$ is updated during the inner iteration only if $\tilde{x}_j$ differs too much from x_j. If a limited number of components of $\tilde{x}$ are updated at a given iteration, a reduced computational cost can be achieved using the Sherman–Morrison formula. Of course one does not obtain the exact projected Newton direction p, but an approximation $\tilde{p}$ of it.

The purpose of this paper is to show that by performing a safeguarded linesearch along $\tilde{p}$, one can achieve the double goal of enforcing a significant decrease of the barrier function at each iteration, while maintaining a relatively small number of updates in the components of $\tilde{x}$, thereby achieving a computational saving in solving (14).

In order to work out these ideas we introduce the diagonal matrix D, with diagonal element d_j, defined by

$$\tilde{X} = XD.$$

Let $\rho > 1$ be some fixed number. The algorithm is designed so as to maintain the inequality

$$\frac{1}{\rho} \leq d_i \leq \rho, \qquad 1 \leq i \leq n. \tag{15}$$

Karmarkar [8] already used approximate solutions and partial updating to reduce the complexity bound for his algorithm. Using these approximate solutions for X, we will show that on the average only $\sqrt{n}$ rank-one modifications are needed, without increasing the complexity bound for the required number of iterations. This can be reached by submitting the linesearch to a Goldstein–Armijo condition.

To measure the distance to the central path, we shall now use a slightly different metric. We define

$$\tilde{\delta}(x,\mu) := \min_{y,s}\left\{\left\|D\left(\frac{Xs}{\mu} - e\right)\right\| : A^T y + s = c\right\}. \tag{16}$$

Again, there is a close relationship between this measure and the approximate Newton direction $\tilde{p}$. It can easily be verified that

$$\tilde{\delta}(x,\mu) = \|\tilde{X}^{-1}\tilde{p}\|.$$

A closed-form solution for $\tilde{p}$ is

$$\tilde{p} = -\tilde{X}P_{A\tilde{X}}D\left(\frac{Xc}{\mu} - e\right).$$

It is clear from the definition that $\tilde{\delta}(x,\mu) = 0$ if and only if $x = x(\mu)$. In other words, we will have

$$\delta(x,\mu) = 0 \Longleftrightarrow \tilde{\delta}(x,\mu) = 0.$$

It is easy to verify that

$$\frac{1}{\rho}\delta(x,\mu) \leq \tilde{\delta}(x,\mu) \leq \rho\delta(x,\mu). \tag{17}$$

Consequently, if $\tilde{\delta}(x,\mu) \leq \frac{1}{\rho}$, then we have $\delta(x,\mu) \leq 1$, and then the lemmas proved in the previous section hold.

The Goldstein–Armijo condition can be formulated as follows:

$$\frac{\Delta f}{\alpha} \geq -\zeta\frac{df(x+\alpha\tilde{p},\mu)}{d\alpha}|_{\alpha=0}, \tag{18}$$

where Δf is the change in the barrier function value and $0 < \zeta < 1$. This condition is a well-known rule in nonlinear programming. It permits significant decreases of $f(x, \mu)$, but prevents excessively large steps. Note that we have

$$(19) \qquad \frac{df(x+\alpha\tilde{p},\mu)}{d\alpha}|_{\alpha=0} = \left(\frac{c}{\mu} - X^{-1}e\right)^T \tilde{p} = -\|\tilde{X}^{-1}\tilde{p}\|^2 = -\tilde{\delta}(x,\mu)^2.$$

We will now describe the revised algorithm.

Revised long-step algorithm

Input:

μ_0 is the initial barrier value, $\mu_0 \leq 2^L$;
θ is the reduction parameter, $0 < \theta < 1$;
ρ is the coordinate update parameter, $\rho > 1$;
ζ is the Goldstein–Armijo factor, $\zeta \leq \frac{1}{2}$;
x^0 is a given interior feasible point such that $\tilde{\delta}(x^0, \mu_0) \leq \frac{1}{2\rho}$;

```
begin
  $x := x^0$; $\tilde{x} := x^0$; $\mu := \mu_0$;
  while $x^T s(x,\mu) > 2^{-L}$ do
  begin (outer step)
    while $\tilde{\delta}(x,\mu) > \frac{1}{2\rho}$ do
    begin (inner step)
      $D := \tilde{X}X^{-1}$
      $\tilde{\alpha} := \arg\min_{\alpha>0} \left\{ f(x+\alpha\tilde{p},\mu) : x+\alpha\tilde{p} > 0,\ \frac{\Delta f}{\alpha} \geq \zeta\tilde{\delta}(x,\mu)^2 \right\}$
      $x := x + \tilde{\alpha}\tilde{p}$
      for $j := 1$ to $n$ do if $(\tilde{x}_j/x_j) \notin (\frac{1}{\rho}, \rho)$ then $\tilde{x}_j := x_j$
    end (inner step)
    $\mu := (1-\theta)\mu$;
  end (outer step)
end.
```

For finding the initial point that satisfies the input assumptions of the algorithm, we refer the reader to, e.g., Renegar [12].

4. Convergence analysis of the revised long-step algorithm. We first give an upper bound for the total number of outer and inner iterations. Finally we derive an upper bound for the total number of coordinate updates of $\tilde{x}$.

Henceforth we shall denote $\{x^j\}$, $j = 0, 1, 2, \cdots$, the sequence of inner iterates and $\{\mu_k\}$, $k = 0, 1, 2, \cdots$, the sequence of parameter values during the successive outer iterations. Suppose that x^j is the current iterate when μ_k is calculated. Then set $p_k = j$. Take $p_0 = 0$. Then for any $j > 0$ there is a k such that $p_k < j \leq p_{k+1}$, and the value of μ used in the calculation of x^j is $\mu_k = (1-\theta)^k \mu_0$.

THEOREM 4.1. *After at most $K = O(\frac{L}{\theta})$ outer iterations, the algorithm ends up with a primal and a dual solution such that $x^T s \leq 2^{-L}$.*

Proof. Since $\tilde{\delta}(x,\mu) \leq \frac{1}{2\rho}$ implies $\delta(x,\mu) \leq \frac{1}{2}$, we can derive an upper bound for the duality gap after K outer iterations from Lemma 2.3:

$$x^T s \leq \mu_K \left(n + \frac{1}{2}\sqrt{n}\right),$$

where $\mu_K = (1-\theta)^K \mu_0$. This means that $x^T s \leq 2^{-L}$ certainly holds if

$$(1-\theta)^K \mu_0 \left(n + \frac{1}{2}\sqrt{n} \right) \leq 2^{-L}.$$

Taking logarithms we obtain

$$K \geq \frac{L + \ln(n + \frac{1}{2}\sqrt{n}) + \ln \mu_0}{-\ln(1-\theta)}.$$

Since we have assumed that $\mu_0 \leq 2^L$, and since $\theta \leq -\ln(1-\theta)$, this certainly holds if $K = O(\frac{L}{\theta})$. □

The following two lemmas are needed to derive an upper bound for the number of inner iterations in each outer iteration. The first lemma estimates the difference in barrier function value between the starting point and end point of an outer iteration. The proof is in essence due to Gonzaga [7]. The second lemma states that a sufficient decrease in the barrier function value can be obtained by taking a step along the direction $\tilde{p}$. Moreover, it shows that for any $\zeta \leq \frac{1}{2}$, the Goldstein–Armijo rule (18) can be enforced with the default value

$$\overline{\alpha} = \frac{1}{\rho(\tilde{\delta} + \rho)}.$$

Thus the algorithm is well defined.

LEMMA 4.2. *One has*

$$f(x^{p_k}, \mu_k) - f(x^{p_{k+1}}, \mu_k) \leq \frac{\theta}{1-\theta} \left(\theta n + 3\sqrt{n}\right) + \frac{1}{3}.$$

Proof. The definition of $f(x,\mu)$, $x > 0$, implies that

$$\begin{aligned} f(x,\mu_k) &= f(x,\mu_{k-1}) + \frac{c^T x}{\mu_k} - \frac{c^T x}{\mu_{k-1}} \\ &= f(x,\mu_{k-1}) + \frac{c^T x}{\mu_{k-1}} \left(\frac{1}{1-\theta} - 1 \right) \\ &= f(x,\mu_{k-1}) + \frac{\theta}{1-\theta} \frac{c^T x}{\mu_{k-1}}. \end{aligned}$$

Using this we obtain

$$\begin{aligned} f(x^{p_k}, \mu_k) - f(x^{p_{k+1}}, \mu_k) &= f(x^{p_k}, \mu_{k-1}) - f(x^{p_{k+1}}, \mu_{k-1}) \\ &\quad + \frac{\theta}{1-\theta} \frac{1}{\mu_{k-1}} (c^T x^{p_k} - c^T x^{p_{k+1}}). \end{aligned} \tag{20}$$

First note that because x^{p_k} and $x^{p_{k+1}}$ are approximately centered with respect to $x(\mu_{k-1})$ and $x(\mu_k)$, respectively, using Lemma 2.6 and $\delta = \frac{1}{2}$ for the first and Lemma 2.1 for the second inequality, we find

$$\begin{aligned} c^T x^{p_k} - c^T x^{p_{k+1}} &\leq c^T x(\mu_{k-1}) + \frac{3}{2}\mu_{k-1}\sqrt{n} - c^T x(\mu_k) + \frac{3}{2}\mu_k \sqrt{n} \\ &= c^T x(\mu_{k-1}) - c^T x(\mu_k) + \frac{3}{2}(2-\theta)\mu_{k-1}\sqrt{n} \end{aligned}$$

$$\begin{aligned}&\leq \left(c^T x(\mu_{k-1}) - b^T y(\mu_{k-1})\right)\\ &\quad - \left(c^T x(\mu_k) - b^T y(\mu_k)\right) + 3\mu_{k-1}\sqrt{n}\\ &= \mu_{k-1}n - \mu_k n + 3\mu_{k-1}\sqrt{n}\\ &= \theta\mu_{k-1}n + 3\mu_{k-1}\sqrt{n}\\ &= \mu_{k-1}\left(\theta n + 3\sqrt{n}\right).\end{aligned}$$

Second, using the fact that $x(\mu_{k-1})$ minimizes $f(x(\mu_{k-1}), \mu_{k-1})$ and using Lemma 2.5 and $\delta = \frac{1}{2}$, we obtain

$$\begin{aligned}f(x^{p_k}, \mu_{k-1}) - f(x^{p_{k+1}}, \mu_{k-1}) &= f(x^{p_k}, \mu_{k-1}) - f(x(\mu_{k-1}), \mu_{k-1})\\ &\quad + f(x(\mu_{k-1}), \mu_{k-1}) - f(x^{p_{k+1}}, \mu_{k-1})\\ &\leq f(x^{p_k}, \mu_{k-1}) - f(x(\mu_{k-1}), \mu_{k-1})\\ &\leq \frac{1}{3}.\end{aligned}$$

Hence, substitution of the last two inequalities into (20) yields

$$f(x^{p_k}, \mu_k) - f(x^{p_{k+1}}, \mu_k) \leq \frac{\theta}{1-\theta}\left(\theta n + 3\sqrt{n}\right) + \frac{1}{3}.$$

This proves the lemma. □

The following lemma will be used in Lemma 4.4.

LEMMA 4.3. *For $v \geq 0$ we have:*

$$\ln(1+v) \leq v - \frac{v^2}{2(1+v)}.$$

Proof. First note that $-\ln(1+v) = \ln(1 - \frac{v}{1+v})$. Now using Karmarkar's [8] well-known inequality, we have for $v \geq 0$

$$\ln\left(1 - \frac{v}{1+v}\right) \geq -\frac{v}{1+v} - \frac{1}{2}\frac{\left(\frac{v}{1+v}\right)^2}{1 - \frac{v}{1+v}} = -\frac{v}{1+v} - \frac{v^2}{2(1+v)}.$$

This means that

$$\ln(1+v) \leq \frac{v}{1+v} + \frac{v^2}{2(1+v)} = v - \frac{v^2}{2(1+v)}.$$ □

LEMMA 4.4. *Let $\tilde{\delta} := \tilde{\delta}(x, \mu)$, $\overline{\alpha} := [\rho(\tilde{\delta} + \rho)]^{-1}$. Then*

$$\bar{\Delta} f := f(x, \mu) - f(x + \overline{\alpha}\tilde{p}, \mu) \geq \frac{\tilde{\delta}}{\rho} - \ln\left(1 + \frac{\tilde{\delta}}{\rho}\right).$$

Moreover,

$$\frac{\bar{\Delta} f}{\overline{\alpha}} \geq \zeta\tilde{\delta}^2 \quad \textit{for } \zeta \leq \frac{1}{2}.$$

Proof. We write down the Taylor expansion for f:

$$f(x+\alpha\tilde{p},\mu) = f(x,\mu) + \alpha\tilde{p}^T\nabla f(x,\mu) + \frac{1}{2}\alpha^2\tilde{p}^T\nabla^2 f(x,\mu)\tilde{p} + \sum_{k=3}^{\infty} t_k,$$

where t_k again denotes the kth-order term in the Taylor expansion.

Using the fact that

$$t_k = \frac{(-\alpha)^k}{k}\sum_{i=1}^{n} x_i^{-k}\tilde{p}_i^k,$$

we derive by the definition of $\tilde{x}$ and $\tilde{\delta}$,

$$|t_k| \le \frac{\alpha^k}{k}\sum_{i=1}^{n}|x_i^{-1}\tilde{p}_i|^k = \frac{\alpha^k}{k}\sum_{i=1}^{n}|d_i\tilde{x}_i^{-1}\tilde{p}_i|^k \le \frac{\alpha^k}{k}\left(\sum_{i=1}^{n}|d_i\tilde{x}_i^{-1}\tilde{p}_i|^2\right)^{k/2} = \frac{\alpha^k}{k}\rho^k\tilde{\delta}^k.$$

Further,

$$\tilde{p}^T\nabla^2 f(x,\mu)\tilde{p} = \tilde{p}^T X^{-2}\tilde{p} = \|D\tilde{X}^{-1}\tilde{p}\|^2 \le \rho^2\tilde{\delta}^2,$$

and, using the fact that $A\tilde{X}\tilde{X}^{-1}\tilde{p} = A\tilde{p} = 0$,

$$\begin{aligned}\tilde{p}^T\nabla f(x,\mu) &= (\tilde{X}^{-1}\tilde{p})^T\tilde{X}\nabla f(x,\mu)\\ &= (\tilde{X}^{-1}\tilde{p})^T P_{A\tilde{X}}(\tilde{X}\nabla f(x,\mu))\\ &= -(\tilde{X}^{-1}\tilde{p})^T\tilde{X}^{-1}\tilde{p}\\ &= -\tilde{\delta}^2.\end{aligned}$$

So if $\alpha\rho\tilde{\delta} < 1$, we find

$$\begin{aligned}f(x+\alpha\tilde{p},\mu) &\le f(x,\mu) - \alpha\tilde{\delta}^2 + \frac{1}{2}\alpha^2\rho^2\tilde{\delta}^2 + \sum_{k=3}^{\infty}\frac{\alpha^k}{k}\rho^k\tilde{\delta}^k\\ &= f(x,\mu) - \alpha\tilde{\delta}^2 - \ln(1-\alpha\rho\tilde{\delta}) - \alpha\rho\tilde{\delta}.\end{aligned}$$

Hence

$$f(x,\mu) - f(x+\alpha\tilde{p},\mu) \ge \alpha\tilde{\delta}^2 + \alpha\rho\tilde{\delta} + \ln(1-\alpha\rho\tilde{\delta}).$$

The right-hand side is maximal if $\alpha = \overline{\alpha} = [\rho(\tilde{\delta}+\rho)]^{-1}$. Note that $\overline{\alpha}\rho\tilde{\delta} < 1$. Substitution of this value finally gives

$$\bar{\Delta}f \ge \frac{\tilde{\delta}^2}{\rho(\tilde{\delta}+\rho)} + \frac{\rho\tilde{\delta}}{\rho(\tilde{\delta}+\rho)} + \ln\left(1 - \frac{\rho\tilde{\delta}}{\rho(\tilde{\delta}+\rho)}\right) = \frac{\tilde{\delta}}{\rho} - \ln\left(1+\frac{\tilde{\delta}}{\rho}\right). \tag{21}$$

This proves the first part of the lemma. The second part follows immediately from (21) and Lemma 4.3:

$$\bar{\Delta}f \ge \frac{\tilde{\delta}}{\rho} - \ln\left(1+\frac{\tilde{\delta}}{\rho}\right) \ge \frac{\tilde{\delta}}{\rho} - \frac{\tilde{\delta}}{\rho} + \frac{\tilde{\delta}^2/\rho^2}{2(1+\frac{\tilde{\delta}}{\rho})} = \frac{\tilde{\delta}^2}{2\rho(\tilde{\delta}+\rho)} = \frac{1}{2}\overline{\alpha}\tilde{\delta}^2. \tag{22}$$

□

THEOREM 4.5. *The number of inner iterations for each outer iteration, denoted by P, is bounded by*

$$P \leq 12\frac{\theta\rho^4}{1-\theta}\left(\theta n + 3\sqrt{n}\right) + 4\rho^4.$$

Proof. Let us consider the $(k+1)$st outer iteration. Let P denote the number of inner iterations. For each inner iteration we know, by the definition of $\tilde{\alpha}$ and (22), that the decrease in the barrier function value is larger than

$$\frac{\tilde{\delta}^2}{2\rho(\tilde{\delta}+\rho)}.$$

Since this expression is an increasing function of $\tilde{\delta}$, and since during each iteration $\tilde{\delta} \geq \frac{1}{2\rho}$, we have

$$\frac{\tilde{\delta}^2}{2\rho(\tilde{\delta}+\rho)} \geq \frac{1}{12\rho^4}.$$

Consequently, we have

$$f(x^{p_{k+1}}, \mu_k) \leq f(x^{p_k}, \mu_k) - \frac{1}{12\rho^4}P.$$

Equivalently,

$$\frac{1}{12\rho^4}P \leq f(x^{p_k}, \mu_k) - f(x^{p_{k+1}}, \mu_k).$$

Now using Lemma 4.2, the theorem follows. □

Consequently, using an additional Goldstein–Armijo rule and approximate solutions does not influence the order of the total number of outer and inner iterations.

The last theorem will give an upper bound for the total number of coordinate updates in $\tilde{x}$. For the proof of this theorem we make use of some results obtained by Anstreicher [1]. The following lemma will be used in the theorem.

LEMMA 4.6. *Let $w \in \mathbb{R}$, $0 < w < 1$, and $v \in \mathbb{R}$, $v \geq w$. Then*

$$|\ln v| \leq \frac{|1-v||\ln w|}{1-w}.$$

Proof. Defining

$$f(z) := \begin{cases} \frac{\ln z}{z-1} & \text{if } z \neq 1, \\ 1 & \text{if } z = 1, \end{cases}$$

it is easy to see that $f(z)$ is monotonically decreasing and positive for $z \in (0, \infty)$. Hence

$$\left|\frac{\ln v}{v-1}\right| \leq \left|\frac{\ln w}{w-1}\right| = \frac{|\ln w|}{1-w}.$$

This implies the lemma. □

THEOREM 4.7. *The total number M of coordinate updates of $\tilde{x}$ up to the last inner iteration N is bounded by*

$$M \leq \frac{2\rho^3\sqrt{n}}{\zeta(\rho-1)}\left(\frac{\theta n + 3\sqrt{n}}{1-\theta} + \frac{1}{3\theta}\right) O(L).$$

Proof. Let k_1 be an iteration at which an update of $\tilde{x}_i$ is performed. Let $k_2 > k_1$ be the first iteration at which $\tilde{x}_i$ is updated again. Then we have

$$\begin{aligned}\prod_{j=k_1+1}^{k_2} \max\left(\frac{x_i^j}{x_i^{j-1}}, \frac{x_i^{j-1}}{x_i^j}\right) &\geq \max\left(\prod_{j=k_1+1}^{k_2} \frac{x_i^j}{x_i^{j-1}}, \prod_{j=k_1+1}^{k_2} \frac{x_i^{j-1}}{x_i^j}\right)\\ &= \max\left(\frac{x_i^{k_2}}{x_i^{k_1}}, \frac{x_i^{k_1}}{x_i^{k_2}}\right)\\ &\geq \rho.\end{aligned}$$

Taking logarithms and defining

$$r_i^j := 1 + \tilde{\alpha}_j (x_i^j)^{-1} \tilde{p}_i^j = \frac{x_i^{j+1}}{x_i^j},$$

we obtain

$$\ln\rho \leq \sum_{j=k_1}^{k_2-1} |\ln r_i^j|. \tag{23}$$

Let

$$\hat{r}_i^j = \max\left\{r_i^j, \frac{1}{\rho}\right\}.$$

Inequality (23) can be sharpened to

$$\ln\rho \leq \sum_{j=k_1}^{k_2-1} |\ln \hat{r}_i^j|. \tag{24}$$

To prove (24), first assume that for some ℓ, $k_1 \leq \ell \leq k_2 - 1$, $r_i^j < \frac{1}{\rho}$. Then

$$\ln\rho = |\ln \hat{r}_i^\ell| \leq \sum_{j=k_1}^{k_2-1} |\ln \hat{r}_i^j|.$$

Otherwise, $\hat{r}_i^j = r_i^j$, $k_1 \leq j \leq k_2 - 1$, and (24) holds because of (23). Hence (24) has been proved.

We deduce from (24) a bound on the total number m_i of updates of coordinate i of $\tilde{x}$:

$$m_i \ln\rho \leq \sum_{j=0}^{N-1} |\ln \hat{r}_i^j|.$$

Consequently the total number of coordinate updates is bounded by

$$M \ln \rho \le \sum_{j=0}^{N-1} \sum_{i=1}^{n} |\ln \hat{r}_i^j|. \tag{25}$$

In view of Lemma 4.6, with $v = \hat{r}_i^k$ and $w = \frac{1}{\rho}$,

$$|\ln \hat{r}_i^j| \le \frac{\ln \rho}{1 - \frac{1}{\rho}} |1 - \hat{r}_i^j|. \tag{26}$$

Since $\hat{r}_i^j = r_i^j$ if $r_i^j \ge \frac{1}{\rho}$, and $\hat{r}_i^j > r_i^j$ if $r_i^j < \frac{1}{\rho}$, we always have

$$|1 - \hat{r}_i^j| \le |1 - r_i^j| = \tilde{\alpha}_j |(x_i^j)^{-1} \tilde{p}_i^j|. \tag{27}$$

Substitution of (26) and (27) into (25) gives

$$M \le \frac{\rho}{\rho - 1} \sum_{j=0}^{N-1} \tilde{\alpha}_j \sum_{i=1}^{n} |(x_i^j)^{-1} \tilde{p}_i^j|.$$

From the inequality between the l_1 and l_2 norms,

$$\sum_{i=1}^{n} |(x_i^j)^{-1} \tilde{p}_i^j| \le \sqrt{n} \|(X^j)^{-1} \tilde{p}^j\| \le \rho \sqrt{n} \|(\tilde{X}^j)^{-1} \tilde{p}^j\|.$$

Hence

$$M \le \frac{\rho^2 \sqrt{n}}{\rho - 1} \sum_{j=0}^{N-1} \tilde{\alpha}_j \|(\tilde{X}^j)^{-1} \tilde{p}^j\|. \tag{28}$$

Since the Goldstein–Armijo condition is satisfied at each inner iteration, for any j and k such that $p_k < j \le p_{k+1}$ (we will write $k(j)$ instead of k to denote its dependence on j),

$$\tilde{\alpha}_j \le \frac{f(x^j, \mu_{k(j)}) - f(x^{j+1}, \mu_{k(j)})}{\zeta \|(\tilde{X}^j)^{-1} \tilde{p}^j\|^2}. \tag{29}$$

Substituting this into (28) we obtain

$$M \le \frac{\rho^2 \sqrt{n}}{\zeta(\rho - 1)} \sum_{j=0}^{N-1} \frac{f(x^j, \mu_{k(j)}) - f(x^{j+1}, \mu_{k(j)})}{\|(\tilde{X}^j)^{-1} \tilde{p}^j\|}.$$

Since $\|(\tilde{X}^j)^{-1} \tilde{p}^j\| \ge \frac{1}{2\rho}$, this implies

$$\begin{aligned} M &\le \frac{2\rho^3 \sqrt{n}}{\zeta(\rho - 1)} \sum_{j=0}^{N-1} (f(x^j, \mu_{k(j)}) - f(x^{j+1}, \mu_{k(j)})) \\ &= \frac{2\rho^3 \sqrt{n}}{\zeta(\rho - 1)} \sum_{k=0}^{K-1} (f(x^{p_k}, \mu_k) - f(x^{p_{k+1}}, \mu_k)). \end{aligned}$$

Now using Theorem 4.1 and Lemma 4.2 we obtain

$$M \leq \frac{2\rho^3\sqrt{n}}{\zeta(\rho-1)}\left(\frac{\theta n + 3\sqrt{n}}{1-\theta} + \frac{1}{3\theta}\right) O(L). \qquad \square$$

Theorems 4.1 and 4.5 imply that N, the total number of inner iterations needed by the algorithm, is bounded by

$$N \leq \left(12\frac{\rho^4}{1-\theta}\left(\theta n + 3\sqrt{n}\right) + 4\frac{\rho^4}{\theta}\right) O(L).$$

The total number of arithmetic operations in each iteration, aside from the work due to coordinate updates, is $O(n^2)$. The same amount of work must be done for one coordinate update. Consequently, the total number of arithmetic operations needed by the algorithm is $(N+M)O(n^2)$.

COROLLARY 4.8.

- *If $0 < \theta < 1$, independent of n and L, then the total number of iterations is bounded by $O(nL)$ and the total number of coordinate updates by $O(n^{1.5}L)$. Consequently, the total complexity is $O(n^{3.5}L)$.*
- *If $\theta = \nu/\sqrt{n}$, $\nu > 0$ and independent of n and L, then the total number of iterations is bounded by $O(\sqrt{n}L)$ and the total number of coordinate updates by $O(nL)$. Consequently, the total complexity is $O(n^3L)$.*

Acknowledgments. We would like to thank Kurt Anstreicher for the valuable suggestions for the proof of Theorem 4.7 that he made during a visit to Delft, from May 14–24, 1990. Thanks are also due to the two referees for their comments.

REFERENCES

[1] K. M. ANSTREICHER (1991), *A standard form variant, and safeguarded linesearch, for the modified Karmarkar algorithm*, Math. Programming, 47, pp. 337–351.

[2] K. M. ANSTREICHER AND R. A. BOSCH (1988), *Long steps in an $O(n^3L)$ algorithm for linear programming*, Preprint, Yale School of Organization and Management, New Haven, CT.

[3] A. V. FIACCO AND G. P. MCCORMICK (1968), *Nonlinear Programming, Sequential Unconstrained Minimization Techniques*, John Wiley and Sons, New York.

[4] R. M. FREUND (1988), *Polynomial-time algorithms for linear programming based only on primal scaling and projected gradients of a potential function*, Preprint, Sloan School of Management, Massachusetts Institute of Technology, Cambridge, MA.

[5] P. E. GILL, W. MURRAY, M. A. SAUNDERS, J. A. TOMLIN, AND M. H. WRIGHT (1986), *On projected Newton barrier methods for linear programming and an equivalence to Karmarkar's projective method*, Math. Programming, 36, pp. 183–209.

[6] C. C. GONZAGA (1989), *An algorithm for solving linear programming problems in $O(n^3L)$ operations*, in Progress in Mathematical Programming, Interior Point and Related Methods, N. Megiddo, ed., Springer-Verlag, New York, pp. 1–28.

[7] ——— (1989), *Large-steps path-following methods for linear programming, Part I: Barrier function method*, Report ES–210/89, Department of Systems Engineering and Computer Sciences, COPPE–Federal University of Rio de Janeiro, Rio de Janeiro, Brasil; also in SIAM J. Optimization, 1 (1991), pp. 268–279.

[8] N. KARMARKAR (1984), *A new polynomial-time algorithm for linear programming*, Combinatorica, 4, pp. 373–395.

[9] M. KOJIMA, S. MIZUNO, AND A. YOSHISE (1989), *A polynomial-time algorithm for a class of linear complementarity problems*, Math. Programming, 44, pp. 1–26.

[10] S. MIZUNO AND M. J. TODD (1989), *An $O(n^3L)$ long step path following algorithm for a linear complementarity problem*, Res. Report 23, Tokyo Institute of Technology, Tokyo, Japan.

[11] R. D. C. MONTEIRO AND I. ADLER (1989), *Interior path following primal-dual algorithms, Part II: Linear programming*, Math. Programming, 44, pp. 27–41.

[12] J. RENEGAR (1988), *A polynomial-time algorithm, based on Newton's method, for linear programming*, Math. Programming, 40, pp. 59–93.

[13] C. ROOS AND J.-PH. VIAL (1988), *A polynomial method of approximate centers for linear programming*, Math. Programming, to appear.

[14] ——— (1990), *Long steps with the logarithmic penalty barrier function in linear programming*, in Economic Decision-Making: Games, Economics and Optimization, dedicated to Jacques H. Drèze; J. Gabszevwicz, J.-F. Richard, and L. Wolsey, eds., Elsevier Science Publisher B.V., Amsterdam, pp. 433–441.

[15] P. M. VAIDYA (1990), *An algorithm for linear programming which requires $O(((m+n)n^2 + (m+n)^{1.5}n)L)$ arithmetic operations*, Math. Programming, 47, pp. 175–201.

[16] Y. YE (1991), *An $O(n^3L)$ potential reduction algorithm for linear programming*, Math. Programming, 50, pp. 239–258.

SIAM J. OPTIMIZATION
Vol. 2, No. 1, pp. 88–120, February 1992

007

LARGE-SCALE OPTIMIZATION OF EIGENVALUES[*]

MICHAEL L. OVERTON[†]

Abstract. Optimization problems involving eigenvalues arise in many applications. Let x be a vector of real parameters and let $A(x)$ be a continuously differentiable symmetric matrix function of x. We consider a particular problem that occurs frequently: the minimization of the maximum eigenvalue of $A(x)$, subject to linear constraints and bounds on x. The eigenvalues of $A(x)$ are not differentiable at points x where they coalesce, so the optimization problem is said to be nonsmooth. Furthermore, it is typically the case that the optimization objective tends to make eigenvalues coalesce at a solution point.

There are three main purposes of the paper. The first is to present a clear and self-contained derivation of the Clarke generalized gradient of the max eigenvalue function in terms of a "dual matrix." The second purpose is to describe a new algorithm, based on the ideas of a previous paper by the author [*SIAM J. Matrix Anal. Appl.*, 9 (1988), pp. 256–268], which is suitable for solving large-scale eigenvalue optimization problems. The algorithm uses a "successive partial linear programming" formulation that should be useful for other large-scale structured nonsmooth optimization problems as well as large-scale nonlinear programming with a relatively small number of nonlinear constraints. The third purpose is to report on our extensive numerical experience with the new algorithm, solving problems that arise in the following application areas: the optimal design of columns against buckling; the construction of optimal preconditioners for numerical linear equation solvers; the bounding of the Shannon capacity of a graph. We emphasize the role of the dual matrix, whose dimension is equal to the multiplicity of the minimal max eigenvalue. The dual matrix is computed by the optimization algorithm and used for verification of optimality and sensitivity analysis.

Key words. nonsmooth optimization, nondifferentiable optimization, generalized gradient, eigenvalue perturbation

AMS(MOS) subject classifications. 65F15, 65K10, 49K99, 90C26

1. Introduction.

Eigenvalues of symmetric matrices play important roles in many different areas of applied mathematics. For perhaps the large majority of true applications, it is not the case that a fixed matrix, say A, is known, and its eigenvalues are needed. It is more typical that A depends on many parameters, and that the eigenvalues are desired for many different choices of the parameters. In many cases the choice of parameters is dictated by some optimization objective. For example, in a control application, where the size of the largest eigenvalue represents system stability, it may be desirable to minimize the largest eigenvalue, while in a structure analysis application, where the smallest eigenvalue represents a buckling load, it may be desirable to maximize the smallest eigenvalue. Other applications might have an optimization objective that does not involve eigenvalues (e.g., cost of a material), but include constraints on eigenvalues (e.g., ensure all eigenvalues are in a safe frequency interval).

In our work on optimization problems involving eigenvalues, we have found it very useful to concentrate on a particular model problem, namely, minimizing the maximum eigenvalue of a symmetric $n \times n$ matrix $A(x)$, where $A(x)$ depends smoothly on a vector of parameters $x \in \Re^m$. It is useful and not significantly more complicated to allow the imposition of linear constraints on x. A common variation is to minimize the maximum eigenvalue in absolute value. (We avoid the term spectral radius, since this suggests complex eigenvalues; nonsymmetric matrices are not discussed in this

[*] Received by the editors May 18, 1990; accepted for publication (in revised form) March 4, 1991. This work was supported in part by National Science Foundation grant CCR-9101640.

[†] Courant Institute of Mathematical Sciences, New York University, 251 Mercer Street, New York, New York 10012.

paper, but see [37].) The model problem is directly applicable to many applications, including the first two mentioned above, while for other problems, e.g., those where only the constraints involve the eigenvalues, it is fairly clear how the main ideas should be extended.

The feature of eigenvalue optimization problems that makes them both particularly interesting and particularly difficult to solve is that the eigenvalues of a differentiable matrix function are not themselves differentiable at points where they coalesce. Furthermore, it is often the case that the optimization objective tends to make the eigenvalues coalesce at a solution point. For example, consider the model problem with

$$A(x) = \begin{bmatrix} 1+x_1 & x_2 \\ x_2 & 1-x_1 \end{bmatrix}.$$

The eigenvalues are

$$1 \pm \sqrt{x_1^2 + x_2^2},$$

so the maximum eigenvalue is minimized by $x = 0$. Clearly the maximum eigenvalue is not a smooth function at $x = 0$. More importantly, though, the max eigenvalue function cannot be written as the pointwise maximum of two smooth functions at $x = 0$; in other words, the eigenvalues themselves cannot be labeled, say, κ_1 and κ_2, each a smooth function of $x \in \Re^2$. Thus standard minmax optimization techniques (e.g., [31]) cannot be applied. Suggestions for transforming the problem into a standard nonlinear programming form by means of determinants have been made [18], but these methods perform poorly [41]; for other comments on the use of determinants, see [15].

In the example given above, the maximum eigenvalue is convex in x. This is true in general when A depends linearly on x, since the Rayleigh principle can be used to show that the maximum eigenvalue is a convex function of the matrix elements. Because of this fact, it has been recognized for some time that the techniques of convex analysis (e.g., [45]) are applicable to eigenvalue optimization problems; optimality conditions and/or first-order algorithms for various problem classes have been given by [7], [43], [9], [49], [19], and [1]. See also [34] and [4] for discussion of problems arising in structural engineering.

In [36], a quadratically convergent algorithm was given to solve the model problem, using a "dual matrix" formulation of the optimality conditions to fully exploit the nonsmooth problem structure. Two papers that greatly influenced this work were [15] and [12]. Numerical examples were given, demonstrating quadratic convergence to nonsmooth solutions. The assumption was made that $A(x)$ was affine, although it was indicated that this was not essential for the main ideas to apply. The reason for this is that the eigenvalues are nonsmooth, nonlinear functions of the matrix, so whether $A(x)$ depends linearly or nonlinearly on x is not of great importance, provided $A(x)$ is a smooth function. If $A(x)$ is nonlinear, the maximum eigenvalue is not necessarily convex in x, but it is a composition of a convex function with a smooth function. Optimality conditions for nonlinear $A(x)$, for the more general case of minimizing sums of largest eigenvalues (algebraically or in absolute value), are given by [38]. These optimality conditions are derived by characterizing Clarke's generalized gradient [5] in terms of a dual matrix. Proofs of the local quadratic convergence of the successive quadratic programming algorithm used in [36] are being developed in [39].

There are three main purposes of the present paper. The first is to present a clear and self-contained derivation of the generalized gradient of the max eigenvalue functional in terms of a dual matrix. An understanding of this is essential for the appreciation of the main ideas underlying our optimization algorithms. Our second contribution is to describe a new algorithm, based on the ideas of [36], that is suitable for solving large-scale eigenvalue optimization problems. The third purpose of the paper is to report on our extensive numerical experience with the new algorithm, solving eigenvalue optimization problems that arise in three very interesting and quite different application areas.

The paper is organized as follows. Section 2 derives the generalized gradient of the max eigenvalue, and consequent optimality conditions for the model problem, using the dual matrix formulation. Section 3 discusses the role of the dual matrix in eigenvalue splitting and sensitivity analysis. Section 4 summarizes the eigenvalue optimization algorithm of [36] and relates this to the generalized gradient derived in §2. Section 5 explains how to extend the main ideas of [36] to solve problems with large numbers of variables. The ideas of this section should also be useful for solving other structured large-scale nonsmooth optimization problems as well as nonlinear programming problems with a relatively small number of nonlinear constraints—both active areas of current research. Section 6 discusses how to efficiently compute the eigenvalues of the matrix iterates generated by the optimization algorithm when the dimension of the matrices is large. Section 7 explains how all of the foregoing may be generalized to apply to eigenvalue problems of the form $A(x)q = \lambda Bq$, where B is a fixed symmetric positive definite matrix. Section 8 discusses the case where several different matrix families are involved. Section 9 summarizes numerical results that have been obtained for a fascinating classical problem of Lagrange, finding the shape of the strongest column. Here the task is to maximize the smallest eigenvalue of a fourth-order differential equation. Section 10 discusses results obtained for finding optimal preconditioners for the solution of linear systems of equations. Section 11 discusses the application of our large-scale algorithm to a problem arising in graph theory, computing the Lovász number of a graph. Section 12 makes some concluding remarks.

2. Optimality conditions, the generalized gradient, and dual matrices.

We start with some notation. Let $\Re^{n\times m}$ denote the set of n by m real matrices, and let $S\Re^{n\times n}$ denote the set of n by n real symmetric matrices. By $A \geq 0$, where A is symmetric, we mean that A is positive semidefinite. The notation $\|\ .\ \|$ will always denote the Euclidean vector norm. Let $\langle , \rangle$ denote the Frobenius inner product on the set of rectangular matrices, namely,

$$\langle B,C\rangle \ = \ \operatorname{tr} B^TC = \operatorname{tr} C^TB = \sum b_{ij}c_{ij},$$

where the dimensions of the matrices depend on the context. (For example, B and C could be vectors.)

We now give a simple but important lemma.

LEMMA 1. *The convex hull of the set*

$$\{ww^T : w \in \Re^n, \| w \|= 1\}$$

is the set

$$\{\tilde{U} : \tilde{U} \in S\Re^{n\times n}, \operatorname{tr} \tilde{U} = 1, \tilde{U} \geq 0\}.$$

Furthermore, the elements in the first set are the extreme points of the second set.

Proof. Any convex combination of the first set is clearly contained in the second. Furthermore, any matrix in the second set has a spectral decomposition

$$\tilde{U} = \sum \theta_i w_i w_i^T,$$

where the eigenvalues θ_i are nonnegative by the positive semidefinite condition and sum to one by the trace condition, and the eigenvectors w_i have unit norm, i.e., the right-hand side is a convex combination of elements in the first set. Clearly, any element of the first set is an extreme point of the second set. Also, any element of the second set that is not rank-one can be written as a nontrivial convex sum of elements in the first set and is therefore not an extreme point.

THEOREM 1. *Let $A \in S\Re^{n\times n}$, and let $\lambda_1(A)$ be the largest eigenvalue of A. The following characterizations hold:*

$$\lambda_1(A) = \max\{\langle q, Aq\rangle \ : \ \| q \| = 1\}; \tag{1}$$

$$\lambda_1(A) = \max\{\langle qq^T, A\rangle \ : \ \| q \| = 1\}; \tag{2}$$

$$\lambda_1(A) = \max\{\langle \tilde{U}, A\rangle \ : \ \tilde{U} \in S\Re^{n\times n}, \ \operatorname{tr} \tilde{U} = 1, \ \tilde{U} \geq 0\}. \tag{3}$$

Consequently, λ_1 is a convex function of A.

Proof. Equation (1) is the well-known Rayleigh quotient characterization. Equation (2) follows immediately from properties of the inner product. Equation (3) follows from Lemma 1, since maximizing a linear function over a set gives the same result as maximizing over its convex hull. The convexity follows from any of the characterizations, since the pointwise maximum of a set of linear functions is always convex. □

The characterization of a convex function as a pointwise maximum of a set of linear functions leads directly to the definition of the *subdifferential* of f. For example, suppose that $z \in \Re^k$, and

$$f(z) = \max\{\langle a_i, z\rangle + \beta_i \ : \ i \in \mathcal{I}\},$$

where $\mathcal{I}$ is a discrete index set. Then the subdifferential of f at z may be defined as

$$\partial f(z) = \operatorname{conv}\{a_i : i \in \mathcal{I}, \ f(z) = \langle a_i, z\rangle + \beta_i\}$$

where "conv" denotes convex hull. An important property of ∂f that immediately follows from this definition is that z minimizes f if and only if $0 \in \partial f(z)$; note also that f is differentiable at z if and only if the subdifferential contains only one element, namely, the gradient of f at z. It is a fact [45, Cor. 23.5.3] that the subdifferential may be defined in this way for general convex functions, giving, as a consequence of (2),

$$\partial\lambda_1(A) = \operatorname{conv}(\{qq^T \ : \ q \text{ is a normalized eigenvector for } \lambda_1(A)\}). \tag{4}$$

This leads to the following theorem.

THEOREM 2. *Suppose the maximum eigenvalue $\lambda_1(A)$ has multiplicity t, i.e., the eigenvalues of A are*

$$\lambda_1 = \cdots = \lambda_t > \lambda_{t+1} \geq \cdots \geq \lambda_n.$$

Then the subdifferential of λ_1 *at* A *is the set*

$$\partial\lambda_1(A) = \operatorname{conv}(\{Q_1 w w^T Q_1^T : w \in \Re^t, \| w \| = 1\}), \tag{5}$$

where the columns of Q_1 *form an orthonormal set of eigenvectors for* $\lambda_1(A)$. *Another equivalent form is*

$$\partial\lambda_1(A) = \{\tilde{U} = Q_1 U Q_1^T : \; U \in S\Re^{t\times t}, \; \operatorname{tr} U = 1, \; U \geq 0\}. \tag{6}$$

Proof. Equation (5) follows directly from (4), and (6) then follows from Lemma 1. Alternatively, writing the eigendecomposition of A as

$$A = Q \operatorname{Diag}(\lambda_i)\, Q^T, \qquad Q = [Q_1 \; Q_2],$$

we see that (6) follows from directly applying the definition of the subdifferential to (3) since the matrices on the right-hand side of (6) are those that achieve the maximum in (3), with

$$\tilde{U} = Q \begin{bmatrix} U & 0 \\ 0 & 0 \end{bmatrix} Q^T.$$

No convex hull operation is necessary, since the set is already convex. □

We now change notation, introducing $A(x) \in S\Re^{n\times n}$, a continuously differentiable function of $x \in \Re^m$, with eigenvalues

$$\lambda_1(x) \geq \cdots \geq \lambda_n(x),$$

and partial derivatives

$$A_k(x) = \frac{\partial A}{\partial x_k}(x).$$

It is convenient to use the symbol λ_1 for two purposes, with

$$\lambda_1(x) \equiv \lambda_1(A(x)),$$

and the distinction should be clear from the context. The function $\lambda_1(x)$ is not generally convex, but it is the composition of the convex function $\lambda_1(A)$ with the smooth function $A(x)$. The Clarke *generalized gradient* of $\lambda_1(x)$ may therefore be defined by means of a chain rule [5, p. 42], [13, p. 366]. We obtain the following theorem.

THEOREM 3. *Suppose the maximum eigenvalue of* $A(x)$ *has multiplicity* t, *with a corresponding orthonormal basis of eigenvectors* $Q_1(x) = [q_1(x), \cdots, q_t(x)]$. *The generalized gradient of* $\lambda_1(x)$ *is the set*

$$\begin{aligned} \partial\lambda_1(x) \;\; = \;\; & \{v \in \Re^m \; : \; v_k = \; \langle U, Q_1(x)^T A_k(x) Q_1(x)\rangle, \\ & \textit{for some } U \in S\Re^{t\times t}, U \geq 0, \operatorname{tr} U = 1\}. \end{aligned} \tag{7}$$

Proof. By the chain rule just cited,

$$\partial\lambda_1(x) = \{v \in \Re^m \; : \; v_k = \; \langle G, A_k(x)\rangle \;\text{ for some } G \in \partial\lambda_1(A)\}.$$

The proof is completed by using (6) and noting that

$$\langle Q_1 U Q_1^T, A_k \rangle \; = \; \langle U, Q_1^T A_k Q_1 \rangle. \qquad \square$$

Equation (4) is well known; see [7], [43], and [5]. The equivalent form (6) is much less known and much more useful, as we shall see shortly; the earliest reference we know for this explicit form is Fletcher [12], where a different proof was given. Equation (7) was given in the case that $A(x)$ is affine in [36], using a proof based on Fletcher's work. The proofs given here make more use of the machinery of [5] and [45]. A referee has pointed out that Clarke's powerful theory is not required for Theorem 3 and subsequent results, which could in fact be obtained from the theory of "locally convex" functions; see [24] and [49]. We prefer to refer to Clarke's work so that we may use the beautifully simple notion of a chain rule developed there.

The matrix $\tilde{U}$ may be viewed as a "dual matrix"; indeed, a "dual problem" is formulated at the end of this section. The $t \times t$ matrix U may be called a "reduced dual matrix," but since it is the one we shall need as a computational tool we shall also refer to it simply as the dual matrix. (The term "Lagrange matrix" was used in [36].) The distinction between $\tilde{U}$ and U is analogous to the notational question of whether inactive constraints in a nonlinear program should be assigned zero Lagrange multipliers.

Theorem 3 gives a form of the generalized gradient that is particularly useful for computation, since it does not involve taking a convex hull. Indeed, it characterizes the generalized gradient using *structure functionals*, to use a term introduced by Osborne [35] for some other nonsmooth optimization problems. In our case, the structure functionals may be taken to be the $t(t+1)/2$ quantities

$$q_i^T A(x) q_j, \qquad 1 \le i \le j \le t, \tag{8}$$

assuming the eigenvectors $q_1, \cdots, q_t$ are fixed. Theorem 3 then states that the generalized gradient of $\lambda_1(x)$ consists of particular linear combinations of the gradients of the structure functionals, namely, those with coefficients u_{ii} and $2u_{ij}$ $(j \neq i)$ making up a positive semidefinite dual matrix U with trace one. (A better definition of the structure functionals, which would allow statements about second-order effects, would presumably use the matrix exponential formulation mentioned in §4.)

Note that the eigenvector basis Q_1 for $\lambda_1(x)$ is not unique if $t > 1$ (and even if $t = 1$ the sign is not unique). However, replacing Q_1 by any other valid choice, which must have the form $Q_1 V$ for some $t \times t$ orthogonal matrix V, simply transforms the dual matrix U into VUV^T, preserving its eigenvalues.

The directional derivative of λ_1 is easily deduced from the generalized gradient formula. We have the following theorem.

THEOREM 4. *Under the assumptions of Theorem* 3, *the directional derivative*

$$\lambda_1'(x; d) = \lim_{\alpha \to 0^+} \frac{\lambda_1(x + \alpha d) - \lambda_1(x)}{\alpha}$$

is the largest eigenvalue of

$$B(d) = \sum_{k=1}^{m} d_k Q_1(x)^T A_k(x) Q_1(x). \tag{9}$$

Proof. Because $\lambda_1(x)$ is the composition of a convex function with a smooth function,

$$\lambda_1'(x;d) = \max_{v\in\partial\lambda_1(x)} \langle v, d\rangle$$

(see [13, p. 369] or [5, Chap. 2]). By (7), we therefore obtain

$$\lambda_1'(x;d) = \max_U \langle U, B(d)\rangle,$$

where the max is taken over positive semidefinite matrices with trace one. The result therefore follows from Theorem 1. □

The formula for the directional derivative may alternatively be obtained from the classical results in [25], which state that the multiple eigenvalue $\lambda_1 = \cdots = \lambda_t$ of $A(x)$ splits into t eigenvalues of $A(x+\alpha d)$, for α near 0, with corresponding derivatives equal to the eigenvalues of $B(d)$. However, the proof of this basic fact is not at all straightforward, especially in the case that $A(x)$ cannot be extended to an analytic function of complex variables.

We now consider optimality conditions for a constrained version of the model problem.

THEOREM 5. *Consider the problem:*

$$\min_x \lambda_1(x) \tag{10}$$

subject to

$$Cx = b; \qquad \ell \le x \le u, \tag{11}$$

where $C = [c_1, \cdots, c_m] \in \Re^{n_c\times m}, b \in \Re^{n_c}, \ell$ *and* $u \in \Re^m$. *Then a necessary condition for x to solve* (10)–(11) *is, in addition to* (11), *that there exists a dual matrix* $U \in S\Re^{t\times t}$, *where t is the multiplicity of* $\lambda_1(x)$, *and vectors of Lagrange multipliers* $\mu \in \Re^{n_c}$ *and* $\gamma \in \Re^m$, *satisfying*

$$\langle U, Q_1(x)^T A_k(x) Q_1(x)\rangle \;=\; \langle \mu, c_k\rangle \;+\; \gamma_k, \qquad k = 1, \cdots, m, \tag{12}$$

$$\operatorname{tr} U = 1, \tag{13}$$

$$U \ge 0, \tag{14}$$

and

$$\begin{array}{rcl} \gamma_k & = & 0 \;\text{ if } \ell_k < x_k < u_k; \\ \gamma_k & \ge & 0 \;\text{ if } x_k = \ell_k; \\ \gamma_k & \le & 0 \;\text{ if } x_k = u_k. \end{array} \tag{15}$$

Here the columns of $Q_1(x)$ *form an orthonormal basis of t eigenvectors for* $\lambda_1(x)$. *The necessary condition* (*together with the satisfaction of* (11)) *is also sufficient for optimality if* $A(x)$ *is affine.*

Proof. The proof follows from the standard Lagrange multiplier rule for nonsmooth optimization [5, pp. 228, 240], which reduces to $0 \in \partial\lambda_1(x)$ in the case that there are no constraints. The last statement holds because if $A(x)$ is affine, $\lambda_1(x)$ is a composition of a convex with an affine function, and is therefore convex. □

We complete this section with a discussion of a duality result, which clarifies the terminology "dual matrix." By (3), the "primal problem" (10)–(11) is equivalent to

$$\min_{Cx=b;\ \ell\leq x\leq u} \quad \max_{\tilde{U}:\ \mathrm{tr}\,\tilde{U}=1,\ \tilde{U}\geq 0} \langle \tilde{U}, A(x)\rangle.$$

(Here, as before, $x \in \Re^m$ and $\tilde{U} \in S\Re^{n\times n}$.) Now define a "dual problem"

$$\max_{\tilde{U}:\ \mathrm{tr}\,\tilde{U}=1,\ \tilde{U}\geq 0} \quad \min_{Cx=b;\ \ell\leq x\leq u} \langle \tilde{U}, A(x)\rangle.$$

The following theorem, motivated originally by [3], is a standard saddle point result and follows from [45, Thm. 36.3]. For closely related results, see [10] and [48].

THEOREM 6. *Suppose that $A(x)$ is an affine function, so that $A_k(x)$ is constant (independent of x) for all k. If the primal problem has a solution, say, defined by $(x^*, \tilde{U}^*)$, then the same pair solves the dual problem.*

Note that in the unconstrained affine case the dual problem can have a solution $\tilde{U}$ with corresponding objective greater than $-\infty$ only if

$$\langle \tilde{U}, A_k\rangle \ = \ 0, \qquad k = 1, \cdots, m.$$

Consequently, the dual version of the unconstrained affine primal problem is

$$\max\{\langle \tilde{U}, A(0)\rangle \ : \ \mathrm{tr}\,\tilde{U} = 1,\ \tilde{U} \geq 0,\ \langle \tilde{U}, A_k\rangle = 0,\ k = 1, \cdots, m\}. \tag{16}$$

3. Eigenvalue splitting and sensitivity analysis. The following theorem shows the importance of the eigenvalues of the $t \times t$ dual matrix U.

THEOREM 7. *Suppose that x, U, μ, and γ satisfy all the conditions* (11)–(15) *except possibly the semidefinite condition* (14), *and let κ be an eigenvalue of U with corresponding normalized eigenvector $v \in \Re^t$. If $d \in \Re^m, \delta \in \Re$ satisfy the following linear system of equations,*

$$\sum_{k=1}^{m} d_k Q_1^T A_k(x) Q_1 - \delta I = -vv^T, \tag{17}$$

$$Cd = 0, \tag{18}$$

$$d_k = 0 \quad \textit{if} \quad x_k = \ell_k \quad \textit{or} \quad x_k = u_k, \tag{19}$$

then d is a feasible direction with directional derivative

$$\lambda_1'(x; d) = \kappa.$$

Proof. It is clear that d is a feasible direction. The eigenvalues of the first matrix term on the left-hand side of (17) are, by construction, all equal to δ except one that has the value $\delta - 1$. It follows from Theorem 4 that the desired directional derivative has the value δ. Taking an inner product of U with both sides of (17) yields, using (12),

$$\sum_{k=1}^{m} d_k(\langle \mu, c_k\rangle + \gamma_k) - \delta = -\kappa,$$

i.e.,

$$\mu^T C d + \gamma^T d - \delta = -\kappa,$$

which gives, using (18)–(19) and (15),

$$\delta = \kappa. \qquad \square$$

This theorem was given in the unconstrained affine case by [36]. It was also explained there that for unconstrained problems, the multiplicity t of the multiple eigenvalue λ_1 is generically restricted by

$$\frac{t(t+1)}{2} \leq m+1, \tag{20}$$

the right-hand side being regarded as the number of degrees of freedom available. (The "1" reflects the fact that the value of the multiple eigenvalue is free.) This restriction is known as the von Neumann–Wigner crossing rule and is well known in quantum mechanics; it is further motivated in [15]. For problems with the linear constraints and bounds (11), it is clearly necessary to replace (20) by

$$\frac{t(t+1)}{2} \leq m+1-n_c-n_b, \tag{21}$$

where n_b is the number of active bounds, i.e., the number of variables x_k which are equal to either ℓ_k or u_k. Note, then, that with this nondegeneracy assumption on t, the linear system (17)–(19), which consists of $t(t+1)/2+n_c+n_b$ linear equations in $m+1$ variables, is generically solvable.

Theorem 7 shows how a descent direction may generically be computed in the event that a point x satisfies all the optimality conditions except the positive semidefinite condition on U. This direction splits the multiple eigenvalue into two clusters, one of unit multiplicity and one of multiplicity $t-1$, to first order. (See the discussion following Theorem 4.) Clearly, other splitting choices are possible; the one given here may be regarded as a generalization of the standard procedure for moving off constraints associated with multipliers of the wrong sign in linear or nonlinear programming, namely, moving off only one constraint at one time. Note that the coefficient matrix of the linear system (17)–(19) is the transpose of the coefficient matrix describing the active optimality conditions (12), (13), and (15).

Theorem 7 also shows how the eigenvalues of the dual matrix U describe the sensitivity of an optimal solution along directions that split the multiple eigenvalue λ_1 to first order. In particular, the theorem shows how to quantify first-order changes in λ_1 along these directions. If equality holds in (21), then, generically, all feasible directions in $\Re^m$ split the multiple eigenvalue to first order; in this case an optimal solution is characterized by first-order information and is generically "strongly unique." However, (21) cannot usually be expected to hold with equality, in which case there exists a nontrivial subspace of feasible directions d along which λ_1 does not split to first order, i.e., feasible directions tangent to the nontrivial manifold along which the eigenvalue retains multiplicity t. Since the function λ_1 is smooth along this manifold, it exhibits only second-order changes away from an optimal point x along these directions. The magnitude of these second-order changes is determined by the eigenvalues of the appropriate reduced Lagrangian Hessian, just as in nonlinear programming.

4. The successive quadratic programming algorithm. Let x^* be a local minimum of $\lambda_1(x)$; if $A(x)$ is affine, x^* is also a global minimum. Suppose that $\lambda_1(x^*)$ has multiplicity t^*. We wish to generate a sequence of iterates x^ν converging to x^*, but even if $t^* > 1$, $A(x^\nu)$ usually has distinct eigenvalues for any finite value of ν. (A similar remark applies to nonlinear programming problems; nonlinear constraints are generally both active and satisfied only in the limit.) In order for an algorithm to have good convergence properties, therefore, it is important for it to exploit the structure of the generalized gradient estimated to apply at the limit point, not just the gradient information at the current iterate. This observation is the basis for the so-called "ϵ-subgradient" methods found in [27], and similarly it is the estimated optimal active constraint structure that underlies successive quadratic programming (SQP) methods for nonlinear programming. In the latter case this estimated structure is usually defined by the active set found at the solution of the approximating quadratic program.

The algorithm presented in [36] takes full advantage of the structure of the generalized gradient that is estimated to apply at the optimal point. To do so, it requires an estimate of t^*, say t, which is obtained and revised as the algorithm proceeds. One way of doing this was suggested in [36], but more recent numerical experience suggests that a simpler approach is better. Let x be the current iterate, with $A(x)$ having eigenvalues

$$\lambda_1(x) \geq \cdots \geq \lambda_n(x),$$

with a corresponding orthonormal set of eigenvectors $\{q_i(x)\}$, and define t in terms of a tolerance τ by

$$(22) \quad \lambda_1(x) - \lambda_t(x) \leq \tau \max(1, |\lambda_1(x)|); \qquad \lambda_1(x) - \lambda_{t+1}(x) > \tau \max(1, |\lambda_1(x)|).$$

Define

$$(23) \qquad Q_1(x) = [q_1(x), \cdots, q_t(x)].$$

It will usually be necessary to adjust τ during the course of the minimization process.

The basic iteration of the method of [36] is defined by solving the following quadratic program (QP):

$$(24) \qquad \min_{d,\delta} \delta + d^T W d$$

subject to

$$(25) \; \delta I - \sum d_k Q_1(x)^T A_k(x) Q_1(x) = \mathrm{Diag}(0, \lambda_2(x) - \lambda_1(x), \cdots, \lambda_t(x) - \lambda_1(x)),$$

$$(26) \qquad \delta - \sum d_k q_i(x)^T A_k(x) q_i(x) \geq \lambda_i(x) - \lambda_1(x), \qquad i = t+1, \cdots, n$$

$$(27) \qquad \| d \|_\infty \leq \rho,$$

where d and δ are variables in $\Re^m$ and $\Re$, respectively; W is a positive definite matrix; and ρ is a trust region radius updated by the algorithm.

The motivation for the constraint (25) is that it results from linearizing a differentiable system of $t(t+1)/2$ nonlinear equations characterizing the condition $\lambda_1(x) =$

$\cdots = \lambda_t(x) = \omega$, for some $\omega \in \Re$. Actually, as was pointed out by [56], the form of the nonlinear system given by (4.1) of [36] is not correct. The correct system uses a matrix exponential formulation based on Theorem 3.1 of [15], as is explained in more detail in [39]. Constraints (26) ensure that linearizations of $\lambda_{t+1}, \cdots, \lambda_n$ give values no greater than the linearized value for the approximate multiple eigenvalue $\lambda_1, \cdots, \lambda_t$. Both (26) and (27) prevent d from having too large a norm, particularly during the early part of the iteration. Ideally, they will not be active near the solution.

The constraint (25) is imposed as $t(t+1)/2$ scalar constraints, each of which has a QP multiplier associated with it. These multipliers make up the QP dual matrix estimate U, with diagonal elements of U equal to the corresponding multipliers for the diagonal equations in (25) and off-diagonal elements of U equal to half the corresponding multipliers for the off-diagonal equations in (25).

Constraints on the variables were not considered in [36] for simplicity, but let us explicitly include linear constraints and bounds in the present discussion, i.e., address the problem (10)–(11). Assume that the present iterate x satisfies (11); then the corresponding restrictions that should be added to the QP are

$$Cd = 0, \tag{28}$$

$$\ell \leq x + d \leq u. \tag{29}$$

The following theorems clarify some points that were not made in [36].

THEOREM 8. *Suppose the quadratic program* (24)–(29) *has solution d, δ with the property that constraints* (26)–(27) *are not active. Then the solution has an associated dual matrix U and vectors of multipliers μ and γ satisfying*

$$(Wd)_k \;+\; \langle U, Q_1^T A_k(x) Q_1 \rangle \;=\; \langle \mu, c_k \rangle \;+\; \gamma_k, \qquad k = 1, \cdots, m, \tag{30}$$

$$\operatorname{tr} U = 1, \tag{31}$$

and

$$\begin{array}{lcll} \gamma_k & = & 0 & \textit{if } \ell_k < x_k + d_k < u_k; \\ \gamma_k & \geq & 0 & \textit{if } x_k + d_k = \ell_k; \\ \gamma_k & \leq & 0 & \textit{if } x_k + d_k = u_k. \end{array} \tag{32}$$

Furthermore, U, μ, and γ are unique if the $t(t+1)/2 + n_c$ linear constraints (25), (28) *on d, δ, together with the active bound restrictions on d, are linearly independent.*

Proof. The proof follows immediately from the standard optimality conditions for quadratic programs (see, e.g., [17]). □

THEOREM 9. *Assume $\tau = 0$, so that $\lambda_1(x)$ has exact multiplicity t. The quadratic program* (24)–(29) *yields a vector d, which is a descent direction for λ_1, unless $d = 0$. Furthermore, if $\rho > 0$, then $(d = 0, \delta = 0)$ solves the QP if and only if* (12), (13), *and* (15) *are satisfied for some U, μ, and γ, i.e., the optimality conditions* (12)–(15) *are satisfied, with the possible exception of the positive semidefinite condition on U.*

Proof. By (25) combined with Theorem 4, we have

$$\lambda_1'(x; d) = \delta. \tag{33}$$

Also, the QP solution (d, δ) satisifies

$$\delta + \frac{1}{2} d^T W d \leq 0$$

since the value zero is achievable with $(d = 0, \delta = 0)$. Thus

$$\lambda_1'(x; d) \leq -\frac{1}{2} d^T W d.$$

Since W is positive definite, the right-hand side is nonpositive, with zero value if and only if $d = 0$. The last statement follows from Theorem 8. □

If it happens that $d = 0$, so that the optimality conditions are satisfied with the possible exception of the positive semidefinite condition on U, and if indeed U has a negative eigenvalue, then it is necessary to split the multiple eigenvalue $\lambda_1(x)$ as explained in Theorem 7 in order to obtain a decrease in the maximum eigenvalue. In nonlinear programming, an analogous situation occurs when x satisfies all optimality conditions except the sign constraints on the Lagrange multipliers.

Whether d is zero or not, (30)–(31) define a matrix U which is unique as long as the active constraint gradients of the QP are linearly independent. (Note that (21) is a necessary condition for such independence.) If the dual matrix estimate U generated by the QP is not positive semidefinite, this is a clear indication that the multiplicity estimate t is too large and that the tolerance τ should be reduced if possible. This strategy is used in the current version of our programs. Consequently, we do not generally expect to converge to points x where it is necessary to split a multiple eigenvalue. This is indeed the case in practice, with the notable exception of the graph problems to be described in §11.

THEOREM 10. *Suppose that the QP* (24)–(29) *yields a solution d, δ with the property that the constraints* (26) *are not active, and suppose that U defined by* (30)–(31) *is positive semidefinite. Then d is a descent direction for λ_1 (unless d = 0), regardless of the value of τ.*

Proof. The exact multiplicity of $\lambda_1(x)$ is less than or equal to the multiplicity t defined by (22). Consequently, (33) holds, just as in Theorem 9. However, $(d = 0, \delta = 0)$ does not generally satisfy (25). Let

$$E\bar{d} = e$$

represent the combined linear system (25) and (28), where $\bar{d} = (d^T, \delta)^T$. It follows that equations (30)–(31) may be written

$$\begin{bmatrix} Wd \\ 1 \end{bmatrix} = E^T v + \begin{bmatrix} \gamma \\ 0 \end{bmatrix},$$

where $v = (U_{11}, 2U_{12}, \cdots, U_{tt}, \mu_1, \cdots, \mu_{n_c})$. (Actually, this system needs modification if the trust radius constraint (27) is active, but this is easily done by modifying the corresponding lower and upper bounds ℓ_k or u_k to impose the trust radius bound.) Taking an inner product with $\bar{d}$ we have

$$d^T W d + \delta = v^T e + \gamma^T d.$$

We have $\gamma^T d \leq 0$ by (32) together with feasibility of x. Since e has nonpositive entries corresponding to diagonal elements of U in v and zero entries elsewhere, we therefore have

$$\delta \leq 0$$

from the semidefiniteness of U and W, with $\delta = 0$ only if $d = 0$ (since W is positive definite). □

It follows that if the dual matrix estimate U is positive semidefinite and τ and ρ are both sufficiently small,

$$\lambda_1(x + d) < \lambda_1(x), \tag{34}$$

provided d is nonzero. (If τ is too large relative to ρ the QP may not be feasible, while if ρ is too large, $\| d \|$ may be too large for the negative directional derivative to yield (34).) The best automatic way to adjust τ and ρ is not clear, but in practice, given a reasonable estimate for τ, obtaining the reduction (34) by decreasing ρ is usually straightforward unless λ_1 is very near its optimal value. Provided (34) holds, the new iterate may be set to $x + d$. (The difficulty of a possibly infeasible subproblem is eliminated in the large-scale algorithm described in the next section.)

It is explained in [36] that, in order to obtain a quadratically convergent method, W should be set to the Hessian of the appropriate Lagrangian function. We emphasize that W is not the Hessian of the max eigenvalue function, which does not exist at x^* if $t^* > 1$. The correct form of the Lagrangian is not (4.9) of [36], but a modification using the matrix exponential formulation mentioned above. The formula for W given by (4.12) of [36] is correct. Its derivation was omitted, but it is given in [39]. In the case $t = 1$, the formula reduces to a fairly well known expression for the second derivative of a distinct eigenvalue; see [26], [20]. In the case that $A(x)$ is nonlinear, an additional term

$$Q_\ell^T \frac{\partial^2 A}{\partial x_j \partial x_k} Q_\ell$$

must be added to (4.12) of [36], assuming that $A(x)$ is twice continuously differentiable.

We make here an observation not made in [36], namely, in some cases the reduction condition (34) may not hold for ρ large enough that (27) is inactive, even when W is set to the correct Hessian matrix and x is very close to an optimal solution. Such a situation is known as the Maratos effect and it prevents quadratic convergence of the algorithm, since the trust radius ρ must be reduced until it yields (34). This difficulty has indeed occurred on some of our test problems, but it has been overcome by implementing Fletcher's second-order correction technique, making use of our knowledge of the Hessian matrix W to avoid additional gradient evaluations, as does Fletcher in [12].

Clearly, it is important to develop a precise version of the algorithm for which global convergence can be guaranteed. As yet, we have not attempted to do this, but we do not see any inherent difficulty. Trust region convergence proofs are by now rather well understood; the essential ingredients in this case are given by the theorems above.

The SQP algorithm summarized in this section has been used to solve a wide variety of problems, some of which will be mentioned in later sections of the paper. Our Fortran implementations use Eispack subroutines [50] to obtain the eigenvalues and eigenvectors of each matrix $A(x)$ and either the Stanford code LSSOL [16] or the equivalent NAG routine [32] to solve the quadratic programs. Using current workstation technology, only a moderate amount of computer time is typically required to obtain a very accurate solution, including verification of the optimality conditions, for, say, $\max(n, m) \leq 40$. However, the algorithm is very inefficient for much larger values of n, m. The next two sections discuss how to modify the algorithm for large-scale problems.

5. The optimization algorithm when m is large. In this section we discuss our approach to modifying the successive quadratic programming algorithm when m is large, say, $m > 40$. The discussion of how to efficiently compute the eigenvalues when n is also large is deferred to the next section.

The first observation is that the benefits of quadratic convergence are far outweighed by the cost of computing and factoring the Lagrangian Hessian W when m is large. We shall therefore consider a first-order algorithm based on successive linear programming instead of successive quadratic programming, replacing W by zero in (24). First-order algorithms, which generally converge at a first-order rate, can be very satisfactory in some applications; in other cases they can be excruciatingly slow. If it happens that equality holds in (21), then, generically, the solution is "strongly unique," which implies that a first-order method is quadratically convergent. However, this is not generally to be expected.

A successive linear programming method retains the key feature of the SQP algorithm of [36], namely, the algorithm estimates the eigenvalue multiplicity t and uses the appropriate $t(t+1)/2$ linear constraints to approximate the condition $\lambda_1(x+d) = \cdots = \lambda_t(x+d) = \omega$, generating the corresponding $t \times t$ dual matrix U. Consequently, verification of the optimality conditions for the model problem is possible. We have the following theorem.

THEOREM 11. *Assume that $\tau = 0$, so that $\lambda_1(x)$ has exact multiplicity t. Then the linear program* (24)–(29), *where $W = 0$, yields a vector d, which is a direction of nonascent for $\lambda_1(x)$. Furthermore, if $\rho > 0$, then $(d = 0, \delta = 0)$ is a (not necessarily unique) solution of the linear program if and only if* (12), (13), *and* (15) *are satisfied for some U, μ, and γ.*

Proof. The proof is a straightforward modification of the proof of Theorem 9. The solution $(d = 0, \delta = 0)$ cannot be unique when $t(t+1)/2 + n_c + n_b < m+1$, since it is not a vertex of the feasible region. □

However, even solving the linear program (24)–(29), where $W = 0$, is not a justifiable expense when m is large, especially if $t(t+1)/2 + n_c + n_b << m$, which is usually the case. Usually the LP has only a few active general linear constraints, i.e., (25) and (28), so that obtaining a vertex solution requires most of the elements of d to be on their bounds. Often, aside from perhaps a few "genuine" active bounds arising in (29), most of the active bounds are trust radius bounds in (27). If the simplex method is used to solve the LP, most of the work involves finding the active set of bounds. Since there are only a few general linear constraints, the work per simplex step need only be $O(m)$, but $O(m)$ steps are required. This is not acceptable, especially since the exact set of active trust radius bounds is of little importance; the purpose of the trust radius is simply to restrict d so that its norm is not too large.

In view of these remarks we have implemented the following "partial linear programming" solver. (For a related idea, see [23].)

PLP Algorithm to partially solve the LP

$$\min g^T \bar{d} \tag{35}$$

subject to

$$E\bar{d} = e, \tag{36}$$

$$F\bar{d} \geq f, \tag{37}$$

$$\bar{d}_k = 0, \qquad k \in K, \tag{38}$$

$$\bar{\ell} \leq \bar{d} \leq \bar{u}, \tag{39}$$

$$\| \bar{d} \|_\infty \leq \rho, \tag{40}$$

where $\bar{d} = (d^T, \delta)^T \in \Re^{m+1}$, K is an index set, and $g, E, e, F, f, \bar{\ell}, \bar{u}$ are defined so that (35)–(40) is equivalent to (24)–(29), with $W = 0$, except that the additional constraints (38) have been introduced (for reasons to be explained shortly) and that, for convenience, the trust radius restriction applies to $\bar{d}$ instead of d. Thus

$$g = [0, \cdots, 0, 1]^T;$$

E and e, respectively, contain the $t(t+1)/2$ rows

$$[-q_i(x)^T A_1(x) q_j(x), \cdots, -q_i(x)^T A_m(x) q_j(x), \delta_{ij}]; \qquad \delta_{ij}(\lambda_i(x) - \lambda_1(x)),$$

$1 \leq i \leq j \leq t$ (where δ_{ij} is the (i,j) element of the identity matrix), together with the additional n_c rows

$$[C \ 0]; \ \ 0;$$

F and f contain the rows

$$[-q_i(x)^T A_1(x) q_i(x), \cdots, -q_i(x)^T A_m(x) q_i(x), 1]; \quad \lambda_i(x) - \lambda_1(x),$$

$i = t+1, \cdots, n$; and

$$\bar{\ell} = [(\ell - x)^T, -\infty]^T; \qquad \bar{u} = [(u - x)^T, \infty]^T.$$

It is assumed that $\ell \leq x \leq u$, so that $\bar{\ell} \leq 0$, $\bar{u} \geq 0$. Note also that $f \leq 0$, so $\bar{d} = 0$ satisfies all constraints except (36). It is assumed that $t(t+1)/2 + n_c << m$. It is not necessary to store or even fully compute the derivative matrices $A_k(x)$; rather, a subroutine is required to perform the matrix vector product $A_k(x)q$ for given index k and vector q.

Step 0. Set $\nu = 0$. Set $\bar{d}^0$ to the least norm solution of the underdetermined linear system (36), (38). This is obtained by a QR factorization of G, a matrix defined initially to contain the columns of E^T, with rows corresponding to the indices in K removed. Let the QR factorization of G be given by

$$G = YR,$$

where R is upper triangular and Y, which has the same dimensions as G, satisfies $Y^T Y = I$. Then solve

$$R^T d_Y = e$$

and set

$$\tilde{d}^0 = Y d_Y, \tag{41}$$

the least norm solution of $G^T \tilde{d} = e$. Set $\bar{d}^0$ to the vector containing $\tilde{d}^0$ interspersed with zeros corresponding to the entries in K. (We use the Linpack software for computing the QR factorization; the range space basis Y is stored only as a product of Householder transformations. For details, see [8] and, for information on how to update the factorization and use it in the context of optimization, see [17].) Then set $\bar{d} = \alpha^0 \bar{d}^0$, where α^0 is defined as follows. If $\bar{d}^0$ is a feasible point for the LP, set $\alpha^0 = 1$. Otherwise, if $\bar{d}^0$ violates the constraints (37), the bounds (39) or the trust radius restriction (40), set α^0 to the maximum value possible so that $\bar{d}$ satisfies (37)–(40). (This effectively modifies the equality constraints of the LP. The rationale here is that if the least norm step to the equality constraints of the LP is infeasible, most likely the approximations underlying the definition of the LP are not good enough to justify its solution, should it indeed have a feasible solution.)

Step 1. Let $\tilde{g}$ be g with rows corresponding to the indices in K removed. Set $\tilde{d}$ to the least squares projection of $\tilde{g}$ onto the null space of G^T. This is obtained by using the QR factorization of G to solve the least squares problem

$$\min_v \parallel Gv - \tilde{g} \parallel_2, \tag{42}$$

i.e., solving

$$Rv = Y^T \tilde{g},$$

and setting $\tilde{d}$ to the residual $Gv - \tilde{g}$. Note that a null space basis is *not* computed. If $\parallel \tilde{d} \parallel \leq \epsilon$, go to Step 3. Otherwise increment ν, and set $\bar{d}^\nu$ to the vector containing $\tilde{d}$ interspersed with zeros corresponding to the entries in K.

Step 2. Compute the maximum step α^ν so that

$$\bar{d} = \sum_{\kappa=0}^{\nu} \alpha^\kappa \bar{d}^\kappa \tag{43}$$

satisfies the constraints of the LP, consequently making a new general linear constraint or bound active. In the former case, append the corresponding row of F as a new column of G. In the latter case, if the new active bound is one of the bounds in (39), add the corresponding index to K and remove the corresponding row from G. In either of these cases, update the QR factorization of G accordingly and go back to Step 1. Finally, if the new active bound is one of the bounds in (40), go to Step 3.

Step 3. Set v to the final vector of constraint multipliers, by permuting the elements of the last solution of (42) to correspond to the row order in E and F, interspersing zeros corresponding to inactive constraints in (37). Set γ to the final vector of bound multipliers, by setting

$$\gamma_k = (g - [E^T \; F^T]v)_k, \qquad k \in K,$$

and $\gamma_k = 0$ otherwise. (See [17, p. 189].) Exit with $\bar{d}$ (defined by (43)), v, γ, and K.

The basic idea of the PLP algorithm is that once one active trust radius bound is encountered, there is little to be gained by going through the computationally expensive process of adding all the other active trust radius bounds making up a vertex solution to the LP. Of course, since the PLP method neither checks multiplier signs nor allows a constraint or bound, once active, to become inactive, it will not generally produce an optimal solution of the LP.

Note that when $t = 1$, $\bar{d}^0 = 0$ and the vector consisting of the first m components of $\alpha^1\bar{d}^1$, say $\alpha^1 d^1$, is the steepest descent step for the differentiable function $\lambda_1(x)$, projected to satisfy the linear constraints $Cd = 0$ and (38), and with steplength restricted by (37), (39), or (40) (if the last case applies, the algorithm terminates immediately with $d = \alpha^1 d^1$). When $t > 1$, the algorithm certainly does not yield a steepest descent direction; such a direction would violate (25) and hence split the current approximate multiple eigenvalue. However, the first m components of $\bar{d}^1$ may be viewed as a projected steepest descent direction, where by this we mean projected to satisfy the additional $t(t+1)/2 - 1$ conditions in (36).

The selection rule for α^0 in Step 0 eliminates one potential difficulty with the SQP method, namely, the possibility of an infeasible subproblem.

Instead of using the PLP algorithm, which is based on QR factorizations of matrices with a small number of columns, an alternative approach would be to use an affine scaling interior point method to partially solve the LP.

We now define the successive partial linear programming (SPLP) method whose purpose is to solve the constrained model problem when m is large by a sequence of calls to the PLP algorithm. Each of these calls partially solves an LP of the form (35)–(40). The number of equality constraints in (36) is determined by the multiplicity estimate t. As with the SQP algorithm, the hope is that, once t is determined correctly, the inequality constraints (37) will become permanently inactive. However, since bounds in (11) may be active at a solution x^*, it is not adequate to begin the PLP algorithm with all bounds on the elements of d inactive, since then the same active set of bounds would have to be repeatedly built up every time the PLP algorithm is executed. This inefficiency is avoided by the use of the bound active set K. Bounds are added to K when they are encountered during a PLP execution; they are removed from K after a PLP execution if the corresponding multiplier signs indicate that they should not be active. Also, if the dual matrix U defined by the multipliers characterizing a PLP "solution" is indefinite, the multiplicity tolerance τ is reduced. The updating of the trust radius ρ is based on recommendations in [13].

SPLP Algorithm to solve (10)–(11).

Step 0. Initialize the trust radius ρ and the multiplicity tolerance τ. Define a convergence tolerance ϵ. Set x to an initial value satisfying (11). Compute the eigenvalues and eigenvectors of $A(x)$. Initialize K to the empty set.

Step 1. Define the multiplicity estimate t and associated block of eigenvectors Q_1 by (22)–(23). Set $K' = K$. Partially solve the LP (35)–(40), using the PLP Algorithm, producing $\bar{d} = (d^T, \delta)^T, v, \gamma$, and (a possibly modified) K.

Step 2. Construct U and μ from v, by setting diagonal elements of U to corresponding multipliers for diagonal equations of (25), off-diagonal elements of U to half the corresponding multipliers for the off-diagonal equations of (25), and elements of μ to corresponding multipliers for the constraint $Cd = 0$. If U is not positive semidefinite, reduce τ by a factor of two. If $\| d \| \leq \epsilon$, go to Step 5.
Step 3. Compute the eigenvalues of $A(x + d)$. If $\lambda_1(x + d) \geq \lambda_1(x)$, then set $K = K'$, divide ρ by two, and go to Step 1.
Step 4. Define

$$\psi = \frac{\lambda_1(x) - \lambda_1(x + d)}{-\delta},$$

the ratio of the actual to predicted reduction in the minimization objective. If $\psi > 0.75$, double ρ; if $\psi < 0.25$, divide ρ by two. Compute the eigenvectors of $A(x + d)$, if they were not already obtained, and replace x by $x + d$. If γ does not satisfy (32), remove indices from K corresponding to violated bounds in (32). Go to Step 1.
Step 5. If U is positive semidefinite and γ satisfies (32), stop. If U is not positive semidefinite, then obtain a reduction in λ_1 by splitting the multiple eigenvalue $\lambda_1(x) = \cdots = \lambda_t(x)$, as explained in Theorem 7; then reduce τ by a factor of 10 and go to Step 1. Otherwise, if γ violates (32), remove indices from K corresponding to violated bounds in (32), divide ρ by two, and go to Step 1.

The following theorem provides one justification for the SPLP method; to avoid unnecessary complication, some simplifying assumptions are made.

THEOREM 12. *Suppose that the PLP algorithm called by Step* 1 *of the SPLP method generates* $\bar{d} = (d^T, \delta)^T$ *with the property that*

$$\bar{d} = \alpha^0 \bar{d}^0 + \alpha^1 \bar{d}^1,$$

i.e., no bound in (39) *or constraint in* (37) *becomes active. Suppose also that U defined by the subsequent Step* 2 *of the SPLP method is positive semidefinite. Then d is a direction of nonascent for* λ_1.

Proof. By construction, we have $E\bar{d}^0 = e$, $E\bar{d}^1 = 0$, so $E\bar{d} = \alpha^0 e$. Therefore, by the same argument used in Theorem 10, (33) holds. We therefore wish to show that $\delta = \bar{d}_{m+1}$ is nonnegative. Let us first look at the second term of $\bar{d}$. We have

$$(\bar{d}^1)_{m+1} = \tilde{g}^T \tilde{d} = -\tilde{g}^T (I - YY^T) \tilde{g} \leq 0,$$

where $\tilde{g}$, $\tilde{d}$ are defined by Step 1 of the PLP algorithm, since YY^T is the orthogonal projector onto the range space of G and $I - YY^T$ is the orthogonal projector onto the null space of G^T. Now consider the first term of $\bar{d}$. We have

$$Gv = YY^T \tilde{g},$$

so taking an inner product with (41) gives

$$v^T e = \tilde{g}^T Y d_Y = \tilde{g}^T \bar{d}^0 = (\bar{d}^0)_{m+1}.$$

The proof is now complete, since $v^T e \leq 0$ for the same reason as given in Theorem 10. □

Theorem 12 is based to some extent on [31, Thm. 4]; as a point of comparison, note that the dual matrix estimate U generated by the SPLP method is obtained from least squares approximation.

We expect that the algorithm described above will be modified in the future with further computational experience and theoretical development. In particular, we have no theoretical guarantee that the algorithm will converge to an optimal solution; we have not yet attempted any convergence analysis. However, in its present form, the algorithm has been used to obtain very satisfactory solutions to the problems to be described in §§9, 10, and 11.

Although it is not practical to compute W when m is large, we note that the SPLP method can probably be improved by approximating the second-order information in some way. The expression for W given by (4.12) of [36] is actually a sum of terms, one corresponding to each eigenvalue smaller than λ_t. Since the denominator of each term is the separation of the eigenvalue from λ_1, one idea is to approximate W by a low-rank approximation, consisting of terms corresponding to eigenvalues immediately lower than λ_t. It is not clear exactly how the low-rank approximation would be exploited, but note that an SLP method may be regarded as an SQP method with a zero-rank approximation to the quadratic term. An alternative idea is to approximate W using a limited memory quasi-Newton method; see [28]. In either case it seems probable that a practical SQP method could be devised that would converge faster than the SPLP method unless it had difficulty identifying the optimal multiplicity t^*.

We complete this section by noting that if n is large, the number of inequalities in (26), and therefore (37), should be substantially reduced. Indeed, as discussed in the next section, it is not practical to compute all the eigenvalues of $A(x)$ when n is large.

6. Computation of the eigenvalues when n is large. When n is large, the QR algorithm used by Eispack is not an efficient way to solve the eigenvalue problem. Indeed, it is particularly inappropriate for our purposes for two reasons:

1. Since only the largest eigenvalues are of any relevance to the optimization, it is grossly inefficient to compute all the eigenvalues of each matrix iterate $A(x)$.
2. Typically, each matrix iterate $A(x)$ generated by the optimization calculation does not differ much from the previous matrix iterate, whose eigenvalues and eigenvectors have already been computed.

For both of these reasons, it is clear that the eigenvalues should be computed by an iterative method. Possibilities are power methods, inverse power methods, and Lanczos methods. The best choice depends on a number of considerations. In all cases, however, it is essential to iterate with a block of r vectors, which are approximate eigenvectors for $\lambda_1, \cdots, \lambda_r$, where $r \geq t^*$, the multiplicity of λ_1 at the optimal solution. Otherwise it will not be possible to verify the multiplicity t^* or to generate the dual matrix U. Indeed, unless an a priori upper bound on t^* is known, it is necessary that $r > t^*$ to be sure that the correct multiplicity is calculated. The number r can be adjusted during the iteration according to the value of the current multiplicity estimate t. It is important to maintain orthogonality of the r vectors during the iteration. The orthogonalized block versions of the power and inverse power methods are generally called subspace iteration; see Parlett [42] for details. The block of eigenvectors computed for the *previous* matrix iterate is a very valuable starting block for each subspace iteration after the first few optimization steps.

The simplest variation of subspace iteration is that based on the ordinary power method, which requires repeated multiplication of $A(x)$ onto the block of approximate eigenvectors. To be applicable, it is necessary that $\lambda_r(x) > | \lambda_n(x) |$; usually this method is used only when $A(x)$ is positive definite. The convergence of $\lambda_i(x), 1 \leq i \leq r$, depends on the separation of its magnitude from $\lambda_{r+1}(x)$. In particular, convergence of $\lambda_1(x)$ is fast if

$$(44) \qquad \lambda_1(x) >> \lambda_{r+1}(x); \qquad \lambda_1(x) >> | \lambda_n(x) | .$$

Whether or not (44) holds, a Lanczos method generally converges faster than the power method. However, a block Lanczos method is needed, for the reason just explained. We have not tried using block Lanczos since the necessary software has not advanced beyond an experimental stage.

Suppose now that $| \lambda_1(x) | < | \lambda_n(x) |$. This happens, in particular, if $A(x)$ is negative definite; equivalently, the optimization objective is to maximize the smallest eigenvalue of the positive definite matrix $-A(x)$, as in the column problem to be discussed in §9. In this case, an inverse block power method (subspace iteration) is appropriate. Convergence of $\lambda_1(x)$ is fast if

$$| \lambda_1(x) | << | \lambda_{r+1}(x) |; \qquad | \lambda_1(x) | << | \lambda_n(x) | .$$

This is the case for the column problem. The inverse power method, unlike the power method, requires factorization of $A(x)$ at each step of the optimization iteration, i.e., once per subspace iteration, as well as two triangular "solves" at each step of the subspace iteration.

If the power or inverse power methods converge slowly an attractive alternative is the shifted inverse power method, commonly known as inverse iteration. As before, the iteration must be carried out on a block of vectors. Each step requires the block of vectors to be multiplied by the inverse of $sI - A(x)$; this is implemented by a factorization of $sI - A(x)$ and several triangular "solves." An excellent shift s is available, namely, the value of λ_1 from the previous matrix iterate. After the first few optimization steps, the shift is usually so good that only one shifted inverse multiplication is needed. If $sI - A(x)$ is discovered not to be positive definite during its factorization, the iterate x may be rejected immediately and the optimization trust radius ρ reduced, since $\lambda_1(x)$ is necessarily greater than the previous value s. This is a very valuable observation.

Whatever iterative method is chosen to compute the eigenvalues, caution must be used. In particular, if the iteration is terminated too soon with an inaccurate underestimate of λ_1, which is lower than the previous best value, the optimization algorithm may be unable to obtain a further reduction in $\lambda_1(x+d)$ for the simple reason that its estimate of $\lambda_1(x)$ is wrong. Thus a good implementation of the algorithm needs to allow recomputation of $\lambda_1(x)$ when necessary. We have not yet incorporated this automatically, instead restarting the algorithm when necessary. This is usually needed only at the beginning of the optimization if shifted inverse iteration is used, since the excellent choice of shift available makes this method very accurate. Note one fortunate fact: whatever form of block iteration is used, it is λ_1 that is the most accurately computed of $\lambda_1, \cdots, \lambda_r$; this is the eigenvalue whose accuracy is the most critical.

If factorizations are not practical, inverse or shifted inverse subspace iteration is still possible by the incorporation of a third nested iteration for, e.g., the conjugate gradient method to solve the linear systems required for each step of each subspace iteration. In the case of shifts, this inner iteration may be terminated if indefiniteness is detected, for the same reason as explained above. We note, however, that the performance of the conjugate gradient method on the nearly singular systems that result from a good choice of shift is not very well understood. Most of our numerical experiments have used factorizations but some (not very extensive) experiments with a conjugate gradient version suggest that the method may give poor results when the shift is good, perhaps because of instability resulting from the near singularity of $sI - A(x)$. An alternative idea, following Szyld [52], is to use the Paige and Saunders method SYMMLQ [40]. This may give better results than conjugate gradient for nearly singular positive definite systems. Szyld gives an argument explaining why the near singularity does not cause difficulty for SYMMLQ; he did not consider the conjugate gradient method, since he was concerned with interior eigenvalues and therefore needed to operate with indefinite systems. However, the disadvantage of using SYMMLQ is that the shifted inverse iteration may converge to a subdominant eigenvalue, since the iteration is not terminated when $sI - A(x)$ is indefinite. We have not yet experimented with a preconditioned conjugate gradient method, for example, using a factorization of an earlier matrix iterate for a number of steps of the optimization.

7. The generalized eigenvalue problem. All of the preceding sections may easily be generalized to apply to the eigenvalue problem

$$A(x)q = \lambda Bq, \tag{45}$$

where B is a symmetric positive semidefinite matrix independent of x, not necessarily the identity matrix, as has been implicitly assumed up to this point. We have the following modifications to Lemma 1 and Theorem 1 (proofs are omitted).

LEMMA 2. *Let Q be a matrix $\in \Re^{n\times n}$ such that*

$$Q^TBQ = I. \tag{46}$$

Then the convex hull of the set

$$\{ww^T : w \in \Re^n, w^TBw = 1\}$$

is the set

$$\{\tilde{U} = Q\hat{U}Q^T : \ \hat{U} \in \Re^{n\times n},\ \hat{U} = \hat{U}^T,\ \operatorname{tr}\hat{U} = 1,\ \hat{U} \geq 0\}.$$

Furthermore, the elements in the first set are the extreme points of the second set.

Note that the trace of $\tilde{U}$ is generally not equal to one.

THEOREM 13. *As above, let Q be any matrix $\in \Re^{n\times n}$ satisfying $Q^TBQ = I$. Now let $\lambda_1(A,B)$ denote the largest eigenvalue of the pencil (A,B), i.e., largest root λ of* (45) *for nontrivial q, ignoring for the moment the dependence of A on x. The following characterizations of λ_1 hold:*

$$\begin{aligned}\lambda_1(A,B) &= \max\{\langle q, Aq\rangle \ : q^TBq = 1\};\\ \lambda_1(A,B) &= \max\{\langle qq^T, A\rangle \ : \ q^TBq = 1\};\end{aligned}$$

$$\lambda_1(A,B) = \max\{\langle\tilde{U}, A\rangle \ : \ \tilde{U} = Q\hat{U}Q^T,\ \hat{U} \in S\Re^{n\times n},\ \operatorname{tr}\hat{U} = 1,\ \hat{U} \geq 0\}. \tag{47}$$

Now take $Q = [q_1, \cdots q_n]$ to be a matrix of eigenvectors of (A, B), normalized so that (46) holds. Thus, in addition to (46), we have

$$Q^T AQ = \mathrm{Diag}(\lambda_i).$$

Assume that the largest eigenvalue λ_1 has multiplicity t, with corresponding eigenvectors, $q_1, \cdots, q_t$ making up a matrix $Q_1 \in \Re^{n \times t}$. We see then that the set of matrices achieving the max in (47) is, as before, the right-hand side of (6). Indeed, Theorem 2 and all subsequent theorems, remarks, and algorithm statements then apply exactly as before provided only that *the normalization* (46) *is consistently used for the eigenvectors.*

Note that the details of subspace iteration are well known for the generalized problem; see [2], [42, Chap. 15]. If a shift s is used, it is of course understood that $A(x)$ is to be shifted by sB instead of sI.

8. Several matrix families. Suppose it is desired to minimize

$$\phi(x) = \max_{1 \le l \le p} \lambda_1^{(l)}(x) \tag{48}$$

subject to (11), where each $\lambda_1^{(l)}(x), l = 1, \cdots, p$, is the largest eigenvalue of a matrix-valued function $A^{(l)}(x)$. The necessary optimality conditions are easily extended to this case by introducing a dual matrix for each matrix family. Given x, let t_l be the multiplicity of $\lambda_1^{(l)}(x)$ if the latter quantity equals $\phi(x)$, and zero otherwise. Let $Q_1^{(l)}$ be an orthonormal set of t_l corresponding eigenvectors if $t_l > 0$, and the empty matrix otherwise.

THEOREM 14. *A necessary condition for x to solve* (48), (11) *is, in addition to* (11), *that there exist dual matrices $U^{(l)} \in S\Re^{t_l \times t_l}$, $l = 1, \cdots, p$, and vectors of Lagrange multipliers $\mu \in \Re^{n_c}$ and $\gamma \in \Re^m$, satisfying*

$$\sum_{l=1}^{p} \langle U^{(l)}, (Q_1^{(l)})^T A_k(x) Q_1^{(l)} \rangle \;=\; \langle \mu, c_k \rangle \;+\; \gamma_k, \qquad k = 1, \cdots, m, \tag{49}$$

$$\sum_{l=1}^{p} \operatorname{tr} U^{(l)} = 1, \tag{50}$$

$$U^{(l)} \ge 0, \qquad l = 1, \cdots, p, \tag{51}$$

as well as (15). *The necessary condition is also sufficient in the affine case.*

The proof is a straightforward generalization of the proof of Theorem 5.

Similarly, the SQP and SPLP algorithms are easily adapted to minimize $\phi(x)$ by including, in the QP or LP, constraints of the form (25)–(26) for each of the p matrix families. Multiplicity estimates $t_l, l = 1, \cdots, p$, may be defined as the largest integer t such that

$$\phi(x) - \lambda_t^{(l)}(x) \le \tau \max(1, |\,\phi(x)\,|),$$

with $t_l = 0$ if no positive integer satisfies the inequality. Note that it is *not* recommended to simply define $A(x)$ to be a block diagonal matrix with blocks $A^{(l)}(x), l = 1, \cdots, p$. Such an approach loses some of the structure of the generalized gradient of $\phi(x)$, since it does not take account of the fact that eigenvalues corresponding to different diagonal blocks of a block diagonal matrix do not interact with each other.

One application of (48) is minimizing the maximum eigenvalue of $A(x)$ in absolute value by taking $A^{(1)}(x) = A(x), A^{(2)}(x) = -A(x)$; see [36] as well as §10 below.

9. The column problem. A classical problem that goes back to Lagrange is to find the shape of the strongest column with given volume. Mathematically, the problem is to determine a function $\sigma(x)$, the cross-sectional area of the column, from an admissible set

$$\sigma \in L^{\infty} : 0 < \alpha \leq \sigma(x) \leq \beta, \qquad \int_0^1 \sigma(x)dx = 1 \tag{52}$$

to maximize the least eigenvalue of

$$-\left(\sigma^p(x)u''(x)\right)'' = \lambda u''(x), \qquad u \in H_0^2, \tag{53}$$

on the interval $[0, 1]$, where $p \geq 1$ (usually $p = 1$ or $p = 2$). Here p has a different meaning from the previous section and x refers not to unknown parameters but to a spatial dimension along the axis of the column. The function $u(x)$ measures the displacement of the column when deflected from its equilibrium position. The case $p = 2$ models columns with circular (or equivalently square) cross-sections of uniform material. The case $p = 1$ models "thin-wall" beams or columns, where a variable thickness shell of one kind of material surrounds a uniform core of another material. The significance of the least eigenvalue of the differential equation is that it corresponds to the critical buckling load in the Euler–Bernoulli model of the column. (The load is applied at the ends of the column, in the direction of its axis.)

The problem is a controversial one that has been addressed by many applied mathematicians and structural engineers, including [53] and [33]. Our work on this problem is a joint effort with Steve Cox; the details of our theoretical and computational contributions may be found in [6]. Here we briefly summarize some of the computational results. We discretized the problem, approximating $\sigma(x)$ by a piecewise constant function σ_h, where h is the mesh size. Following the standard approach in [51], we approximated u by u_h, using piecewise cubic Hermite finite elements, and constructed the corresponding finite-dimensional bending matrix $A(\sigma_h)$ and stiffness matrix B such that the eigenvalues of the generalized problem (45),

$$A(\sigma_h)q = \lambda Bq, \tag{54}$$

converge to the eigenvalues of the differential equation as h decreases to zero. (Only the smallest eigenvalues are well approximated by the discretization; these are also the eigenvalues of physical interest.) The eigenvector q consists of the values of u_h and its derivative at the mesh points. There is a slight conflict of notation; σ_h refers both to a piecewise constant function and to the vector of variables that defines it. The boundary conditions of (53) are "clamped-clamped"; thus A and B are defined so that u_h and its derivative are zero at 0 and 1. Note that, as in (45), A depends on the unknown variables while B does not. The integral constraint in (52) becomes a linear constraint on σ_h. Regarding the bounds on σ: a solution of the mathematical problem is known to exist only for $\alpha > 0$, $\beta < \infty$ [6]; however, in practice, these requirements do not seem to be necessary and for most experiments we used $\alpha = 0$, $\beta = \infty$.

We then applied the SPLP algorithm of §4 to find that σ_h which maximizes the smallest eigenvalue of (54), or equivalently, negating the signs of the eigenvalues, minimizes the largest one, subject to the linear integral constraint. The order of the matrices A and B, n, is $2N - 4$, and the number of variables, m, is $N - 1$, where $N = h^{-1} + 1$. We used the inverse version of subspace iteration without shifts to compute the eigenvalues, which requires the factorization of a band matrix at each optimization step, as explained in §6. Since it is known that the extremal eigenvalue cannot have multiplicity greater than 2, we computed only $r = 2$ eigenvalues. Most of the papers in the literature do not take this direct optimization approach. Of the few that do, we do not know of any that compute the dual matrix approximation U, which is the key to verifying optimality. (When $p = 1$, $A(\sigma_h)$ is linear, so the minimum eigenvalue is concave; when $p > 1$, concavity is lost, and satisfaction of the necessary conditions does not guarantee optimality, but comparison of the results for varying p indicates that our computed solutions are most likely global maxima.)

The results show that when $p > 1$, the optimal σ_h is bounded away from zero as $h \to 0$, but for $p = 1$ apparently the optimal σ_h converges to zero at two points as $h \to 0$. Presumably, the optimal column has zero thickness at two points if $p = 1$, but not if $p > 1$. This has been a subject of great controversy in the literature, especially when $p = 2$; see [6] for details. Plots of the optimal cross-sectional area $\sigma_h(x)$ are shown in Fig. 1 for $N = 513$ with $p = 1$ and $p = 2$, respectively. The functions plotted are piecewise constant with 512 pieces, with no interpolation. The strongest column is about 33 percent stronger than the uniform column with the same volume in the case $p = 2$ and about 25 percent stronger in the case $p = 1$.

In all cases $1 \leq p \leq 3$, the first eigenvalue is double at optimality. It is this double eigenvalue that has caused most of the debate in the literature; indeed, some authors have expressed doubt about the multiplicity even when giving the correct result for the optimal σ. Even more interesting, the 2×2 dual matrix U that demonstrates optimality has minimum eigenvalue bounded away from zero as $h \to 0$ for all $p > 1$, but for $p = 1$ the dual matrix is apparently singular in the limit as $h \to 0$. We conclude that the double multiplicity of the eigenvalue of the differential equation is "strongly stable" for $p > 1$, but not for $p = 1$.

The performance of the SPLP algorithm was very good. The results shown in Fig. 1 were obtained using a convergence tolerance $\epsilon = 10^{-3}$, with the multiplicity tolerance and trust radius set initially to $\tau = 0.1$ and $\rho = 5$ and the variables initialized to 1, i.e., the uniform column. The number of calls to the subspace iteration routine, i.e., the number of times the computation of the eigenvalues was required, was 60 for $p = 2$ and 27 for $p = 1$, with a total computation time of less than 1.5 hours on a Sparc station in each case. The residual of equations (12)–(13) was reduced to about 10^{-3} in the case $p = 2$ and about 10^{-2} in the case $p = 1$. The accuracy of the optimal λ_1 was approximately four decimal figures, with the gap between the first and second eigenvalues reduced to about 10^{-6}. Such fast convergence indicates a well-conditioned optimization problem, since the method is only first-order. We also performed some experiments with α, the lower bound on σ_h, set to a positive number, e.g., 0.25. The active bound strategy used by the SPLP algorithm worked very effectively. Typically, most of the active bounds were identified in just a few steps, with fine tuning of the active set taking place subsequently.

10. Design of optimal preconditioners. Greenbaum and Rodrigue [21] have used our optimization programs to solve the following problem: given a positive definite symmetric matrix B, find the positive definite symmetric matrix M with prescribed sparsity pattern which minimizes the two-norm condition number of $M^{-1/2}BM^{-1/2}$. They show that M equivalently minimizes the maximum eigenvalue (in absolute value) of

$$I - M^{-1/2}BM^{-1/2}$$

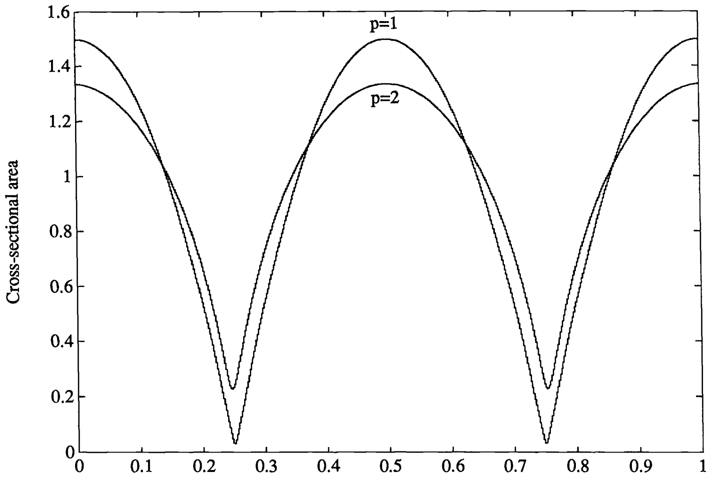


FIG. 1. *The shape of the strongest column.*

or

$$I - L^{-1}ML^{-T}, \tag{55}$$

where LL^T is a Cholesky factorization of B. The latter formulation is preferable, since the variables, the nonzero elements of the sparse matrix M, enter linearly. Since a factorization of B is used, finding the optimal preconditioner is clearly much more costly than solving a system $Bx = b$; the idea is that finding such optimal preconditioners gives insight that can then be widely applied.

The work reported in [21] was done before the SPLP version of the algorithm was available, so the SQP method described in §3 was used, the eigenvalues being computed by Eispack. The primary interest was in matrices B arising from elliptic partial differential equations, but only very coarse meshes could be handled. Nonetheless, it was found that the experiments gave a substantial amount of insight. For example, the optimal tridiagonal preconditioner M for B equal to the five-point finite difference approximation to the Laplacian on the square was computed. The results led to the conjecture that the optimal condition number is $O(h^{-2})$, where h is the mesh size in each direction, and that the optimal tridiagonal preconditioner is only slightly better than simply setting M to be the tridiagonal part of B. It was also found that the optimal solution yields (55) with a double eigenvalue at each end of its spectrum, these two double eigenvalues having the same magnitude. Further experiments involving domain decomposition were also done; this is a promising area for further investigation.

A better way to formulate the optimization problem is to minimize the maximum eigenvalue, in absolute value, of the generalized eigenvalue problem

$$(M - B)q = \lambda Bq.$$

Note that, as in §7, the variables, i.e., the elements of M, appear only on the left-hand side. Using this formulation, we have now performed further experiments with the SPLP version of our algorithm. Our first idea was to compute the extreme eigenvalues of the pencil $(M - B, B)$ by direct subspace iteration. This requires only one Cholesky factorization of B before the optimization iteration begins. However, convergence was

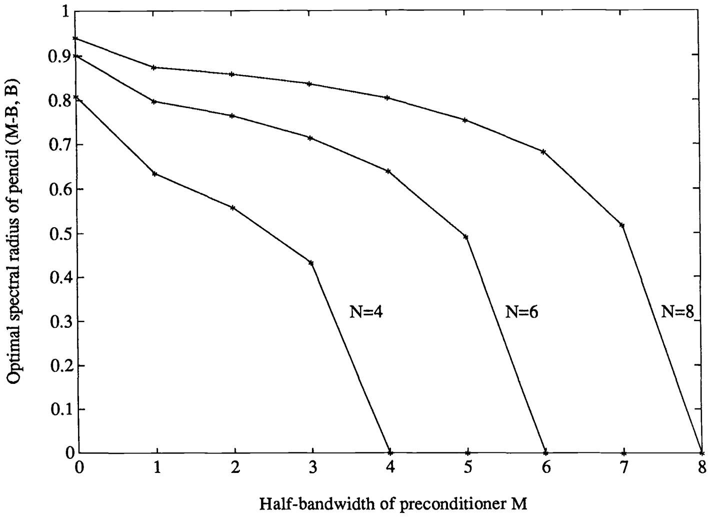


FIG. 2. *Optimal banded preconditioners for B =discrete Laplacian.*

much too slow for this approach to be practical. We therefore used shifted inverse iteration to independently compute the algebraically largest eigenvalues of the pencils

$$(A^{(1)},B) = (M-B,B) \quad \text{and} \quad (A^{(2)},B) = (B-M,B).$$

This required factorizations of $(s+1)B-M$ and $(s-1)B+M$ at each optimization step, for which we used the Linpack band matrix subroutines. At the optimal solution of all test problems, and indeed usually after a few optimization steps, the largest and smallest eigenvalues of $(M-B,B)$ were approximately equal in magnitude and opposite in sign. As explained in §8, two dual matrices $U^{(1)}$ and $U^{(2)}$ are generated by the SPLP algorithm, with dimensions t_1 and t_2, which are the computed multiplicities of each end of the spectrum of $(M-B,B)$. Note that instead of (13), we have the condition

$$\operatorname{tr} U^{(1)} + \operatorname{tr} U^{(2)} = 1.$$

We computed the optimal banded preconditioner M for B equal to the finite difference negative Laplacian on a unit square with mesh size h in each direction. We assumed Dirichlet boundary conditions, so that B and M are $n \times n$ matrices, where $n = N^2$, $N = h^{-1} - 1$. The matrix M is said to have half-bandwidth k if its total bandwidth is $2k+1$; thus, for $k=0$, M is restricted to be diagonal, while if $k = N$, the optimal solution is $M = B$. The dimension of the optimization problem, m, is approximately $(k+1)N^2$. The results support the following conjecture: the optimal preconditioner M with half-bandwidth k gives a pencil $(M-B,B)$ with eigenvalues of multiplicity $k+1$ at each end of its spectrum for all $k < N$. However, computing accurate optimal preconditioners for even moderate mesh sizes was very difficult for the simple reason that, like the discrete Laplacian itself, the eigenvalue optimization problem is increasingly ill conditioned as N increases. The negative end of the spectrum of $(M-B,B)$ has a cluster of eigenvalues which becomes more dense as N increases. For small mesh sizes ($N \leq 6$) there was not much difficulty identifying the apparently correct optimal multiplicity $k+1$, but this became more difficult for larger N, since the gap between the extremal eigenvalue and the interior eigenvalues becomes smaller as N increases. Furthermore, it is apparently the case that $\operatorname{tr} U^{(1)} \to 0$ and $\operatorname{tr} U^{(2)} \to 1$ as $N \to \infty$, showing that the positive end of the spectrum of $(M-B, B)$ becomes more and more irrelevant as the discrete Laplacian B becomes closer to being singular.

The situation is quite different from that reported for the column problem as we allow the mesh size to go to zero. The column problem is well posed in infinite dimensions and the finite-dimensional optimization problem is well conditioned as $N \to \infty$. By contrast, the optimal preconditioning problem for the Laplacian is not a well-posed problem in infinite dimensions. The reason for this is that the column problem is concerned only with one end of the spectrum of the differential operator, namely, the lowest eigenvalue that corresponds (in the case that it is simple) to a positive eigenfunction, while the optimal preconditioning problem is concerned with both ends of the spectrum, including eigenvalues corresponding to highly oscillatory eigenfunctions.

The computed optimal spectral radius of $(M-B, B)$ is plotted in Fig. 2 for various k and N. The trend is clear. The optimal tridiagonal preconditioner represents a significant improvement over the optimal diagonal preconditioner (which is a scalar multiple of the identity matrix). However, increasing k gives successively smaller improvements until k starts to approach N. This, of course, reflects the fact that the discrete Laplacian has only five nonzero diagonals, namely, the three main diagonals and the Nth sub- and super-diagonal.

11. A graph problem. The following problem was communicated to us by Schramm and Zowe; its origin may be found in [29] and [22]. Given an undirected graph G, with vertices $1, \cdots, n$, let M be an $n \times n$ symmetric matrix with the restriction that its diagonal elements are zero and its offdiagonal elements (i,j) are zero if i and j are not adjacent in the graph, and let x be the vector whose components are the nonrestricted lower triangular elements of M. The problem is to choose M, or equivalently x, to minimize the largest eigenvalue of

$$A(x) = M + ee^T, \tag{56}$$

where $e = [1, \cdots, 1]^T$. The minimum value for the max eigenvalue is known to give an upper bound for the Shannon capacity of the graph [29]. (The upper bound is sometimes called the Lovasz number of the graph.)

We applied our eigenvalue optimization algorithm to a test problem suggested by [46]. Given integers $\alpha \geq 1$ and $\omega \geq 3$, let $n = \alpha\omega + 1$ and define G to have the property that vertices i and j are adjacent if $j - i < \omega$ or $i + n - j < \omega$. The class of graphs with this property is denoted $C_n^{\omega-1}$. We tried solving the optimization problem for various values $\alpha \leq 10$ and $\omega \leq 6$. For these examples the order of the matrix n is moderate (≤ 61), but the number of variables m, which is the number of pairs of adjacent vertices in the graph, is large (≤ 305). Consequently, it is important to use the SPLP version of the optimization algorithm, but it is reasonable (though not very efficient) to compute the eigenvalues using Eispack. (Unshifted subspace iteration would not work since the smallest eigenvalue, which is of no interest, is negative and sometimes has a larger magnitude than the largest eigenvalue.)

The test problems are certainly very interesting. In all cases the algorithm *immediately* generated a point, say $\hat{x}$, where the max eigenvalue is multiple to machine precision, with the two optimality conditions (12)–(13) satisfied to machine precision.

TABLE 1
Summary of results for graph problem.

α	ω	m	λ_1	t	min e.v.(U)	# λ-evals.
3	4	39	3.106027	7	.0532	1
4	4	51	4.132934	7	.0545	1
5	4	67	5.151476	7	.0556	217
8	4	99	8.183308	7	.0575	130
10	4	123	10.195149	7	.0584	219
3	6	95	3.055559	11	.0195	235
4	6	125	4.073890	11	.0209	238
5	6	155	5.087257	11	.0219	187
6	6	185	6.097343	11	.0227	181
7	6	215	7.105194	11	.0233	227
8	6	245	8.111465	11	.0237	957*
9	6	275	9.116589	11	.0241	478
10	6	305	10.120845	11	.0244	608*

The multiplicity was seven in the cases where $\omega = 4$ and eleven in the cases where $\omega = 6$. (In some cases this required as many as four optimization steps, since successive doubling of the trust radius was needed to make a sufficiently large change in x.) In the case of the first two test problems, the dual matrix U was positive semidefinite and the algorithm terminated with the optimal solution $\hat{x}$. In all other cases, however, the dual matrix U was not positive semidefinite and so it was necessary for the algorithm to split the multiple eigenvalue to obtain a lower point, as described in Theorem 7. The algorithm then took many more steps to converge to the optimal solution x^*. In all these cases, the max eigenvalue had the same multiplicity at the final solution x^* as at the initially generated point $\hat{x}$. This unusual behavior of the algorithm indicates some underlying linear structure of the eigenvalues that is not generic and not well understood at the present.

In general, it seems that the optimal multiplicity is $2\omega - 1$. Another interesting observation is that the minimum eigenvalue of the optimal dual matrix has multiplicity two for all the problems we have run.

The results are summarized in Table 1. The first two columns specify the problem, and the third gives the number of variables. The next three columns give the computed optimal max eigenvalue, its multiplicity, and the smallest eigenvalue of the associated dual matrix U. The last value given is the number of times the eigenvalues of $A(x)$ were computed (using Eispack). The convergence tolerance was set to $\epsilon = 10^{-6}$. The multiplicity tolerance and trust region radius were initialized to $\tau = .01$ and $\rho = 10$, respectively. The variables were all initialized to -1. The norm of the residual of (12)–(13) was reduced in each case to about 10^{-6}, except in the first two cases, where it was reduced to machine precision (about 10^{-14}) in one step. In the two cases marked by an asterisk (*) it was necessary to restart the algorithm at one point (with the original values of τ and ρ) to obtain a satisfactory residual for (12)–(13). It is not clear why the case $\alpha = 8$, $\omega = 6$ was so much more difficult than the others, but in all cases an accurate solution was eventually found. (For the purposes of the graph application, the iteration could have been terminated much sooner, since the integer part of the solution is of primary interest, but we wanted to test the accuracy of the SPLP method.)

It is of some interest to compare our algorithm to that used by Schramm and Zowe, a "bundle trust region" method, which, as the name suggests, combines ideas of trust region methods with those of the early subgradient bundle methods of Lemarechal [27]. This algorithm is intended for general nonsmooth optimization problems, not necessarily involving eigenvalues. The bundle trust region method accumulates a set ("bundle") of subgradients during the course of the optimization. In the version described in [47] and [57], one subgradient is added to the bundle per iteration, namely,

$$[q^T A_1(x)q, \cdots, q^T A_m(x)q]^T, \tag{57}$$

where, as earlier,

$$A_k(x) = \frac{\partial A(x)}{\partial x_k}$$

(in this case a matrix with one nonzero element), and where q is a normalized eigenvector corresponding to $\lambda_1(x)$, arbitrarily chosen from the invariant subspace if the multiplicity of $\lambda_1(x)$ is greater than one. Theorem 2 (together with the chain rule) assures us that this vector is indeed a subgradient of $\lambda_1(x)$, that is, an element of the generalized gradient $\partial\lambda_1(x)$.

The initial comparison of our results with those of Schramm and Zowe showed that, while both algorithms obtained accurate solutions, our algorithm usually required fewer steps to achieve the same accuracy [46]. However, a revised version of Schramm and Zowe's algorithm has now been tested, where at each iteration, if $\lambda_1(x)$ has approximate multiplicity t, then t subgradients of the form (57) are added to the bundle of subgradients, for q equal to the t different columns of the matrix of eigenvectors $Q_1(x)$. This strategy substantially improved the algorithm, which now requires far fewer steps than ours for the same accuracy [46]. The reason for the dramatic improvement is not completely clear, but it may be related to the surprising initial behavior of our algorithm. Considering (6) in Theorem 2 again, we see that the first version of Schramm and Zowe's algorithm computes the subgradient defined by $U = e_1 e_1^T$, while the second version computes the t subgradients defined by $U = e_k e_k^T, k = 1, \cdots, t$ (here e_k is the kth column of the identity matrix). Clearly, then, one could add more subgradients to the bundle, using other permissible values for U; there is nothing special about the choice $U = e_k e_k^T$, since the basis Q_1 has been arbitrarily chosen by Eispack. The feature of our algorithm which we believe to be very attractive is that it efficiently computes $t(t+1)/2$ generically linearly independent subgradients at each iteration, namely, the gradients of the structure functionals (8), while the dual matrix estimate U defines the linear combination of these subgradients that satisfies the optimality condition (12) in the limit. This dual matrix is the key not only to the verification of optimality but also to any sensitivity analysis of the solution (see Theorem 7).

It would be premature to draw conclusions as to whether the bundle trust region algorithm or ours is more efficient, for several reasons: the former requires an estimate of the optimal solution value, which ours does not; the former solves a QP (with dimension equal to the number of subgradients in the bundle), which ours does not; comparisons have been made only on the graph problems just described, which apparently have a rather special structure that is not completely understood. We expect that it should be possible to improve the rate of convergence of our algorithm by approximating second-order information (see §5). We also wonder if the bundle trust region algorithm would have difficulties when the eigenvalues are computed by a shifted iterative method, since the basis Q_1 would tend to be little changed at each iteration. By contrast, when Eispack is used, the basis $Q_1(x)$ does not generally converge as $x \to x^*$ (see the examples in [14]), perhaps giving a bundle that is more "rich" in the various possible values for the subgradients.

Finally, we note that the dual matrix itself appears in the references [29] and [22]. Indeed, the property stated as Theorem 4 in [29] and the third equality in Theorem 9.3.12 of [22] is a special case of Theorem 6 given above, specifically giving the dual formulation (16). It seems likely that the multiplicity of the minimum eigenvalue of the optimal dual matrix U (found to be two in our experiments), as well as the multiplicity of the optimal maximum eigenvalue of $A(x)$ (conjectured to be 2ω), should be significant for the understanding of the original graph capacity problem.

12. Concluding remarks. We have derived optimality conditions for an important eigenvalue optimization model problem, emphasizing the representation of the generalized gradient in terms of a dual matrix U. We have given a practical algorithm for solving large-scale problems of this type, based on successive partial linear programming, which has been applied very successfully in diverse application areas. The behavior of the algorithm was quite different for the three applications described in detail. The column problem described in §9 is a well-posed infinite-dimensional optimization problem; discretized versions were solved very efficiently by the algorithm. The preconditioning problem described in §10 gave rise to very ill conditioned problems, which were nonetheless solved by the algorithm to reasonable accuracy. The algorithm also gave very accurate answers to the graph problems described in §11, which have a rather special structure that is not completely understood.

The SQP algorithm of [36], on which the new algorithm is based, has also been applied to some other applications not discussed in this paper, including the quadratic assignment problem [44], the stability of Runge–Kutta methods for ordinary differential equations [30], and optimal diagonal scaling of nonsymmetric matrices [55]. Another application to which we hope to apply our large-scale algorithm is the computation of structured singular values in control [9], [11], [54].

Perhaps the most important feature of our algorithms is that they compute the optimal dual matrix U, which is the key to the verification of optimality and to sensitivity analysis of the solution. Given the optimal dual eigenspace basis Q_1^*, the dual matrix U is unique if the active linear constraints of the limiting LP or QP are independent (see Theorem 9). If the linear independence assumption fails to hold, the problem is said to be degenerate, since U is then not uniquely defined and verification of optimality is much more difficult; this happens, for example, in the Runge–Kutta problems of [30]. Because the basis Q_1^* may be replaced by any other orthonormal basis spanning the same eigenspace, it is the eigenvalues of U that are of significance. Nonnegativity of the eigenvalues of U is a necessary condition for optimality and, together with the other conditions of Theorem 5, a sufficient condition if $A(x)$ is affine. The eigenvalues of U play essentially the same role in sensitivity analysis of optimal solutions as that well known for dual variables (Lagrange multipliers) in the context of nonlinear programming; see Theorem 7. In particular, if the smallest eigenvalue of U is zero, it may be concluded that the optimal multiplicity of the minimization objective $\lambda_1(x)$ is not strongly stable.

Acknowledgments. The work on the column problem is a joint effort with Steve Cox; much of the development of the large-scale versions of the algorithm described in §§5 and 6 was also joint work with Cox in the course of obtaining solutions to the column problem. The work on optimal preconditioners is joint with Anne Greenbaum. The presentation of the optimality conditions in Theorems 1–3 is related to [38], which is joint work with Rob Womersley. I would like to thank Helga Schramm and Jochem Zowe for providing me with details of the graph problems and the related performance of their bundle trust region method. I have also received much helpful input from many other people, too numerous to list here, which I nonetheless gratefully acknowledge.

REFERENCES

[1] J. C. Allwright, *On maximizing the minimum eigenvalue of a linear combination of symmetric matrices*, SIAM J. Matrix Anal. Appl., 10 (1989), pp. 347–382.

[2] K. Bathe and E. Wilson, *Numerical methods in finite element analysis*, Prentice–Hall, Englewood Cliffs, NJ, 1976.

[3] S. P. Boyd, Private communication, 1987.

[4] A. Bratus, *Multiple eigenvalues in problems of optimizing the spectral properties of systems with a finite number of degrees of freedom*, USSR J. Comput. Math. and Math. Phys., 26 (1986), pp. 1–7.

[5] F. H. Clarke, *Optimization and Nonsmooth Analysis*, John Wiley, New York, 1983.

[6] S. J. Cox and M. L. Overton, *On the optimal design of columns against buckling*, SIAM J. Math. Anal., 23 (1992), to appear.

[7] J. Cullum, W. E. Donath, and P. Wolfe, *The minimization of certain nondifferentiable sums of eigenvalues of symmetric matrices*, Math. Programming Stud., 3 (1975), pp. 35–55.

[8] J. J. Dongarra, C. B. Moler, J. R. Bunch, and G. W. Stewart, LINPACK *Users' Guide*, Society for Industrial and Applied Mathematics, Philadelphia, PA, 1978.

[9] J. Doyle, *Analysis of feedback systems with structured uncertainties*, IEE Proc., Part D, 129 (1982), pp. 242–250.

[10] R. M. Ehrdahl, *Two algorithms for the lower bound method of reduced density matrix theory*, Rep. Math. Phys., 15 (1979), pp. 147–162.

[11] M. K. H. Fan and A. L. Tits, *Characterization and efficient computation of the structured singular value*, IEEE Trans. Automat. Control, 31 (1986), pp. 734–743.

[12] R. Fletcher, *Semi-definite constraints in optimization*, SIAM J. Control Optim., 23 (1985), pp. 493–513.

[13] ———, *Practical Methods of Optimization*, Second Edition, John Wiley, Chichester, New York, 1987.

[14] S. Friedland, J. Nocedal, and M. L. Overton, *Four quadratically convergent methods for solving inverse eigenvalue problems*, in Numerical Analysis, D. Griffiths, ed., New York, 1986, John Wiley, pp. 47–65. Pitman Research Note in Mathematics 140, John Wiley, New York, 1986, pp. 47–65.

[15] ———, *The formulation and analysis of numerical methods for inverse eigenvalue problems*, SIAM J. Numer. Anal., 24 (1987), pp. 634–667.

[16] P. E. Gill, W. Murray, M. A. Saunders, and M. H. Wright, *User's guide for* LSSOL: *A Fortran package for constrained linear least-squares and convex quadratic programming*, Systems Optimization Laboratory Report 86-1, Stanford University, Stanford, CA, 1986.

[17] P. E. Gill, W. Murray, and M. H. Wright, *Practical Optimization*, Academic Press, New York, London, 1981.

[18] C. Goh and K. Teo, *On minimax eigenvalue problems via constrained optimization*, J. Optim. Theory Appl., 57 (1988), pp. 59–68.

[19] B. Gollan, *Eigenvalue perturbations and nonlinear parametric optimization*, Math. Programming Stud., 30 (1987), pp. 67–81.

[20] G. H. Golub and C. Van Loan, *Matrix Computations*, The Johns Hopkins University Press, Baltimore, MD, 1983.

[21] A. Greenbaum and G. H. Rodrigue, *Optimal preconditioners of a given sparsity pattern*, BIT, 29 (1990), pp. 610–634.

[22] M. Grötschel, L. Lovász, and A. Schriver, *Geometric Algorithms and Combinatorial Optimization*, Springer-Verlag, New York, 1988.

[23] C. B. Gurwitz and M. L. Overton, *Sequential quadratic programming methods based on approximating a projected Hessian matrix*, SIAM J. Sci. Statist. Comput., 10 (1989), pp. 631–653.

[24] A. D. Ioffe and V. M. Tihomirov, *Theory of Extremal Problems*, North-Holland, Amsterdam, 1979.

[25] T. KATO, *A Short Introduction to Perturbation Theory for Linear Operators*, Springer-Verlag, New York, 1982.

[26] P. LANCASTER, *On eigenvalues of matrices dependent on a parameter*, Numer. Math., 6 (1964), pp. 377–387.

[27] C. LEMARECHAL AND R. MIFFLIN, EDS., *Nonsmooth Optimization*, Pergamon Press, Oxford, U.K., 1978.

[28] D. C. LIU AND J. NOCEDAL, *On the limited memory* BFGS *method for large scale optimization*, Math. Programming, 45 (1989), pp. 503–528.

[29] L. LOVÁSZ, *On the Shannon capacity of a graph*, IEEE Trans. Inform. Theory, 25 (1979), pp. 1–7.

[30] M. MÜLLER, *Algebraische Stabilitätsbedingungen für Runge–Kutta-Verfahren*, Ph.D. thesis, Universität Karlsruhe, Karlsruhe, FRG, 1990.

[31] W. MURRAY AND M. L. OVERTON, *A projected Lagrangian algorithm for nonlinear minimax optimization*, SIAM J. Sci. Statist. Comput., 1 (1980), pp. 345–370.

[32] NAG *library manual*, Numerical Algorithms Group, Oxford, U.K.

[33] N. OLHOFF AND S. RASMUSSEN, *On single and bimodal optimum buckling loads of clamped columns*, Internat. J. Solids and Structures, 9 (1977), pp. 605–614.

[34] N. OLHOFF AND J. E. TAYLOR, *On structural optimization*, J. Appl. Mech., 50 (1983), pp. 1138–1151.

[35] M. R. OSBORNE, *Finite Algorithms in Optimization and Data Analysis*, John Wiley, Chichester, New York, 1985.

[36] M. L. OVERTON, *On minimizing the maximum eigenvalue of a symmetric matrix*, SIAM J. Matrix Anal. Appl., 9 (1988), pp. 256–268.

[37] M. L. OVERTON AND R. S. WOMERSLEY, *On minimizing the spectral radius of a nonsymmetric matrix function: optimality conditions and duality theory*, SIAM J. Matrix Anal. Appl., 9 (1988), pp. 473–498.

[38] ———, *Optimality conditions and duality theory for minimizing sums of the largest eigenvalues of symmetric matrices*, Computer Science Department Report 566, New York University, New York, 1991. Submitted to Math. Programming.

[39] ———, *Second derivatives for optimizing eigenvalues of symmetric matrices*, manuscript in preparation.

[40] C. C. PAIGE AND M. A. SAUNDERS, *Solution of sparse indefinite systems of linear equations*, SIAM J. Numer. Anal., 12 (1978), pp. 617–629.

[41] E. R. PANIER, *On the need for special purpose algorithms for minimax eigenvalue problems*, Tech. Report, Department of Electrical Engineering, University of Maryland, College Park, MD, 1989.

[42] B. N. PARLETT, *The Symmetric Eigenvalue Problem*, Prentice–Hall, Englewood Cliffs, NJ, 1980.

[43] E. POLAK AND Y. WARDI, *A nondifferentiable optimization algorithm for structural problems with eigenvalue inequality constraints*, J. Structural Mech., 11 (1983), pp. 561–577.

[44] F. RENDL AND H. WOLKOWICZ, *Applications of parametric programming and eigenvalue maximization to the quadratic assignment problem*, Math. Programming, (1992), to appear.

[45] R. T. ROCKAFELLAR, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.

[46] H. SCHRAMM, Private communication, 1989.

[47] ———, *Eine Kombination von Bundle- und Trust-Region-Verfahren zur Lösung nichtdifferenzierbarer Optimierungsprobleme*, Ph.D. thesis, Universität Bayreuth, Bayreuth, FRG, 1989.

[48] A. SHAPIRO, *Extremal problems on the set of nonnegative definite matrices*, Linear Algebra Appl., 67 (1985), pp. 7–18.

[49] ———, *Optimal block diagonal l_2-scaling of matrices*, SIAM J. Numer. Anal., 22 (1985), pp. 81–94.

[50] B. T. SMITH, J. M. BOYLE, J. DONGARRA, B. GARBOW, Y. IKEBE, V. C. KLEMA, AND C. B. MOLER, *Matrix Eigensystem Routines—EISPACK Guide*, Lecture Notes in Computer Science 6, Springer-Verlag, New York, 1967.

[51] G. STRANG AND G. J. FIX, *An Analysis of the Finite Element Method*, Prentice–Hall, Englewood Cliffs, NJ, 1973.

[52] D. B. SZYLD, *A two-level iterative method for large sparse generalized eigenvalue calculations*, Ph.D. thesis, Department of Mathematics, New York University, 1983.

[53] I. TADJBAKHSH AND J. B. KELLER, *Strongest columns and isoperimetric inequalities for eigenvalues*, J. Appl. Mech., 29 (1962), pp. 159–164.

[54] G. A. WATSON, *Computing the structured singular value, and related problems*, in Numerical Analysis 1989, D. Griffiths, ed., Pitman Research Notes in Mathematics 228, John Wiley,

New York, 1990, pp. 258–275.

[55] ———, *An algorithm for optimal l_2 scaling of matrices*, IMA J. Numer. Anal., 11 (1991), pp. 481–492.

[56] J. Zhou, Private communication, 1988.

[57] J. Zowe, *The BT-algorithm for minimizing a nonsmooth functional subject to linear constraints*, in Nonsmooth Optimization and Related Topics, F. H. Clarke, V. F. Demyanov, and F. Gianessi, eds., Plenum Press, New York, 1989, pp. 459–480.

SIAM J. OPTIMIZATION
Vol. 2, No. 1, pp. 121–152, February 1992

008

A VERSION OF THE BUNDLE IDEA FOR MINIMIZING A NONSMOOTH FUNCTION: CONCEPTUAL IDEA, CONVERGENCE ANALYSIS, NUMERICAL RESULTS*

HELGA SCHRAMM† AND JOCHEM ZOWE†

Abstract. During recent years various proposals for the minimization of a nonsmooth functional have been made. Amongst these, the bundle concept turned out to be an especially fruitful idea. Based on this concept, a number of authors have developed codes that can successfully deal with nonsmooth problems. The aim of the paper is to show that, by adding some features of the trust region philosophy to the bundle concept, the end result is a distinguished member of the bundle family with a more stable behaviour than some other bundle versions. The reliability and efficiency of this code is demonstrated on the standard academic test examples and on some real-life problems.

Key words. nondifferentiable optimization, bundle methods

AMS(MOS) subject classifications. 90C30, 65K05

1. Introduction and exposition of conceptual ideas.

This paper deals with the minimization of a *nonsmooth* functional (i.e., $f \notin C^1$)

$$\text{minimize } f(x) \text{ where } f : \mathbb{R}^n \to \mathbb{R}. \tag{1.1}$$

Additional constraints in (1.1) do not cause difficulties—at least in theory; they can be added to f as (nonsmooth) exact penalty terms. As usual in *nonsmooth optimization* (NSO), we require throughout that

$$f \text{ is locally Lipschitzian.} \tag{1.2}$$

For such f the *subdifferential* of f at x,

$$\partial f(x) := \operatorname{conv}\{g \in \mathbb{R}^n \mid g = \lim_{i\to\infty} \nabla f(x_i),\, x_i \to x, \nabla f(x_i) \text{ exists, } \nabla f(x_i) \text{ converges}\}, \tag{1.3}$$

is a well-defined, nonempty, convex, and compact subset of $\mathbb{R}^n$; this and other standard facts from convex analysis and NSO can be found, e.g., in the textbooks by Clarke [4] and Rockafellar [35]. The elements of $\partial f(x)$ are called *subgradients* of f at x. Quite naturally, these subgradients serve in NSO as substitute for the gradients. Hence, parallel to what is standard in *smooth optimization*, we require in the following that we dispose of a subroutine that

$$\text{computes } f(x) \text{ and one (arbitrary) } g \in \partial f(x) \text{ for given } x. \tag{1.4}$$

This seems to be a modest (and minimal) requirement. Sections 4.2–4.4, however, will show that in many real-life situations, the computation of only one $g \in \partial f(x)$ is all but easy and is the time-consuming job per iteration.

* Received by the editors March 26, 1990; accepted for publication (in revised form) December 18, 1990. This research has been supported by Deutsche Forschungsgemeinschaft, Forschungsschwerpunkt "Anwendungsbezogene Optimierung und Steuerung."

† Mathematisches Institut, Universität Bayreuth, Postfach 101251, 8580 Bayreuth, Germany.

1.1. Subgradient methods. Apart from the cutting plane method (see below), the first methods which could deal with (1.1) under assumption (1.4) and for convex f were the *Russian subgradient methods* (also called *Kiev methods*); see, e.g., Ermoliev [9], Poljak [34], and Shor [39]. At iterate x_k one makes a step along a negative subgradient with some off-line chosen steplength

$$x_{k+1} := x_k + \lambda_k d_k \quad \text{where } d_k := -g_k/\|g_k\| \quad \text{with } g_k \in \partial f(x_k)\,. \tag{1.5}$$

It can be shown that, under rather suggestive assumptions and for $\lambda_k \downarrow 0$ and $\sum_{k=1}^{\infty} \lambda_k = \infty$, the x_k from (1.5) converge to an optimal point. The simple structure of these subgradient methods still makes them widely used, although they suffer from some serious drawbacks: The methods do not guarantee a descent at each step, they lack an implementable stopping criterion, and the convergence speed is extremely poor (less than linear). The last disadvantage can be partly overcome by premultiplying d_k in (1.5) with some variable metric matrix H_k, which is updated in a simple way at each iteration. Linear convergence in the function values can be established for a member of this class (see, e.g., Shor [39]); the additional H_k, however, makes the method very cumbersome for large n.

1.2. Bundle concept. Lemaréchal [23] and Wolfe [41] initiated a giant stride forward in NSO by the *bundle concept*, which can handle convex and nonconvex f. Since the motivating ideas come from the convex situation, we assume a convex f throughout this motivating section.

All bundle methods carry two distinctive features (for some other views in this section, see Lemaréchal [24] and Zowe [43]):

(i) They make use at the iterate x_k of the *bundle* of information $(f(x_k), g_k)$, $(f(x_{k-1}), g_{k-1}), \cdots$ collected so far to build up a model of f;

(ii) If, due to the kinky structure of f, this model is not yet an adequate one, then they mobilize even more subgradient information close to x_k.

Recipe (i) leads in a natural way to the *cutting plane* (CP) *approximation* of f at x_k:

$$\max_{1\le i\le k} \{g_i^T(x - x_i) + f(x_i)\}\,. \tag{1.6}$$

Equation (1.6) is a piecewise linear approximation of the convex f from below, which coincides with f at all x_i. For short, we put $d := x - x_k$ in (1.6) and use the notation

$$f_{\mathrm{CP}}(x_k; d) := \max_{1\le i\le k} \{g_i^T d + g_i^T(x_k - x_i) + f(x_i)\} \quad \text{for } d \in \mathbb{R}^n\,. \tag{1.7}$$

Obviously there is no reason to trust this substitute for f far away from x_k. Therefore a stabilizing term $(1/2t_k)d^T d$ with positive t_k is added in (1.7), when minimizing this CP-model of f. If f_{CP} models f well enough close to x_k, then the minimizer d_k of

$$f_{\mathrm{CP}}(x_k; d) + \frac{1}{2t_k} d^T d$$

is a descent direction for f and a linesearch along $x_k + \lambda d_k$ for $\lambda \ge 0$ provides some x_{k+1} with $f(x_{k+1}) < f(x_k)$. For a nonsmooth f it may happen, however, that f_{CP} is such a poor approximation of f that d_k is not a descent direction for f (or that the linesearch only leads to a marginal decrease in f); think, e.g., of $f(x) = |x|$, $x_i < 0$ for $i = 1, \cdots, k$ and x_k close to the kink 0. Here strategy (ii) comes up: Obviously f_{CP} does not copy f on the halfline $x_k + \lambda d_k$, $\lambda \ge 0$; to master this lack of information one stays at x_k and enriches the model by including one more subgradient from $\partial f(x_k + \lambda d_k)$ for small $\lambda > 0$. Omitting all details, we obtain the following.

(1.8) **Iteration** $x_k \to x_{k+1}$**:**
- (1) Compute $d_k := d(t_k) := \arg\min\{f_{\mathrm{CP}}(x_k; d) + (\frac{1}{2t_k})d^T d \mid d \in \mathbb{R}^n\}$.
- (2) Perform a linesearch for f along $x_k + \lambda d_k$, $\lambda \geq 0$.
  - (a) If the linesearch leads to a "sufficient decrease" in f, then make a **Serious Step**: Put $x_{k+1} := x_k + \lambda_k d_k$ with $\lambda_k \in \arg\min_{\lambda>0} f(x_k + \lambda d_k)$ and compute $g_{k+1} \in \partial f(x_{k+1})$.
  - (b) If the linesearch yields only an "insufficient decrease," then make a **Null Step**: Put $x_{k+1} := x_k$ and compute $g_{k+1} \in \partial f(x_k + \lambda d_k)$ for suitable small $\lambda > 0$.

Unlike the subgradient approach, the above iteration guarantees a decrease for each (Serious) Step. Further, one disposes of an implementable stopping criterion: x_k is "optimal" as soon as d_k in (1) is "close" to 0. And, since the linesearch adjusts the steplength λ_k to the chosen d_k, one has a considerably faster convergence speed. All this can be made precise and a detailed convergence analysis exists for convex and nonconvex f; see Lemaréchal, Strodiot, and Bihain [26]; Mifflin [31]; or the monograph by Kiwiel [18].

The above concept has been implemented by a number of authors. We mention in particular the advanced and sophisticated Fortran code M1FC1 by Lemaréchal [27], which is widely used in NSO. Numerous test runs proved the efficiency of this code; see also §4 below. Needless to say, M1FC1 is only "work of man!" The reader who is familiar with M1FC1 in applications will agree that the code suffers from two weak points: First, the success of M1FC1 depends in a delicate way on the parameter t_k in step (1) of (1.8) (actually some "dual" parameter ε_k is used in M1FC1); a bad guess for t_k (respectively, ε_k) leads to a "bad" search direction d_k and M1FC1 breaks down with linesearch difficulties. Second, for $f \in C^1$ and $t_k \to 0$, (1.8) reduces to the *steepest descent method*, which is only linearly convergent. Numerical experiments confirm this first-order behaviour of M1FC1. We will discuss how one can bypass the first shortcoming in practice; further, it will become obvious how to deal, in principle, with the second problem and how to reach faster convergence.

1.3. Bundle trust region concept. We start with a simple observation: With d_k from step (1) of iteration (1.8) and $\rho_k := \frac{1}{2}d_k^T d_k$, the minimization in (1.8)(1) becomes "equivalent" to

$$\text{(1.9)} \qquad \text{compute } d(\rho_k) := \arg\min\{f_{\mathrm{CP}}(x_k; d) \mid \tfrac{1}{2}d^T d \leq \rho_k\}.$$

This follows by a comparison of the Kuhn–Tucker conditions for the two problems. A closer inspection shows that there is even a monotone correspondence between t_k and ρ_k. Now we replace (1.8)(1) by (1.9). It then becomes obvious how to bypass the first difficulty discussed above. Instead of working with some a priori and more or less randomly chosen ρ_k (respectively, t_k), we follow the *trust region* philosophy: We decrease and/or increase ρ_k in a systematic way (*trust region part*) and improve f_{CP} by Null Steps (*bundle part*), until we reach some f_{CP} together with a ρ_k-ball, on which we can *trust* this model, i.e., the d_k from (1.9) leads to a substantial decrease in f. The advantage of this procedure is twofold: It suggests a way to choose ρ_k, and it releases us at the same time from the need for a linesearch. Obviously we can apply just the same strategy in (1.8) and tune the t_k. The reason for working with (1.8) is purely numerical in nature. We will see that the minimization of (1.8) leads to a quadratic programming problem with a lot of reliable software (e.g., [19]). This is not true for (1.9) because of the quadratic constraint. In schematic terms, we obtain

(1.10) **Iteration** $x_k \to x_{k+1}$:

```
    (1) Compute $d_k := d(t_k) := \arg\min\{f_{CP}(x_k; d) + \frac{1}{2t_k} d^T d \mid d \in \mathbb{R}^n\}$.
    (2) If $f(x_k + d_k)$ is "sufficiently smaller" than $f(x_k)$, then either
        (a) enlarge $t_k$ and go back to (1), or
        (b) make a Serious Step: Put $x_{k+1} := x_k + d_k$, compute $g_{k+1} \in \partial f(x_{k+1})$.
        If $f(x_k + d_k)$ is "not sufficiently smaller" than $f(x_k)$, then either
        (c) reduce $t_k$ and go back to (1), or
        (d) make a Null Step: Put $x_{k+1} := x_k$, compute $g_{k+1} \in \partial f(x_k + d_k)$.
```

How to solve the alternatives (a)–(b) and (c)–(d) will be seen in the precise statement of the algorithm.

Preliminary versions of the above variant of the bundle family have been presented in [38] and [44]. In these versions the t_k were reduced in (2)(c) only as long as they stayed above some fixed positive lower bound $\underline{t}$ (otherwise one had to make a Null Step). This restrictive assumption for the inner iteration (2)(c) can be skipped now by introducing a modified Null Step criterion (see also [21]). The above variant has been implemented by us under the name BT (= "implicit" **b**undle **t**rust region) algorithm (see the remark before Theorem 2.3). Extensive testing (in particular on some real-life problems, which are known as "tough nuts") proved the code to be efficient and reliable so far. We want to convince the reader of this claim and encourage him to work with our code and other bundle implementations.

Let us briefly return to the second drawback of the existing bundle implementations, namely the linear (hence slow) convergence. Obviously, the trust region approach could also help with this difficulty by tuning the bilinear form $d^T d$ in step (1) of (1.10) to account for the compiled knowledge about the level sets of f. There is a whole series of recent papers that address this challenging item and try to gain control of such curvature (hence second-order) information by using ideas from the ellipsoid method; see, e.g., Goffin [11]; Goffin, Haurie, and Vial [13]; Sonnevend and Stoer [40]; Kiwiel [20]. Only some first attempts for implementing these concepts have been made. Some more abstract approaches to second-order ideas in NSO are reviewed in [24].

We mention that our work has benefitted greatly from cooperation with Lemaréchal and from the work of Kiwiel (in particular, [18]). In a recent paper [21], Kiwiel proposed a bundle variant, which is close to our BT-iteration. The difference is that Kiwiel does not adapt the t in some inner iteration as we do in steps (2)(a) and (2)(c), because he does not work with the trust region philosophy. His t is updated *after* having made a Serious Step or a Null Step.

The paper is organized as follows. Since the motivation and the key arguments are based on convexity, we treat the convex case in detail in §2, i.e., we will specify iteration (1.10) together with the overall algorithm and present the convergence analysis for convex f. Section 3 discusses the necessary modifications for nonconvex f and states the convergence results without proofs; the detailed proofs can be found in Schramm [37]. Section 4 will verify our claim that BT behaves well in practice.

Some remarks on the notation: $\|\cdot\|$ denotes the Euclidean norm. The subscript k always refers to the sequence of iterates $x_1, x_2, \cdots$, whereas the superscript j will be used in the inner iteration, which leads from x_k to x_{k+1}. If J is a set of indices, then $|J|$ denotes its cardinality. Further, we put

$$\Lambda(n) := \left\{ \lambda \in \mathbb{R}^n \mid \lambda_i \geq 0,\ 1 \leq i \leq n, \text{ and } \sum_{i=1}^{n} \lambda_i = 1 \right\}.$$

2. BT-algorithm: The convex case. We assume throughout this section that f is convex. Then the elements of the subdifferential can be characterized by an inequality:

$$g \in \partial f(x) \iff g^T(y-x) \leq f(y) - f(x) \quad \text{for all } y \in \mathbb{R}^n. \tag{2.1}$$

This *subgradient inequality* plays a crucial role for the conceptual ideas and in the convergence analysis. For later use we add a continuity result for the set-valued map $x \to \partial f(x)$:

$$\text{the map } x \to \partial f(x) \text{ is locally bounded and upper semicontinuous}. \tag{2.2}$$

Further, let us mention that a convex $f : \mathbb{R}^n \to \mathbb{R}$ is locally Lipschitzian, i.e., our general continuity assumption (1.1) holds.

2.1. The cutting plane model. At the iterate x_k we have at our disposal the sequence $x_1, x_2, \cdots, x_k$ and a collection of auxiliary points y_i together with subgradients $g_i \in \partial f(y_i)$ for $i \in J_k$; here J_k is some nonempty set of indices. On first reading, the reader may think of J_k as a subset of $\{1, \cdots, k\}$ and assume $y_i = x_i$. This bundle of information leads to the *cutting plane model* $\max_{i \in J_k}\{g_i^T(x - y_i) + f(y_i)\}$ of f. With the *linearization errors*

$$\alpha_{k,i} := \alpha(x_k, y_i) := f(x_k) - (f(y_i) + g_i^T(x_k - y_i)) \tag{2.3}$$

and the new variable $d := x - x_k$, we can write this in a condensed form

$$\max_{i \in J_k}\{g_i^T d - \alpha_{k,i}\} + f(x_k) \quad \text{for } d \in \mathbb{R}^n.$$

For convenience, let us skip the constant $f(x_k)$ and put

$$f_{\text{CP}}(x_k; d) := \max_{i \in J_k}\{g_i^T d - \alpha_{k,i}\} \quad \text{for } d \in \mathbb{R}^n. \tag{2.4}$$

Step (1) from iteration (1.10) becomes, for *suitable* t (which still has to be chosen appropriately!):

$$\text{compute } d := d(t) = \arg\min \left\{ f_{\text{CP}}(x_k; d) + \frac{1}{2t}\|d\|^2 \mid d \in \mathbb{R}^n \right\}. \tag{2.5}$$

This can equivalently be written as a quadratic programming problem in $\mathbb{R}^1 \times \mathbb{R}^n$:

$$\begin{aligned} &\text{compute } (v,d) := (v(t), d(t)) \\ &\qquad = \arg\min \left\{ v + \tfrac{1}{2t}\|d\|^2 \mid v \geq g_i^T d - \alpha_{k,i} \text{ for } i \in J_k \right\}. \end{aligned} \tag{2.6}$$

Problem (2.5) is a strictly convex problem with a unique minimizer $d(t)$; the same holds for (2.6), of course. From the Kuhn–Tucker conditions for (2.6), one easily obtains a representation for $d(t)$ and $v(t)$.

LEMMA 2.1. *For the solution* $(v(t), d(t))$ *of* (2.6) *there exists* $\lambda(t) \in \Lambda(|J_k|)$ *such that*

$$(2.7)\quad \lambda_i(t)(-v(t) + g_i^T d(t) - \alpha_{k,i}) = 0 \quad \textit{for } i \in J_k\,,$$

$$(2.8)\quad d(t) = -t \sum_{i \in J_k} \lambda_i(t) g_i\,,$$

$$(2.9)\quad v(t) = -t \left\| \sum_{i \in J_k} \lambda_i(t) g_i \right\|^2 - \sum_{i \in J_k} \lambda_i(t) \alpha_{k,i} = -\frac{1}{t} \|d(t)\|^2 - \sum_{i \in J_k} \lambda_i(t) \alpha_{k,i}\,.$$

Since (2.6) is a convex problem with linear constraints, the Kuhn–Tucker conditions (i.e., (2.7)–(2.9)) are also sufficient for optimality of a feasible x.

Thanks to convexity, all $\alpha_{k,i}$ are nonnegative (a consequence of (2.1)),

$$(2.10)\qquad \alpha_{k,i} \geq 0 \quad \text{for } i \in J_k.$$

Now add $\alpha_{k,i} - [f(x_k) - f(y_i) - g_i^T(x_k - y_i)] = 0$ to the subgradient inequality

$$g_i^T(x - y_i) \leq f(x) - f(y_i);$$

one obtains, after simple reordering,

$$(2.11)\qquad g_i^T(x - x_k) \leq f(x) - f(x_k) + \alpha_{k,i} \quad \text{for all } x \in \mathbb{R}^n \quad \text{and} \quad i \in J_k\,,$$

i.e., $\alpha_{k,i}$ "measures" how good $g_i \in \partial f(y_i)$ satisfies the subgradient inequality at the point x_k. The $\alpha_{k,i}$ take care that the influence of g_i in (2.6) and (2.16) below will be greater the smaller the weight $\alpha_{k,i}$ is.

Now fix some $\lambda \in \Lambda(|J_k|)$, multiply (2.11) by λ_i, and sum up over i. We obtain the useful formula, which holds with arbitrary $\lambda \in \Lambda(|J_k|)$:

$$(2.12)\qquad \left(\sum_{i \in J_k} \lambda_i g_i \right)^T (x - x_k) \leq f(x) - f(x_k) + \sum_{i \in J_k} \lambda_i \alpha_{k,i} \quad \text{for all } x \in \mathbb{R}^n\,.$$

Inequality (2.12) can be interpreted similarly to (2.11) above.

As a direct conclusion from (2.9) and (2.10) we note:

$$(2.13)\qquad v(t) \leq 0 \quad \text{for the optimal } v(t) \text{ from (2.6).}$$

As expected, $v(t) = 0$ characterizes optimality of x_k. This follows immediately from our next result, if we put there $\varepsilon = 0$ and use (2.9). The lemma itself is an immediate consequence of inequality (2.12).

LEMMA 2.2. *Suppose there exists* $\lambda \in \Lambda(|J_k|)$ *with*

$$(2.14)\qquad \left\| \sum_{i \in J_k} \lambda_i g_i \right\| \leq \varepsilon \quad \textit{and} \quad \sum_{i \in J_k} \lambda_i \alpha_{k,i} \leq \varepsilon\,.$$

Then x_k *is* ε*-optimal, i.e.,*

$$f(x_k) \leq f(x) + \varepsilon \|x - x_k\| + \varepsilon \quad \textit{for all } x \in \mathbb{R}^n\,.$$

For later use we add a continuity result on $(v(t), d(t))$, which follows easily from the strict convexity of the objective function in (2.5):

(2.15) The solution $(v(t), d(t))$ of (2.6) depends continuously on $t \in (0, \infty)$.

Due to the simple structure of (2.6), the last statement can be strengthened substantially. We add without proof (a detailed treatment is given in [37]):

- There exists a finite sequence $0 = t^0 < t^1 < \cdots < t^m = \infty$ and $a^i, b^i \in \mathbb{R}^n$, such that $d(t) = a^i + tb^i$ for $t \in (t^{i-1}, t^i]$ and $i = 1, 2, \cdots, m$;
- $a^1 = 0$ and $b^1 =$ projection of the origin onto $\operatorname{conv}\{g_i \mid i \in J_k \text{ and } \alpha_{k,i} = 0\}$;
- There exists a CP-solution d_{CP} (i.e., d_{CP} minimizes $f_{\mathrm{CP}}(x_k; \cdot)$) if and only if $a^m = d_{CP}$ and $b^m = 0$.

Remark. For an efficient implementation of (2.6), two devices become important.

(a) The index set J_k (i.e., the number of subgradients carried along) should be kept at reasonable size as $k \to \infty$. Hence from time to time we clean up the bundle. The convergence analysis requires $|J_k| \geq 3$ together with a certain *reset strategy*.

(b) Problem (2.6) is a quadratic programming problem in $1+n$ variables and $|J_k|$ linear constraints. Since, typically, $|J_k|$ will be much smaller than the dimension n, we replace (2.6) by its *dual* in $|J_k|$ variables and $|J_k| + 1$ constraints:

$$\min \left\{ \frac{1}{2} \left\| \sum_{i \in J_k} \lambda_i g_i \right\|^2 + \frac{1}{t} \sum_{i \in J_k} \lambda_i \alpha_{k,i} \mid \lambda \in \Lambda(|J_k|) \right\}. \tag{2.16}$$

Some standard duality arguments show that the solutions λ of (2.16) and the $\lambda(t)$ from Lemma 2.1 correspond to each other.

In the next section we make clear how to find an appropriate t for (2.6). Section 2.3 summarizes the overall algorithm and §2.4 presents the convergence analysis.

2.2. Inner iteration $x_k \to x_{k+1}$. We fix an upper bound T for t, parameters $0 < m_1 < m_2 < 1$, $0 < m_3 < 1$, some small $\nu > 0$, and a stopping parameter $\varepsilon \geq 0$. Suppose we are at the iterate x_k and let J_k, y_i, $g_i \in \partial f(y_i)$ and $\alpha_{k,i}$ be as discussed above. Then we specialize (1.10) as follows. Here the superscript j is the running index; the subscript k is kept fixed. The stopping rule in step (1) is based on (2.8), (2.9), and Lemma 2.2. Finally, the decisive criteria **SS** and **NS** will be specified below.

(2.17) **Inner iteration $x_k \to x_{k+1}$:**

(0) Choose $t^1 := t_{k-1}$. Set $l^1 := 0$, $u^1 := T$, and $j := 1$.

(1) Compute the solution $(v^j, d^j) = (v(t^j), d(t^j))$ of (2.6). If $(1/t^j)\|d^j\| \leq \varepsilon$ and $-(1/t^j)\|d^j\|^2 - v^j \leq \varepsilon$, then stop: x_k is ε-optimal. Otherwise put $y^j := x_k + d^j$ and compute $g^j \in \partial f(y^j)$.

(2) (a) If **SS**(i) and **SS**(ii) hold, then make a **Serious Step**: Put $x_{k+1} := y_{k+1} := y^j$, $g_{k+1} := g^j$ and stop.

(b) If **SS**(i) holds but not **SS**(ii), then put $l^{j+1} := t^j$, $u^{j+1} := u^j$, $t^{j+1} := \frac{1}{2}(u^{j+1} + l^{j+1})$, $j := j + 1$ and go back to (1).

(c) If **NS**(i) and **NS**(ii) hold, then make a **Null Step**: Put $x_{k+1} := x_k$, $y_{k+1} := y^j$, $g_{k+1} := g^j$ and stop.

(d) If **NS**(i) holds but not **NS**(ii), then put $u^{j+1} := t^j$, $l^{j+1} := l^j$, $t^{j+1} := \frac{1}{2}(u^{j+1} + l^{j+1})$, $j := j + 1$ and go back to (1).

Let v_k, d_k, and t_k be the values, with which we leave (2.17) in case of a Serious Step or a Null Step. Then $d_k = -t_k \sum_{i \in J_k} \lambda_{k,i} g_i$ for suitable $\lambda_k = (\lambda_{k,i}) \in \Lambda(|J_k|)$ (see (2.8)). With this λ_k, we define for later use

$$z_k := \sum_{i \in J_k} \lambda_{k,i} g_i \quad \text{and} \quad \sigma_k := \sum_{i \in J_k} \lambda_{k,i} \alpha_{k,i} \, . \tag{2.18}$$

We now present the criteria that determine whether a Serious Step or a Null Step is taken (for $k = 1$ put in **NS**(ii) $z_0 := g_1$, $\sigma_0 := 0$):

SS: (i) $f(y^j) - f(x_k) < m_1 v^j$,
SS: (ii) $(g^j)^T d^j \geq m_2 v^j$ or $t^j \geq T - \nu$,
NS: (i) $f(y^j) - f(x_k) \geq m_1 v^j$,
NS: (ii) $\alpha(x_k, y^j) \leq m_3 \sigma_{k-1}$ or $|f(x_k) - f(y^j)| \leq \|z_{k-1}\| + \sigma_{k-1}$.

Discussion of SS and NS. *Ad* (2)(a) *and* (b): Condition **SS**(i) ensures, for a Serious Step, a decrease of at least m_1 times $v_k [= f_{\mathrm{CP}}(x_k; d_k) =$ decrease in the CP-model]. The first part of **SS**(ii) takes care of a substantial change in the CP-model; this follows from (we use $x_{k+1} = y_{k+1}$, $v_k < 0$ and $m_2 < 1$)

$$g_{k+1}^T d_k - \alpha_{k+1,k+1} = g_{k+1}^T d_k \geq m_2 v_k > v_k \geq g_i^T d_k - \alpha_{k,i} \quad \text{for } i \in J_k \, , \tag{2.19}$$

which implies that, after a Serious Step, the updated model (2.6) will provide some (v, d) in step $k + 1 \to k + 2$, which differs from the present (v_k, d_k). If the first part of **SS**(ii) does not hold (and such a change in the model cannot be guaranteed) and if t is still smaller than some upper bound T (this is taken care of by the second condition under **SS**(ii)), then we prefer to try some larger t, even if **SS**(i) holds. This motivates steps (2)(a) and (2)(b).

Ad (2)(c) *and* (d): Now suppose **NS**(i) holds. Then either f_{CP} is not yet an adequate model and/or we were too optimistic with respect to t. The obvious way out: Try some smaller t in (1); this is step (2)(d). If, however, the first condition under **NS**(ii) also holds, then a Null Step makes sense as well and we prefer this option. The reason: after such a Null Step, we get from (2.11) for $k + 1$ and $i = k + 1$ (use $x_{k+1} = x_k$)

$$g_{k+1}^T (x - x_k) \leq f(x) - f(x_k) + \alpha_{k+1,k+1} \, ,$$

where $\alpha_{k+1,k+1} = \alpha(x_k, y_{k+1}) \leq m_3 \sigma_{k-1}$ and $m_3 < 1$. We conclude that g_{k+1} is "close" to $\partial f(x_k)$ and thus it makes sense to add g_{k+1} to the bundle at x_k. Condition **NS**(i) guarantees that this g_{k+1} contributes nonredundant information. This follows from the next inequality, which serves the same purpose as (2.19) in case of a Serious Step:

$$\begin{aligned} g_{k+1}^T d_k - \alpha_{k+1,k+1} &= f(y_{k+1}) - f(x_k) \geq m_1 v_k \\ &> v_k \geq g_i^T d_k - \alpha_{k,i} \quad \text{for } i \in J_k; \end{aligned} \tag{2.20}$$

consequently, in iteration $k+1 \to k+2$ the enriched model f_{CP} will yield some direction d which differs from the unsuccessful present d_k. This, taken together, explains one-half of (2)(c); for technical reasons (which will become clear in Proposition 2.7 below) we also make a Null Step, if **NS**(i) holds together with the second condition under **NS**(ii).

We summarize (2.17) in a flow chart (see Fig. 2.1). We state a by-product of the

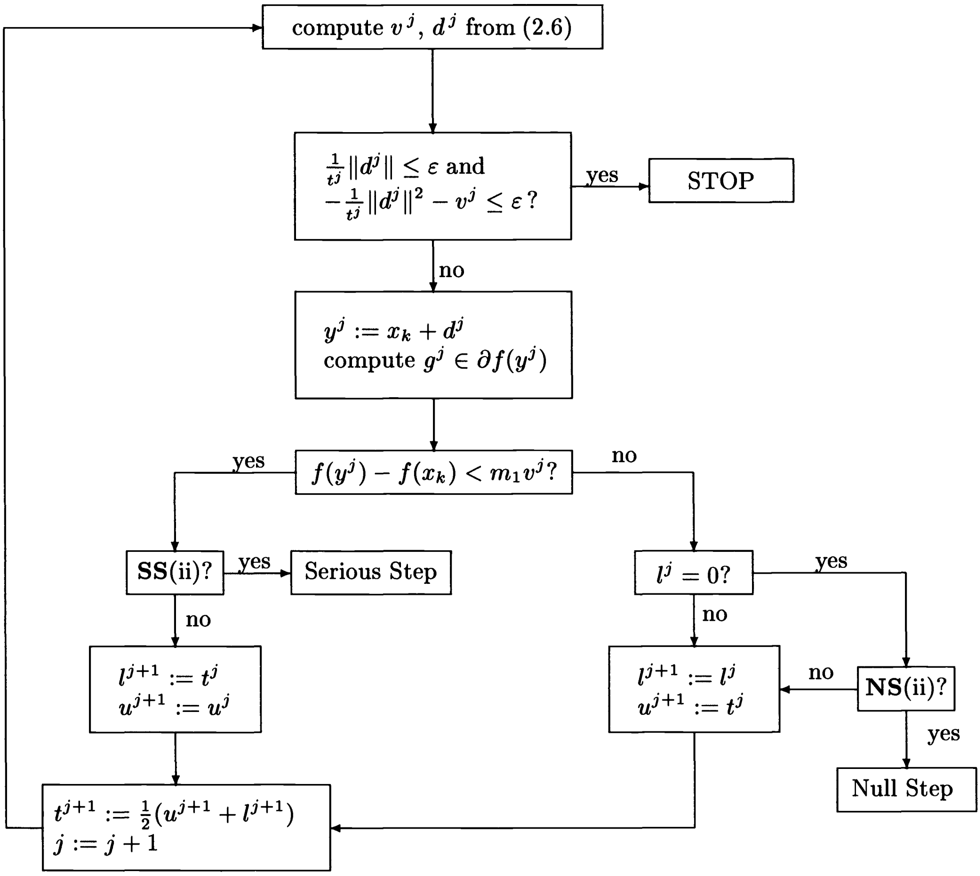

compute v^j, d^j from (2.6)
$\frac{1}{t^j}\|d^j\| \le \varepsilon$ and
$-\frac{1}{t^j}\|d^j\|^2 - v^j \le \varepsilon$?
yes
STOP
no
$y^j := x_k + d^j$
compute $g^j \in \partial f(y^j)$
yes
$f(y^j) - f(x_k) < m_1 v^j$?
no
SS(ii)?
yes
Serious Step
no
$l^{j+1} := t^j$
$u^{j+1} := u^j$
$l^j = 0$?
yes
no
$l^{j+1} := l^j$
$u^{j+1} := t^j$
no
NS(ii)?
yes
Null Step
$t^{j+1} := \frac{1}{2}(u^{j+1} + l^{j+1})$
$j := j + 1$

FIG. 2.1. *Flow chart for inner iteration.*

proof of Theorem 2.3 below:

(2.21) If $f(y^j) - f(x_k) < m_1 v^j$ for some j, then one leaves (2.17) with a Serious Step.

Hence it suffices to check **NS**(ii) in Fig. 2.1 only as long as $l^j = 0$.

In our implementation of (2.17) we replace the simple bisection rule for t by a more sophisticated heuristic strategy. We choose a safeguarded variation of t, which corresponds to the change of the function value. In step (0) we choose the initial $t^1 = t_{k-1}$ only in case of a Null Step; in case of a Serious Step we choose $t^1 \ge t_{k-1}$, as in [21].

The t-variation in (2.17) corresponds to the linesearch in M1FC1. The crucial difference is: In M1FC1 one makes an a priori decision on t_k (respectively, on some dual quantity ε_k). This results in a *fixed direction* d_k and, in the line search, one "minimizes" $f(x_k + \cdot\, d_k)$. In (2.17) t is variable and we thus try *different directions* $d(t)$ when "minimizing" $f(x_k + d(\cdot))$. The examples from §4 show that this can be a decisive advantage.

Remark. Let us mention that the t-adjustment in (2.6) is actually of an *implicit* nature and one should better talk of an *implicit trust region approach* in our context. A similar implicit trust region idea was considered in a recent paper by Bell [2].

The next result supplies the actual justification for what we are doing.

THEOREM 2.3. *Iteration* (2.17) *ends after finitely many cycles, either with a Serious Step or a Null Step or the information that* x_k *is* ε*-optimal.*

Proof (by contradiction). Suppose the algorithm is an endless cycle. Then three cases can occur: (i) $l^j = 0$ for all j; (ii) $u^j = T$ for all j; (iii) neither (i) nor (ii) holds.

Ad (i): We are always on the right branch in Fig. 2.1 and thus $t^{j+1} = \frac{1}{2}(0+t^j) \downarrow 0$ and $y^j \to x_k$ as $j \to \infty$. Hence **NS**(ii) will hold for large enough j; since **NS**(i) is satisfied by construction on the right branch, (2.17) stops with a Null Step in contradiction to our assumption.

Ad (ii): Now we are always on the left branch and thus $t^{j+1} = \frac{1}{2}(t^j + T) \uparrow T$ for $j \to \infty$, i.e., **SS**(ii) holds for large enough j. Since **SS**(i) is automatically satisfied on the left branch, we will stop with a Serious Step in contradiction to our assumption.

Ad (iii): In this case $0 < l^j < u^j < T$ for all sufficiently large j and a monotonicity argument implies $l^j \uparrow t^*$ and $u^j \downarrow t^*$ for some $t^* \in (0, T)$. A continuity argument (recall (2.15)) together with **SS**(i) and **NS**(i) yields for $d^* := d(t^*)$ and $v^* := v(t^*)$

$$f(x_k + d^*) - f(x_k) = m_1 v^* \,. \tag{2.22}$$

Let $j(1), j(2), \cdots$ be the subsequence of indices, for which **SS**(i) holds; this is an infinite sequence since otherwise $l^{j(m)} = t^*$ for some m and then (2.22) would contradict **SS**(i). Since $l^j \uparrow t^*$, the $g^{j(i)}$ have a cluster point g^* that belongs to $\partial f(x_k + d^*)$ (we use (2.2)). Hence

$$(g^*)^T(x_k - (x_k + d^*)) \leq f(x_k) - f(x_k + d^*),$$

and, because of (2.22),

$$(g^*)^T d^* \geq m_1 v^*.$$

Now $v^* < 0$ (otherwise (2.17) would have stopped because of ε-optimality) together with $0 < m_1 < m_2$ shows $g^* d^* > m_2 v^*$. A continuity argument implies, for sufficiently large i,

$$(g^{j(i)})^T d^{j(i)} \geq m_2 v^{j(i)},$$

hence we will stop with a Serious Step in contradiction to our assumption. □

2.3. The overall algorithm. We briefly summarize the overall algorithm with *reset strategy.*

(2.23) **BT-algorithm:** Choose a starting point $x_1 \in \mathbb{R}^n$ and parameters $T > 0$, $0 < m_1 < m_2 < 1$, $0 < m_3 < 1$, $\nu > 0$, $\varepsilon \geq 0$ and an upper bound $J_{\max} \geq 3$ for $|J_k|$.

(0) Compute $f(x_1)$, $g_1 \in \partial f(x_1)$ and put $y_1 := x_1$, $J_1 := \{1\}$ and $k := 1$.

(1) INNER ITERATION: Compute x_{k+1} and g_{k+1} as in (2.17) or realize that x_k is ε-optimal (in which case we stop).

(2) If $|J_k| = J_{\max}$, then go to (3); otherwise put $J := J_k$ and go to (4).

(3) RESET: Choose $J \subset J_k$ with $|J| \leq J_{\max} - 2$ and $\max\{i \mid i \in J_k,\ \alpha_{k,i} = 0\} \in J$. Introduce some additional index $\tilde{k}$ and define with z_k, σ_k from (2.18)

$$g_{\tilde{k}} := z_k, \qquad \alpha_{k,\tilde{k}} := \sigma_k, \qquad J := J \cup \{\tilde{k}\}.$$

(4) UPDATE: If the outcome of (2.17) was a Serious Step, then put

$$\alpha_{k+1,i} := \alpha_{k,i} + f(x_{k+1}) - f(x_k) - g_i^T d_k \quad \text{for } i \in J,\ \alpha_{k+1,k+1} := 0\,.$$

If the outcome of (2.17) was a Null Step, then put

$$\alpha_{k+1,i} := \alpha_{k,i} \quad \text{for } i \in J,\ \alpha_{k+1,k+1} := \alpha(x_k, y_{k+1})\,.$$

Put $J_{k+1} := J \cup \{k+1\}$ and go to (1).

Remark. We add a comment on the index $\tilde{k}$ in step (3) and the update formula in step (4).

(a) The $g_{\tilde{k}}$ defined in the reset step corresponds to the *aggregate subgradient* introduced in [18]. Usually $g_{\tilde{k}}$ will not be a subgradient at some point $y_{\tilde{k}}$ and thus $\alpha_{k,\tilde{k}}$ does not fit into the concept (2.4). It follows, however, from (2.12) that the synthetic $\alpha_{k,\tilde{k}}$ again satisfies

$$g_{\tilde{k}}^T(x - x_k) \leq f(x) - f(x_k) + \alpha_{k,\tilde{k}} \quad \text{for all } x\,,$$

which is actually what is needed from subgradients.

(b) One easily checks that for the indices i which correspond to points y_i, the update formula in (4) is in accordance with (2.4). The update strategy dispenses the need to carry along the x_i's and y_i's.

2.4. Convergence analysis. The proof technique below is largely based on ideas that go back to Kiwiel [18]. Throughout, we work with the stopping parameter $\varepsilon = 0$. Let x_k, $k = 1, 2, \cdots$, be the iterates generated by (2.23) and recall the abbreviations introduced in (2.18):

$$z_k = \sum_{i \in J_k} \lambda_{k,i} g_i \quad \text{and} \quad \sigma_k = \sum_{i \in J_k} \lambda_{k,i} \alpha_{k,i}\,.$$

In terms of z_k and σ_k, the crucial relations (2.8) and (2.9) become

$$d_k = -t_k z_k \quad \text{and} \quad v_k = -t_k \|z_k\|^2 - \sigma_k\,. \tag{2.24}$$

Further, let us denote the minimal value in (2.16) by w_k, i.e.,

$$w_k = \frac{1}{2}\|z_k\|^2 + \frac{1}{t_k}\sigma_k\,, \tag{2.25}$$

and put

$$X^* := \{x^* \in \mathbb{R}^n \mid f(x^*) \leq f(x) \text{ for all } x \in \mathbb{R}^n\}\,.$$

Finally we mention a technical assumption that we will need for our auxiliary results:

$$\text{There exists } \bar{x} \text{ such that } f(\bar{x}) \leq f(x_k) \text{ for all } k. \tag{2.26}$$

As a foretaste of what we will prove, we summarize the result:

$$\begin{aligned}&f(x_k) \text{ converges to } \inf_x f(x) \ (\geq -\infty) \text{ and,}\\ &\text{if } X^* \neq \emptyset, \text{ then } x_k \text{ converges to some } x^* \in X^*\,.\end{aligned}$$

We start with the following observation.

LEMMA 2.4. *If* (2.26) *holds, then for each* $\delta > 0$ *there exists* $n_0(\delta) \in \mathbb{N}$ *such that*

$$\|\bar{x} - x_{k+1}\|^2 \leq \|\bar{x} - x_m\|^2 + \delta \quad \textit{for} \quad k \geq m \geq n_0(\delta)\,. \tag{2.27}$$

Proof. Equation (2.12) becomes, in terms of z_k and σ_k,

$$z_k^T(\bar{x} - x_k) \leq f(\bar{x}) - f(x_k) + \sigma_k$$

and, since $f(\bar{x}) \leq f(x_k)$,

$$z_k^T(\bar{x} - x_k) \leq \sigma_k\,.$$

If we put

$$\delta_k := \begin{cases} 1\,, & \text{if } k \to k+1 \text{ is a Serious Step}\,,\\ 0\,, & \text{if } k \to k+1 \text{ is a Null Step}\,,\end{cases}$$

then $x_{k+1} - x_k = \delta_k d_k = -\delta_k t_k z_k$ for all k and thus

$$-(\bar{x} - x_k)^T(x_{k+1} - x_k) = \delta_k t_k (\bar{x} - x_k)^T z_k \leq \delta_k t_k \sigma_k\,.$$

It follows that

$$\begin{aligned}\|\bar{x} - x_{k+1}\|^2 &= \|\bar{x} - x_k\|^2 + \|x_k - x_{k+1}\|^2 - 2(\bar{x} - x_k)^T(x_{k+1} - x_k)\\ &\leq \|\bar{x} - x_k\|^2 + \|x_k - x_{k+1}\|^2 + 2\delta_k t_k \sigma_k\,.\end{aligned}$$

Hence for all $m \in \mathbb{N}$ and $k \geq m$

$$\|\bar{x} - x_{k+1}\|^2 \leq \|\bar{x} - x_m\|^2 + \sum_{i=m}^{k}(\|x_i - x_{i+1}\|^2 + 2\delta_i t_i \sigma_i)\,. \tag{2.28}$$

Now consider the sum in (2.28). From $f(x_{i+1}) - f(x_i) \leq \delta_i m_1 v_i$ we obtain, for arbitrary $l > 1$,

$$f(x_1) - f(x_l) = f(x_1) - f(x_2) + f(x_2) - \cdots + f(x_{l-1}) - f(x_l) \geq -m_1 \sum_{i=1}^{l} \delta_i v_i\,,$$

and thus for $l \to \infty$ (we use (2.9), (2.8), and (2.24)),

$$\infty > f(x_1) - f(\bar{x}) \geq -m_1 \sum_{i=1}^{\infty} \delta_i v_i = m_1 \sum_{i=1}^{\infty} \delta_i (t_i \|z_i\|^2 + \sigma_i) .$$

Since $\delta_i t_i^2 \|z_i\|^2 = \|x_{i+1} - x_i\|^2$, we can continue

$$\infty > m_1 \sum_{i=1}^{\infty} \left(\frac{1}{t_i} \|x_{i+1} - x_i\|^2 + \delta_i \sigma_i \right) ,$$

and thus (we use that, by construction, $t_i \leq T$)

$$\infty > m_1 \sum_{i=1}^{\infty} (\|x_{i+1} - x_i\|^2 + \delta_i t_i \sigma_i) .$$

Consequently we can make the sum in (2.28) as small as we like by letting $m \to \infty$. This proves the assertion. □

The next lemma is an almost immediate consequence of (2.27).

LEMMA 2.5. *If* (2.26) *holds, then the* x_k *converge to some* $\tilde{x}$, *for which*

$$f(\tilde{x}) \leq f(x_k) \quad \textit{for all } k .$$

Proof. By (2.27) the x_k-sequence is bounded and has a cluster point, say $\tilde{x}$. Since, by construction, $f(x_k)$ is monotonically decreasing, we see

$$f(\tilde{x}) \leq f(x_k) \quad \text{for all } k .$$

Hence Lemma 2.4 applies once more (now with $\bar{x}$ replaced by $\tilde{x}$) and for given $\varepsilon > 0$ we can choose $n_0(\varepsilon/2)$ such that

$$\|\tilde{x} - x_{k+1}\|^2 \leq \|\tilde{x} - x_m\|^2 + \frac{\varepsilon}{2} \quad \text{for } k \geq m \geq n_0\left(\frac{\varepsilon}{2}\right) .$$

Since $\tilde{x}$ is a cluster point of the x_k-sequence, there exists $\tilde{m} \geq n_0(\varepsilon/2)$ with $\|\tilde{x} - x_{\tilde{m}}\|^2 \leq \varepsilon/2$ and we end up with

$$\|\tilde{x} - x_{k+1}\|^2 \leq \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon \quad \text{for all } k \geq \tilde{m} . \qquad \square$$

In the following we will show that $\tilde{x}$ from Lemma 2.5 is indeed optimal. For this aim, we prove for suitable subsequences

$$z_{k(i)} \to 0 \quad \text{and} \quad \sigma_{k(i)} \to 0 \quad \text{for } i \to \infty . \tag{2.29}$$

The optimality of $\tilde{x}$ follows from (2.12) (cf. Theorem 2.10).

We start with the crucial observation that besides x_k, the auxiliary sequences w_k, etc., are also bounded in situation (2.26).

LEMMA 2.6. *If* (2.26) *holds, then the sequences of* w_k, z_k, σ_k, d_k, y_k, g_k, *and* $\alpha_{k,k}$ $(k = 1, 2, \cdots)$ *are bounded.*

Proof. We combine (2.24) and (2.25) to see

$$0 \leq w_k = -\frac{1}{t_k} \left(v_k + \frac{1}{2t_k} \|d_k\|^2 \right) ,$$

i.e.,

$$0 \geq -w_k = \frac{1}{t_k}\left(\max_{i \in J_k}\{g_i^T d_k - \alpha_{k,i}\} + \frac{1}{2t_k}\|d_k\|^2\right).$$

Now choose $i(k) \in J_k$ such that $\alpha_{k,i(k)} = 0$ (such $i(k)$ exists because of our reset strategy); then we can continue the last inequality

$$0 \geq -w_k \geq \frac{1}{t_k} \min_{d \in \mathbb{R}^n}\left\{g_{i(k)}^T d + \frac{1}{2t_k}\|d\|^2\right\}.$$

With the minimizer $d := -t_k g_{i(k)}$ we obtain

$$0 \geq -w_k \geq -\frac{1}{2}\|g_{i(k)}\|^2. \tag{2.30}$$

The choice $\alpha_{k,i(k)} = 0$ guarantees $g_{i(k)} \in \partial f(x_k)$ (a consequence of (2.11) and (2.1)). This, together with the convergence of the x_k (Lemma 2.5) and the boundedness of the map $x \to \partial f(x)$ (see (2.2)) yields the boundedness of $\{g_{i(k)}\}_{k\in\mathbb{N}}$ and because of (2.30), the boundedness of $\{w_k\}_{k\in\mathbb{N}}$. A look at (2.25) and (2.24) convinces the reader that the sequences of z_k, σ_k,d_k, and $y_k = x_k + d_k$ are also bounded (we use $t_k \leq T$). Consequently the $g_k \in \partial f(y_k)$ are also bounded, since $\partial f(\cdot)$ is a bounded map. This, together with the convergence of the x_k and the continuity of f, finally proves the boundedness of the $\alpha_{k,k}$. □

In our two main auxiliary Propositions 2.8, 2.9 below, we will prove $v_k(i) \to 0$ (respectively, $w_k(i) \to 0$) for a suitable subsequence. A glance at (2.24) and (2.25) shows that this implies the crucial relation (2.29), provided $t_k \geq \underline{t} > 0$. The situation $t_k \to 0$ has to be treated as a special case in the next theorem. Here the role of the second condition in **NS**(ii) becomes clear: It is needed to ensure (2.29) even if $t_k \to 0$.

PROPOSITION 2.7. *Suppose* (2.26) *holds and* 0 *is a cluster point of* $\{t_k\}_{k\in\mathbb{N}}$. *Then, for suitable subsequences,*

$$\lim_{i\to\infty} z_{k(i)} = 0, \qquad \lim_{i\to\infty} \sigma_{k(i)} = 0.$$

Proof (by contradiction). Suppose there is $\delta > 0$ with

$$\|z_k\| + \sigma_k \geq \delta \quad \text{for all } k.$$

Now denote by $d_k(t)$ the solution in step (1) of (2.17) for variable t and variable k. The Lipschitz continuity of the convex f, the convergence of the x_k (Lemma 2.5) and (2.8), together with the boundedness of the g_i (Lemma 2.6) imply the existence of $L > 0$, $C > 0$, and $0 < \tilde{T} \leq T$ such that

$$|f(x_k + d_k(t)) - f(x_k)| \leq L\|d_k(t)\| \leq tLC \quad \text{for } k \quad \text{and} \quad t \leq \tilde{T}.$$

By making $\tilde{T}$ smaller, if necessary, we can guarantee that

$$|f(x_k + d_k(t)) - f(x_k)| \leq \delta \leq \|z_{k-1}\| + \|\sigma_{k-1}\| \quad \text{for } k \geq 2 \quad \text{and} \quad t \leq \tilde{T}.$$

Hence, whenever we are on the right branch in Fig. 2.1, then we will leave (2.17) with a Null Step as soon as, for the first time, $t^j \leq \tilde{T}$. Since t is increased on the left branch, we conclude from the bisection update rule for t that $t_k \geq \frac{1}{2}\tilde{T}$ for all k. This contradicts the assumption. □

It is convenient to discuss separately the case of finitely many Serious Steps and of infinitely many Serious Steps.

PROPOSITION 2.8. *Let* (2.26) *hold and suppose that one makes infinitely many Serious Steps in* (2.23). *Then for suitable subsequences,*

$$\lim_{i\to\infty} z_{k(i)} = 0, \qquad \lim_{i\to\infty} \sigma_{k(i)} = 0\,.$$

Proof. We may assume $t_k \geq \underline{t} > 0$ for all k, since otherwise the assertion follows from Proposition 2.7. Now let $\{x_{k(i)}\}_{i\in\mathbb{N}}$ be a subsequence resulting in Serious Steps, i.e.,

$$f(x_{k(i)+1}) - f(x_{k(i)}) < m_1 v_{k(i)}\,;$$

hence for $l \geq 1$ (note, $x_{k+1} = x_k$ for Null Steps)

$$f(x_{k(l)+1}) - f(x_{k(1)}) < m_1 \sum_{i=1}^{l} v_{k(i)}\,.$$

We conclude

$$f(\bar{x}) - f(x_{k(1)}) < m_1 \sum_{i=1}^{\infty} v_{k(i)}$$

and thus $0 \geq \sum_{i=1}^{\infty} v_{k(i)} > -\infty$. The assertion follows from (2.24) since $t_k \geq \underline{t} > 0$ for all k. □

PROPOSITION 2.9. *Suppose* (2.26) *holds and one makes only finitely many Serious Steps. Then for suitable subsequences*

$$\lim_{i\to\infty} z_{k(i)} = 0, \qquad \lim_{i\to\infty} \sigma_{k(i)} = 0\,.$$

Proof. Because of Proposition 2.7, we can again assume that

$$t_k \geq \underline{t} > 0 \quad \text{for all } k\,. \tag{2.31}$$

Further, there exists by assumption some $\bar{k}$ with

$$x_k = x_{\bar{k}} \quad \text{for } k \geq \bar{k}\,.$$

In step (i) we will discuss the change in the minimal value w_k of (2.16) from $w_k \to w_{k+1}$ for $k \geq \bar{k}$; this is used in (ii) to show $w_k \to 0$, which proves the assertion (see (2.25)).

Ad (i): We fix some $k \geq \bar{k}$ and consider the function for $\nu \in [0,1]$

$$Q(\nu) := \frac{1}{2}\|(1-\nu)z_k + \nu g_{k+1}\|^2 + (1-\nu)\frac{1}{t_{k+1}}\sigma_k + \nu\frac{1}{t_{k+1}}\alpha_{k+1,k+1}\,.$$

A glance at (2.16) tells us that

$$w_{k+1} \leq \min\{Q(\nu) \mid 0 \leq \nu \leq 1\} =: \tilde{w}\,. \tag{2.32}$$

To unburden the notation we put

$$\Delta_k := \frac{1}{t_{k+1}} - \frac{1}{t_k}$$

and skip the subscript k in the rest of part (i) and write $+$ for $k+1$. Simple arithmetic shows

$$\begin{aligned} Q(\nu) &= \frac{1}{2}\nu^2\|z-g_+\|^2 + \nu(z^Tg_+ - \|z\|^2) + \frac{1}{2}\|z\|^2 + \frac{1}{t_+}\sigma + \frac{1}{t_+}\nu(\alpha_{+,+}-\sigma) \\ &= \frac{1}{2}\nu^2\|z-g_+\|^2 + \nu(z^Tg_+ - \|z\|^2) + w + \Delta\sigma + \frac{1}{t_+}\nu(\alpha_{+,+}-\sigma)\,. \end{aligned} \tag{2.33}$$

Since we only make Null Steps for $k \geq \bar{k}$, one has, as a consequence of **NS**(i),

$$g_+^T(-tz) - \alpha_{+,+} \geq m_1 v > m_2 v = m_2(-t\|z\|^2 - \sigma_k)$$

and thus

$$g_+^T z \leq -\frac{1}{t}\alpha_{+,+} + m_2\left(\|z\|^2 + \frac{1}{t}\sigma\right)\,.$$

This inequality allows us to continue (2.33) for $\nu \in [0,1]$

$$\begin{aligned} Q(\nu) &\leq \frac{1}{2}\nu^2\|z-g_+\| + \nu\left(-\frac{1}{t}\alpha_{+,+} + m_2\|z\|^2 + m_2\frac{1}{t}\sigma - \|z\|^2\right) + w \\ &\quad +\Delta\sigma + \frac{1}{t_+}\nu(\alpha_{+,+}-\sigma) \\ &= \frac{1}{2}\nu^2\|z-g_+\|^2 - \nu(1-m_2)\left(\frac{1}{t}\sigma + \|z\|^2\right) + w \\ &\quad -\frac{1}{t}\nu(\alpha_{+,+}-\sigma) + \Delta\sigma + \frac{1}{t_+}\nu(\alpha_{+,+}-\sigma) \\ &\leq \frac{1}{2}\nu^2\|z-g_+\|^2 - \nu(1-m_2)w + w + \nu\Delta(\alpha_{+,+}-\sigma) + \Delta\sigma \\ &=: q(\nu)\,. \end{aligned} \tag{2.34}$$

With

$$(C_k =)\, C := \max\left\{\|z\|, \|g_+\|, \frac{1}{t}\sigma, 1\right\} \tag{2.35}$$

we can go on:

$$\begin{aligned} Q(\nu) &\leq q(\nu) \\ &\leq 2\nu^2C^2 - \nu(1-m_2)w + w + \nu\Delta(\alpha_{+,+}-\sigma) + \Delta\sigma \\ &=: \bar{q}(\nu)\,. \end{aligned}$$

For the special $\bar{\nu} := (1-m_2)w/4C^2$, we obtain from (2.33) and the last inequality (note that $\bar{\nu} \in [0,1]$ since $\bar{\nu} \leq (1-m_2)(\frac{1}{2}C^2 + C)/4C^2 < 1$):

$$w_+ \leq \tilde{w} \leq \bar{q}(\bar{\nu}) = w - (1-m_2)^2\frac{w^2}{8C^2} + (1-m_2)\frac{w}{4C^2}\Delta(\alpha_{+,+}-\sigma) + \Delta\sigma\,. \tag{2.36}$$

Ad (ii): We add again the index $k(\geq \bar{k})$ to w, σ, α, t, Δ, and C from (2.35). Since we only make Null Steps for $k \geq \bar{k}$, the t_k are monotonically decreasing from $\bar{k}$ on (see Fig. 2.1) and we conclude from (2.31) that

$$\Delta_k \to 0 \quad \text{as } k \to \infty\,. \tag{2.37}$$

By Proposition 2.7 the terms z_k, g_{k+1}, σ_k in (2.35) are bounded; this, together with (2.31), implies the existence of $\bar{C}$ with $\bar{C} \geq C_k$ for all k. Inequality (2.36) simplifies to

$$\begin{aligned} w_{k+1} \leq\, & w_k - (1-m_2)^2 \tfrac{w_k^2}{8} \bar{C}^{-2} \\ & +(1-m_2)\tfrac{w_k}{4} C_k^{-2} \Delta_k(\alpha_{k+1,k+1} - \sigma_k) + \Delta_k \sigma_k \quad \text{for } k \geq \bar{k}\,. \end{aligned} \tag{2.38}$$

We use Lemma 2.6 once more to see that $\{w_k\}_{k\in\mathbb{N}}$ is bounded. Let a be the greatest cluster point and assume

$$w_{k(i)+1} \to a \quad \text{for } i \to \infty\,.$$

Now let b be any other cluster point of the sequence $w_{k(i)}$, i.e., for a further subsequence we have

$$w_{k(i(j))} \to b \quad \text{for } j \to \infty\,.$$

From (2.37) and (2.38) we obtain for $j \to \infty$

$$a \leq b - [(1-m_2)^2 \tfrac{1}{8} \bar{C}^{-2}] b^2 + 0\,.$$

Since, by choice, $b \leq a$, this can hold only if $a = b = 0$. This proves $w_k \to 0$ and the assertion follows from (2.25) and the boundedness of the t_k. □

Our convergence results now follow easily.

THEOREM 2.10. *If $X^* \neq \emptyset$, then x_k converges to some $x^* \in X^*$ as $k \to \infty$.*

Proof. Obviously (2.26) holds and the x_k converge to some $\tilde{x}$ (Lemma 2.5). From (2.12) we get, for each k and with z_k, σ_k from (2.18)

$$z_k^T(x - x_k) \leq f(x) - f(x_k) + \sigma_k \quad \text{for all } x\,.$$

If we fix x and choose a subsequence as in Propositions 2.7–2.9, then we obtain for $k \to \infty$

$$0 \leq f(x) - f(\tilde{x})\,.$$

Hence $x^* := \tilde{x} \in X^*$. □

The above result can be supplemented as follows.

THEOREM 2.11. *If $X^* = \emptyset$, then $f(x_k)$ converges to* $\inf\{f(x) \mid x \in X\} \in [-\infty, \infty)$.

Proof. By construction, the $f(x_k)$ are monotonically decreasing. Now suppose the assertion not to be true, i.e., for some $\bar{x}$ one has $f(\bar{x}) \leq f(x_k)$ for all k. Just as above, we conclude that $x_k \to \tilde{x} \in X^*$, which contradicts $X^* = \emptyset$. □

2.5. Piecewise linear case. For *piecewise linear* convex functions

$$f(x) := \max\{a_j^T x - b_j \mid 1 \leq j \leq m\} \quad \text{with } a_j \in \mathbb{R}^n, \qquad b_j \in \mathbb{R}\ (1 \leq j \leq m).$$

Theorem 2.10 can be refined substantially. Suppose f is bounded below (which implies $X^* \neq \emptyset$ for piecewise linear f), choose $J_{\max} := n+2$ in (2.23), and organize the algorithm such that each g_i, $i \in J_k$, is some a_j, $1 \leq j \leq m$. For this aim we put in the reset step (3) of (2.23) $J := \{i \mid \lambda_{k,i} > 0\}$, where $d_k = -t_k \sum_{i\in J_k} \lambda_{k,i} g_i$ solves (2.6). By Carathéodory's theorem one can always find λ_k such that $|J| \leq n+1$. A quadratic programming method, which solves (2.6) with this purpose in mind, is given in Kiwiel [19]. Further, some *Haar condition* has to be satisfied. Denote by $I(x^*)$ the set of active indices for given $x^* \in X^*$ (i.e., $I(x^*) = \{i \mid f(x^*) = a_i^T x^* - b_i\}$) and assume:

(2.39) If $I \subset I(x^*)$ and $|I| \leq n$ then the a_i, $i \in I$, are linearly independent.

Then we can establish *finite convergence* for our algorithm.

THEOREM 2.12. *Suppose f is piecewise linear, bounded below, and* (2.39) *holds. Then $x_k \in X^*$ for some $k \in \mathbb{N}$.*

We omit the proof, which the reader can easily copy from the above discussion and the treatment of this topic in Chapter 2 of Kiwiel [18].

The numerical results from §4 will display the finite convergence convincingly.

3. BT-algorithm: The nonconvex case. We discuss the modifications necessary for nonconvex f. Throughout this section we assume that f is locally Lipschitzian and

$$\text{weakly semismooth,} \tag{3.1}$$

i.e., the directional derivative $f'(x;d) := \lim_{t\downarrow 0} t^{-1}[f(x+td) - f(x)]$ exists for all x and d, and $f'(x;d) = \lim_{t\downarrow 0} g(x+td)^T d$ where $g(x+td) \in \partial f(x+td)$.

3.1. Model and algorithm. For nonconvex f, the subgradient inequality (2.1) does not hold and the $\alpha_{k,i}$ may become negative. As a consequence, $f_{\text{CP}}(x_k;\cdot)$ is no longer an approximation of $f(x_k+\cdot) - f(x_k)$ from below; in particular, usually $f_{\text{CP}}(x_k;0) = \max\{-\alpha_{k,i}\} > f(x_k+0) - f(x_k)$. To cope with this difficulty we follow a strategy (also used in M1FC1) and replace $\alpha_{k,i}$ by

$$\beta_{k,i} := \beta(x_k, y_i) := \max\{\alpha_{k,i},\ c_0\|x_k - y_i\|^2\};$$

here c_0 is a fixed small positive real (and $c_0 := 0$ for convex f). By construction, $\beta_{k,i} \geq 0$ and the modified model

$$f_{\text{CP}}(x_k;d) := \max_{i\in J_k}\{g_i^T d - \beta_{k,i}\} \tag{3.2}$$

coincides again with $f(x_k+d) - f(x_k)$, at least at $d = 0$. The $\beta_{k,i}$ copy part of the role of the $\alpha_{k,i}$ in §2: Whenever y_i is "far away" from the current iterate x_k, then $\beta_{k,i}$ is large and thus g_i only plays a minor role in (3.2). However, we have to admit that the above f_{CP} is a much less satisfactory model in the nonconvex case.

Now replace, in §§2.1–2.3, the $\alpha_{k,i}$ by the new weights $\beta_{k,i}$. This does not change the character of (2.6) and (2.16), and thus Lemma 2.1 and the duality between (2.6) and (2.16) remain true. For (2.12), however, convexity was essential and as a consequence (2.14) and the corresponding criterion in (2.17)(1) no longer imply the ε-optimality for x_k. For nonconvex f the condition

$$\left\|\sum_{i\in J_k} \lambda_i g_i\right\| \leq \varepsilon \quad \text{and} \quad \sum_{i\in J_k} \lambda_i \beta_{k,i} \leq \varepsilon$$

merely says that 0 "lies up to ε" in the convex hull of certain $g_i \in \partial f(y_i)$ for which the "y_i are not far away from x_k" (since $\sum_{i\in J_k} \lambda_i \beta_{k,i} \leq \varepsilon$). This corresponds to "almost" stationarity in smooth optimization.

Iteration (2.17) requires two modifications. First, part (iii) in the proof of Theorem 2.3 does not carry over to nonconvex f. This difficulty is easily bypassed: We simply omit condition **SS**(ii) in step (2)(a) in (2.17) and skip step (2)(b). This does not affect the convergence analysis since the purpose of **SS**(ii) in (2.17) was of merely numerical nature.

Second, for nonconvex f, we have to add in (2)(c) the further Null Step condition

NS: (iii) $g^{j^T} d^j - \beta_{k,j} \geq m_2 v^j$.

NS(iii) guarantees that, after a Null Step, the updated model provides a direction d that differs from the unsuccessful previous one. This change in the direction played a crucial role in the convergence analysis (see, e.g., the proof of Proposition 2.9). For convex f, condition **NS**(iii) is automatically satisfied whenever we are on the right branch in Fig. 2.1; cf. (2.20). This is not true for nonconvex f and we have to add this as an additional condition. Unfortunately, this supplementary **NS**(iii) leads to a serious drawback of our method. For nonconvex f we can no longer guarantee the existence of $\tilde{T} > 0$ in the proof of Proposition 2.7 such that **NS**(ii) together with **NS**(iii) holds on the right branch of Fig. 2.1 for $t^j \leq \tilde{T}$. As a consequence we cannot assure any more that $z_k \to 0$ and $\sigma_k \to 0$ for the special case $t_k \to 0$, which had to be separated from the proof of Propositions 2.8 and 2.9.

As things stand now, we can propose only the following emergency exit. We add to **NS**(i) and **NS**(ii) the additional condition **NS**(iii) and split (2)(d) of (2.17) in two branches. Suppose **NS**(i) holds but **NS**(iii) does not: If the second condition of **NS**(ii) is not satisfied, then we are allowed to choose a smaller t^{j+1}; if the second condition of **NS**(ii) holds, then we make a *linesearch* along d^j, just as in M1FC1. More precisely, this yields the following.

(3.3) **Inner iteration** $x_k \to x_{k+1}$**:**

(0) Choose $t^1 := t_{k-1}$. Set $l^1 := 0$, $u^1 := T$, and $j := 1$.

(1) Compute the solution $(v^j, d^j) = (v(t^j), d(t^j))$ of (2.6) with $\alpha_{k,i}$ replaced by $\beta_{k,i}$. If $(1/t^j)\|d^j\| \leq \varepsilon$ and $-(1/t^j)\|d^j\|^2 - v^j \leq \varepsilon$, then stop: x_k is almost stationary. Otherwise put $y^j := x_k + d^j$ and compute $g^j \in \partial f(y^j)$.

(2) (a) If **SS**(i) holds, then make a **Serious Step**: Put $x_{k+1} := y_{k+1} := y^j$, $g_{k+1} := g^j$ and stop.

(b) If **NS**(i), **NS**(ii), and **NS**(iii) hold, then make a **Null Step**: Put $x_{k+1} := x_k$, $y_{k+1} := y^j$, $g_{k+1} := g^j$ and stop.

(c) If **NS**(i), **NS**(ii) hold but **NS**(iii) does not, then:

(i) if the second part of **NS**(ii) holds, then put $d_k := d^j$, $v_k := v^j$ and make a linesearch along $x_k + sd_k$, $s \geq 0$,

(ii) otherwise put $u^{j+1} := t^j$, $l^{j+1} := l^j$, $t^{j+1} := \frac{1}{2}(u^{j+1} + l^{j+1})$, $j := j+1$ and go back to (1).

(d) If **NS**(i) holds but **NS**(ii) does not, then put $u^{j+1} := t^j$, $l^{j+1} := l^j$, $t^{j+1} := \frac{1}{2}(u^{j+1} + l^{j+1})$, $j := j+1$ and go back to (1).

For weakly semismooth f (see (3.1)) the linesearch ends up in finitely many steps with a stepsize $s_k \geq 0$ such that in $y_{k+1} := x_k + s_k d_k$ and $g_{k+1} \in \partial f(y_{k+1})$, either the (short) serious criterion

SSS: (i) $f(y_{k+1}) - f(x_k) < m_1 s_k v_k$,

SSS: (ii) $g_{k+1}^T d_k \geq m_2 v_k$

is satisfied or **NS**(iii) and the first part of **NS**(ii) hold but **SSS**(i) is not satisfied. In case **SSS**(i) and **SSS**(ii) hold, we put $x_{k+1} := y_{k+1}$ and add g_{k+1} to the bundle; if **NS**(iii) and the first part of **NS**(ii) hold, we make a Null Step. All details concerning the linesearch can be found in Lemaréchal [22]. Semismoothness is a property with respect to halflines and this explains the success of a linesearch. The heart of our argument was simply not to restrict the search to a halfline; we wanted to work with various directions. Hence we consider a linesearch only as an emergency step that is against the spirit of our approach. And, indeed, if our method runs into numerical troubles, then usually this is because we had to switch to a linesearch that ends in a collapse.

In the overall algorithm (2.23) the updating of the $\beta_{k,i}$, together with the reset strategy, has to be adapted to the new situation. We do this just as in Kiwiel's aggregate subgradient method, where one avoids again the storing of the previous x_i and y_i; these technicalities are skipped here.

We mention that linear constraints can be added in (2.6) and (2.16) without major difficulties. Proposals on how to handle nonlinear constraints in the bundle framework have been made, e.g., by Kiwiel [18].

3.2. Convergence analysis. Suppose we use the above definition of $\beta_{k,i}$ and do not use a reset strategy. It is easily verified that the inner iteration (3.3) is again a finite process. Lemmas 2.4 and 2.5 rely decisively on the subgradient inequality and do not carry over to nonconvex f. Hence the statement of Lemma 2.5 now becomes an assumption:

$$\{x_k\}_{k\in\mathbb{N}} \text{ is bounded.} \tag{3.4}$$

With (2.26) replaced by (3.4), Lemma 2.6 remains true. The same holds for Propositions 2.7–2.9; of course, eventual linesearch steps have to be taken into account. Only the proof of Proposition 2.9 requires some technical modifications. For nonconvex f and $\varepsilon = 0$, we get the following convergence result (for details, see Schramm [37]).

THEOREM 3.1. *If f is weakly semismooth, bounded below, and* (3.4) *holds, then there exists a cluster point $\bar{x}$ of the sequence $\{x_k\}_{k\in\mathbb{N}}$ such that $0 \in \partial f(\bar{x})$.*

4. Numerical examples and applications. The above concept was implemented in FORTRAN 77 as **BTC** for convex f and **BTNC** for nonconvex f. The implemented reset strategy and a "safeguarded weighting technique" go back to proposals by Kiwiel [18], [21]. Further, we use a subroutine due to Kiwiel (see [19]) to solve the dual quadratic programming problem at each iteration.

Section 4.1 reports our experience with a collection of (non)convex academic test problems. In §4.2 we compute dual bounds for traveling salesman problems, and in §4.3 we deal with minimax eigenvalue problems for matrices coming from special graphs. In §4.4, finally, we present the results for a nonconvex and nonsmooth optimal design problem: The maximization of the area of contact for the deflection of a clamped beam.

All computations were done on a HP9000/330, respectively, on a VAX8600.

4.1. "Academic" testexamples. Convex examples. Table 4.2 presents a comparison of BTC with M1FC1 [27] for a collection of classical convex test examples listed in Table 4.1. The following abbreviations are used in Tables 4.1–4.4.

Dim	dimension of the problem,
f^*	(known) optimal value,
niter	number of iterations,
$\# f/g$	number of function/subgradient-evaluations,
f	computed approximation of f^*.

Many of the test examples are described in detail in [25] or [42]. Test function Mifflinl has been communicated to us by Mifflin [32]:

$$f(x) = -x_1 + 20 \max\{x_1^2 + x_2^2 - 1, 0\},$$

with starting point $(0.8, 0.6)^T$, minimum $x^* = (1,0)^T$, and $f(x^*) = -1$.

TABLE 4.1
List of convex examples.

Nr.	Problem	Dim	f^*
1	CB2 [3]	2	1.952225
2	CB3 [3]	2	2
3	DEM [6]	2	-3
4	QL [42]	2	7.2
5	LQ [42]	2	$-\sqrt{2}$
6	Mifflinl [32]	2	-1
7	Mifflin2 [15]	2	-1
8	Mak [29]	3	-132.0608
9	Rosen [3]	4	-44
10	Shor [39]	5	22.60016
11	Maxquad [25]	10	-0.8414084
12	Maxq [37]	20	0
13	Maxl [37]	20	0
14	Goffin [12]	50	0
15	TR48 [25]	48	-638565

Since BTC and M1FC1 proved to be strongly superior to all subgradient variants, which we tested, we restrict our comparison to BTC and M1FC1. In BTC we put $m_1 := 0.1$, $m_2 := 0.2$, $m_3 := 0.9$, $\varepsilon \leq 10^{-4}$. Furthermore we take $k$max(maximal number of subgradients):= 5 for CB2, CB3, DEM, QL, LQ, and $k$max := 10 for Mifflinl, Mifflin2, Rosen/Suzuki, Shor, Maxquad; for higher-dimensional examples we use $k$max := 200. The parameters in M1FC1 were chosen correspondingly. We mention that BTC works in double precision, whereas M1FC1 requires function- and subgradient-evaluations only in single precision. The function value and the corresponding subgradient are computed in one subroutine.

In all examples, BTNC reached the required accuracy; the same holds for M1FC1 apart from the data marked by "$*$," where M1FC1 broke down with linesearch difficulties. Obviously BTC often shows a better performance for the discussed examples than M1FC1. We note that results similar to ours are reported by Kiwiel [21] for his proximity control algorithm, which is closely related to our approach.

In §2.5 we discussed "finite convergence" of BTC for convex piecewise linear f. This finite convergence can be observed for the piecewise linear examples of Table 4.2. As an example, we give the finite behaviour for Goffin's test function [12]

$$f(x) = 50 \max_{1 \leq i \leq 50} x_i - \sum_{i=1}^{50} x_i ,$$

TABLE 4.2
Convex examples.

	BTC			M1FC1		
Nr.	niter	$\# f/g$	f	niter	$\# f/g$	f
1	13	16	1.952225	11	31	1.952253
2	13	21	2.000000	12	44	2.001415
3	9	13	-3.000000	10	33	-3.000000
4	12	17	7.200009	12	30	7.200018
5	10	11	-1.414214	16	52	-1.141420
6	49	74	-1.000000	143	281	-0.999967
7	6	13	-1.000000	30	71	-0.999993
8	24	28	-132.0608	3	5	-132.0608
9	22	32	-43.99998	22	61	-43.99998
10	29	30	22.60016	21	71	22.60018
11	45	56	-0.8414083	29	69	-0.8413589
12	125	128	0.0	144	207	0.0
13	74	84	0.0	138	213	0.0
14	51	53	0.0	72	94	0.00010
15	165	179	-638565.0	163	284	-633625.5*

with starting point $x^i = i - 25.5$, $i = 1, \cdots, 50$, and optimal value 0. We use the notation:

niter number of iterations,
ncomp number of function/subgradient-evaluations,
f function value at the current iterate,
gn $\| \sum_{i \in J_k} \lambda_{k,i} g_i \|$,
alpha $\sum_{i \in J_k} \lambda_{k,i} \alpha_{k,i}$.

```
BT-Algorithm  --  Goffin
========================


niter  ncomp        f               gn              alpha
  1      1    .12250000E+04   .49497475E+02   .00000000E+00
  2      2    .12250000E+04   .34641019E+02   .24989898E+02
  3      5    .11497487E+04   .33243153E+02   .20320902E+03
  4      6    .11497487E+04   .31880038E+02   .65908627E+02
  5      7    .10744975E+04   .27719626E+02   .23738585E+03
  6      8    .10744975E+04   .29102567E+02   .99746186E+02
  7      9    .99924621E+03   .24138528E+02   .24395184E+03
  8     10    .99924621E+03   .26342143E+02   .12550253E+03
  9     11    .92399495E+03   .21277728E+02   .23862048E+03
 10     12    .92399495E+03   .23604747E+02   .14317764E+03
  .      .          .               .               .
  .      .          .               .               .
 20     22    .57248737E+03   .89001539E+01   .18329243E+03
 30     32    .41997475E+03   .65860451E+01   .22510361E+03
 40     42    .24120581E+03   .42541465E+01   .19390350E+03
  .      .          .               .               .
  .      .          .               .               .
 45     47    .18869318E+03   .30403588E+01   .17830897E+03
 46     48    .18869318E+03   .29196707E+01   .17519938E+03
```

```
   47     49    .16243687E+03     .22448944E+01     .16184848E+03
   48     50    .16243687E+03     .21904615E+01     .15727798E+03
   49     51    .16243687E+03     .19993010E+01     .15532349E+03
   50     52    .13618055E+03     .11116780E+01     .13312204E+03
   51     53    .67968547E-12     .37861010E-14     .60867311E-12
convergence
```

The last step yields a "jump" to the optimum by adding the decisive subgradient information.

Nonconvex examples. In Table 4.4 we compare BTNC to M1FC1 (which can also deal with nonconvex f) for the problems given in Table 4.3. Here "Rosb" is the differentiable Rosenbrock example.

TABLE 4.3
List of nonconvex examples.

Nr.	Problem	Dim	f^*
1	Cres [18]	2	0
2	Mad [28]	2	0.6164324
3	Mabs [30]	2	0
4	ℓ_1 [8]	3	7.894231
5	ℓ_1a [8]	6	0.559814
6	Rosb	2	0

TABLE 4.4
Nonconvex examples.

	BTNC			M1FC1		
Nr.	niter	# f/g	f	niter	# f/g	f
1	24	27	$0.944280 \cdot 10^{-6}$	31	93	$0.225317 \cdot 10^{-5}$
2	21	22	0.6164324	17	41	0.6164330
3	30	39	$0.444089 \cdot 10^{-14}$	37	88	$0.111921 \cdot 10^{-7}$
4	21	23	7.894231	16	39	7.894232
5	73	78	0.559814	116	318	0.559814
6	79	88	$0.130389 \cdot 10^{-11}$	70	121	$0.243610 \cdot 10^{-6}$

Again, our bundle trust region version BTNC yields better results than M1FC1; this is further confirmed by some test runs done by Schittkowski [36]. Since, however, our experience with nonconvex problems is still rather limited, we do not claim this to be a final statement.

4.2. Traveling salesman problems. In many practical applications one has to solve a problem which can be phrased as a (symmetric) *traveling salesman problem*: Given a complete graph $K_n = (V, E)$ and distances c_{ij} for each edge $ij \in E$ (with $c_{ij} = c_{ji}$), find a tour T^* with length $c(T^*)$ as small as possible. Since problems of this type often appear in tremendous size (e.g., drilling problems with several thousands of knots), it is generally not possible to solve them exactly. Widely used tools in combinatorial optimization are therefore heuristics, which compute an approximate solution T rather quickly. To judge the quality of such a tour T, it is important to know a lower bound for the length $c(T^*)$ of the optimal tour. Such a bound can be found via the 1-*tree relaxation* described in Held and Karp [17]. We can formulate the TSP as a linear problem of the form

$$\min\{\langle c,x\rangle \mid Ax=a,\ Bx\le b, x_i\in\{0,1\}\}.$$

The following weak duality relation holds:

$$c(T^*)=\min\{c(T)\mid T \text{ is a tour}\}\ge \max\{\phi(\lambda)\ \mid\ \lambda\in\mathbb{R}^n\}, \tag{4.1}$$

where $\phi:\mathbb{R}^n\to\mathbb{R}$ is defined by

$$\phi(\lambda):=\min\{\langle c,x\rangle+\langle\lambda,Ax-a\rangle\ \mid\ Bx\le b,\ x_i\in\{0,1\}\}.$$

Without the binary constraints $x_i\in\{0,1\}$, one has even equality in (4.1). For our TSPs the gap was never greater than 1–3 percent. Hence we can find a good lower bound by maximizing the function ϕ which, as minimum of finitely many linear functions in λ, is nonsmooth, concave, and piecewise linear. It is known from combinatorics that $\phi(\lambda)$ can be computed via the length $\tilde c$ of a so-called minimum spanning 1-tree $x(\lambda)$ for our graph with the new distances $\tilde c_{ij}:=c_{ij}+\lambda_i+\lambda_j$; it holds $\phi(\lambda)=\tilde c-2\sum_{i=1}^n\lambda_i$. There are efficient algorithms to compute such a tree and thus $\phi(\lambda)$; we used the Prim algorithm. Simple subgradient calculus shows that this tree $x(\lambda)$ provides us, for free, with a subgradient of ϕ at λ: $\phi(\lambda)=\langle c,x(\lambda)\rangle+\langle\lambda,Ax(\lambda)-a\rangle$ and as a byproduct we obtain a subgradient of ϕ at λ

$$g(x(\lambda)):=A\,x(\lambda)-a\in\partial\phi(\lambda).$$

The components of $g(x(\lambda))$ are just the degrees of the knots of the 1-tree as follows:

$$g(x(\lambda))_i=\text{degree}(i)-2,\qquad i=1,\cdots,n.$$

Thus we are precisely in the framework (1.4) and can apply BTC or M1FC1.

The following subgradient variant

$$x_{k+1}:=x_k-M\rho^k(\alpha_kg_k+(1-\alpha_k)g_{k-1})/\|\alpha_kg_k+(1-\alpha_k)g_{k-1}\| \tag{4.2}$$

with fixed $M>0$, $0<\rho<1$ and $0<\alpha_k\le 1$ for all $k\in\mathbb{N}$ is currently the standard method in the TSP context. For the choice $\alpha_k=1$, $k\in\mathbb{N}$, one can establish convergence with geometric convergence speed (with factor $M\rho$) of the x_k to some limit which, however, need not be optimal. We believe that Table 4.6 will convince the reader to also consider more sophisticated methods like BTC or M1FC1. Presented are the results for a collection of synthetic examples (Krolak1, $\cdots$, Krolak5) and for some TSPs which come from drilling problems. Table 4.5 gives a list of the problems we treated. "Dim" in the third column is the number of knots (i.e., the dimension of our optimization problem). The fourth column gives the length of a tour, which is considered a good one (it is not known whether this tour is optimal).

Table 4.6 shows the results. Here "lb" is the lower bound which we obtained from the three methods; for (4.2) we tried several M's and ρ's and report our best results. Finally, "%" gives the remaining gap in percentage. In BTC we take $m_1:=0.01$, $m_2:=0.2$, and $m_3:=0.9$. BTC was stopped when the stopping criterion was satisfied with $\varepsilon:=10^{-4}$ for the smaller problems, respectively, $\varepsilon:=10^{-2}$ for the larger ones. Also, for M1FC1 we could satisfy a corresponding stopping criterion, apart from a few runs where the code broke down with linesearch difficulties close to the optimal point. The subgradient method was stopped when we observed no further progress

TABLE 4.5
List of traveling salesman problems.

Nr.	Problem	Dim	Tour
1	KROL1	100	21282
2	KROL2	100	22141
3	KROL3	100	20749
4	KROL4	100	21294
5	KROL5	100	22068
6	TSP442	442	5069
7	TSP1173	1173	57323
8	V362	362	1966
9	V614	614	2312
10	V1167	1167	5657
11	V2116	2116	6786

TABLE 4.6
Traveling salesman problems.

	Subgradient Method			M1FC1			BTC		
Nr.	#f/g	lb	%	#f/g	lb	%	#f/g	lb	%
1	194	20929	1.66	103	20938	1.62	58	20938	1.62
2	202	21648	2.23	606	21753	1.75	233	21833	1.39
3	264	20451	1.44	156	20473	1.33	79	20473	1.33
4	116	20951	1.61	326	21110	0.86	118	21142	0.71
5	183	21779	1.31	292	21784	1.29	136	21799	1.22
6	229	5043	0.51	248	5033	0.71	378	5051	0.36
7	78	56351	1.70	621	56193	1.97	399	56386	1.63
8	161	1941	1.27	360	1930	1.83	285	1942	1.22
9	129	2253	2.55	255	2250	2.68	179	2254	2.51
10	141	5579	1.38	442	5564	1.64	506	5580	1.36
11	109	6599	2.76	668	6579	3.05	713	6606	2.65

in the leading digits; obviously we have convergence to a nonoptimal point, e.g., for KROL2.

Finite convergence was again observed for many of the TSPs (recall that ϕ is concave and piecewise linear). Below we give the result for KROL1.

```
BT-Algorithm  --  KROL1
======================


niter  ncomp        f                gn              alpha
  1      1    .19094198E+05    .66332496E+01    .00000000E+00
  2      3    .19370920E+05    .52129811E+01    .63650738E+03
  3      4    .19654650E+05    .41342650E+01    .72867439E+03
  4      5    .20150557E+05    .35600250E+01    .52416612E+03
  5      6    .20295988E+05    .23095560E+01    .69738857E+03
  .      .          .                .                .
 10     11    .20551025E+05    .14857001E+01    .30072466E+03
  .      .          .                .                .
 20     21    .20823536E+05    .71434597E+00    .11684214E+03
```

```
        .     .       .                .                .
       30    31    .20893782E+05    .47437971E+00    .69645027E+02
        .     .       .                .                .
       40    42    .20926464E+05    .58577127E+00    .88800444E+01
        .     .       .                .                .
       50    52    .20934447E+05    .17380992E+00    .32980892E+01
       51    53    .20935444E+05    .13554464E+00    .23477955E+01
       52    54    .20935905E+05    .99386622E-01    .19572159E+01
       53    55    .20935905E+05    .96884277E-01    .19192909E+01
       54    56    .20937609E+05    .15142964E-01    .31555350E+00
       55    57    .20937609E+05    .51250884E-02    .31600774E+00
       56    58    .20937926E+05    .39299811E-15    .63493827E-11
   convergence
```

4.3. Minimizing the maximal eigenvalue. Often an application requires the solution of the subproblem

$$(\mathrm{P}_\lambda) \qquad \text{minimize } f(x) := \lambda_{\max}(A(x))\,;$$

here $A(\cdot)$ is a real symmetric $m \times m$-matrix, which depends linearly on $x \in \mathbb{R}^n$, and $\lambda_{\max}(A(x))$ denotes the maximal eigenvalue of $A(x)$. The following properties hold (see, e.g., [5]):

- f is convex;
- f is nonsmooth at x, if the maximal eigenvalue $f(x)$ has multiplicity greater than 1;
- if u is eigenvector of $A(x)$ for the eigenvalue $f(x)$ and $\|u\|_2 = 1$, then a subgradient of f at x can be easily computed from the dyadic product uu^T.

Hence we are again in the situation (1.4) and can attack (P_λ) with bundle-type methods.

We encountered such problems in connection with

(i) stable sets of graphs,

(ii) experimental design (see Gaffke and Mathar [10]).

First, numerical steps for (ii) are reported in [1]. Here we consider (i) more closely. The theoretical background is discussed in detail in a book by Grötschel, Lovász, and Schrijver [14], who brought this subject to our attention. Let $G = [V, E]$ be a graph, $w = (w_1, \cdots, w_{|V|})^T$ a vector in $\mathbb{R}^{|V|}$ with nonnegative components, and put $\bar{w} := (\sqrt{w_1}, \cdots, \sqrt{w_{|V|}})^T$. We want to compute the so-called *theta-function* $\vartheta(G; \cdot)$,

$$\vartheta(G; w) := \min_{A \in M} \lambda_{\max}(A + W)\,,$$

where $W = \bar{w}\bar{w}^T$ and

$$\begin{aligned} M := \{B\,|B &\text{ symmetric } n \times n\text{-matrix}, \\ &b_{ii} = 0 \text{ for } i \in V,\ b_{ij} = 0 \text{ for } i, j \text{ nonadjacent}\}\,. \end{aligned}$$

The theta-function is the support function of the convex set TH(G) (a set that contains the convex hull of the incidence vectors of all stable sets of G). Its value is known for some special cases (let $w = (1, \cdots, 1)^T$):

(a) If G is a circle with an odd number n of knots, then

$$\vartheta(G; w) = \frac{n \cos \dfrac{\pi}{n}}{1 + \cos \dfrac{\pi}{n}}\,;$$

(b) if G is an Erdös–Ko–Rado graph $K(n,r)$, then

$$\vartheta(K(n,r);w) = \binom{n-1}{r-1}.$$

To compute $\vartheta(G;w)$, we have to solve the nonsmooth convex problem

$$(\mathrm{P}^{\vartheta}) \qquad \text{minimize} \quad \lambda_{\max}(A+W) \quad \text{subject to } A \in M.$$

Since W is constant and the constraints only require A to be symmetric and some components of A to be zero, (P^{ϑ}) can be phrased as an unconstrained minimization problem of form (P_{λ}), where $A = A(x)$ and $x \in \mathbb{R}^m$ corresponds to the free components of A. The dimension is equal to

$$m = \frac{n(n-1)}{2} - |\{(ij)\,|\,i,j \text{ nonadjacent }\}|.$$

Table 4.7 shows some results for (a). Note that for circles the dimension of the optimization problem equals the number of knots n. The starting point is always $(-1,\cdots,-1)^T$ and $k$max := 20. The value ϑ in the second column is the precise value of $\vartheta(G;w)$.

TABLE 4.7
Odd circles.

		BTC			M1FC1		
n	ϑ	niter	# *f*/*g*	*f*	niter	# *f*/*g*	*f*
17	8.42701	19	23	8.42705	28	70	8.42706
23	11.44619	26	30	11.44619	42	104	11.44626
39	19.46833	42	44	19.46833	37	102	19.46837
55	27.47756	49	49	27.47756	52	132	27.47756
111	55.48889	50	50	55.48889	58	134	55.48897

Table 4.8 gives the corresponding results for (b). Here we put $k$max := 20 for dimension $M \leq 50$ and $k$max := 50 otherwise. In the last two examples (∗), M1FC1 breaks down with linesearch difficulties.

The results show clearly what was observed above: For convex f the code BTC seems to be superior to M1FC1.

TABLE 4.8
Erdös–Ko–Rado graphs.

				BTC			M1FC1		
n	r	Dim	ϑ	niter	# *f*/*g*	*f*	niter	# *f*/*g*	*f*
5	2	15	4	28	28	4.000003	25	52	4.000047
6	2	45	5	33	33	5.000013	33	62	5.134571
10	2	630	9	63	63	9.000008	75	95	9.000071
9	3	840	28	63	65	28.000095	28	36	53.000003*
10	4	1575	84	125	128	84.000409	37	48	163.333332*

Table 4.9 discusses the improvement of BTC, if we add all subgradients, which we obtain from a system of orthonormal eigenvectors for $f(x)$, to the bundle at x. We considered eigenvalues as equal if they differ in value less than 10^{-9}. The extreme improvement is probably due to some hidden structure of the problem, since the same technique applied to random graphs only leads to no improvement.

TABLE 4.9
Circles—enlarged information.

	BTC			modified BT		
n	niter	$\#f/g$	f	niter	$\#f/g$	f
17	12	15	8.42701	3	6	8.42701
23	29	29	11.44619	2	3	11.44619
39	30	33	19.46833	3	3	19.46833
55	49	49	27.47756	3	3	27.47756
111	50	50	55.48889	3	7	55.48888

4.4. Maximization of the contact area between a clamped beam and a rigid obstacle. Let $y(s)$ for $0 \le s \le 1$ be the deflection (*state variable*) of a clamped beam under the load $u(s)$ (*design variable*). The total amount of load $\int_0^1 u(s)ds$ is given and $u(s)$ is bounded from above by some β for each s; further, the deflection of the beam is limited from below by some rigid obstacle $g(s)$. Then the aim is to find a load density such that the beam comes as "close" to the obstacle as possible. In an abstract setting the problem becomes

$$\text{minimize} \int_0^1 (y(s) - g(s))^2 ds \quad \text{subject to} \tag{4.3}$$
$$\langle Ay, z - y\rangle \ge \langle B(x), z - y\rangle \quad \text{for all } z \in K, \qquad x \in X_{ad} \qquad (x \in X,\ y \in Y).$$

Here $X := L^2((0,1))$ and $Y := H_0^2((0,1))$ are the control space and state space, respectively. $X_{ad} := \{x \in L^\infty((0,1)) \mid 0 \le x(s) \le \beta$ almost everywhere in $(0,1)$ and $\int_0^1 x(s)ds = M\}$ and $K := \{z \in Y \mid z(s) \ge g(s)$ almost everywhere in $(0,1)\}$ are the sets of feasible controls and admissible state variables. The operator $A : Y \to Y'$ is defined by $Ay = y^{iv}$ and $B : X \to H^{-2}((0,1))$ is the natural embedding. The variational inequality (4.3) assigns to a given load x the deflection y; it is known that (4.3) can be rewritten as a quadratic programming problem

$$y \in \arg\min_{z\in K} \frac{1}{2}\langle z, Az\rangle - \langle B(x), z\rangle . \tag{4.4}$$

The above problem is thoroughly discussed in a more general framework in [16] and [33]. The discretization below follows these references.

For a numerical treatment we divide the interval $[0,1]$ into n equidistant subintervals of length $1/n$ and consider design functions which are a constant x_i on each subinterval i. For the controls we use a standard finite-element technique and work with functions

$$y(s) = \sum_{i=1}^{2n-2} y_i \varphi_i(s) ,$$

where the φ_i are third-order polynomials chosen such that $y(i/n) = y_{2i-1}$ and $y'(i/n) = y_{2i}$ for $i = 1, \cdots, n-1$. In this framework K becomes $\{z \in \mathbb{R}^{2n-2} \mid z_{2i-1} \ge g(i/n)$ for $i = 1, \cdots, n-1\}$ and X_{ad} reduces to $\{x \in \mathbb{R}^n \mid \sum(x_i/n) = M,\ 0 \le x_i \le \beta$ for $i = 1, \cdots, n\}$. With the positive-definite $(2n-2) \times (2n-2)$ rigidity matrix H (built up from terms $\int \varphi_i'' \varphi_j''$) and the $(2n-2)$-vector $b(x)$ (with elements $\int x\varphi_i$), the discretized

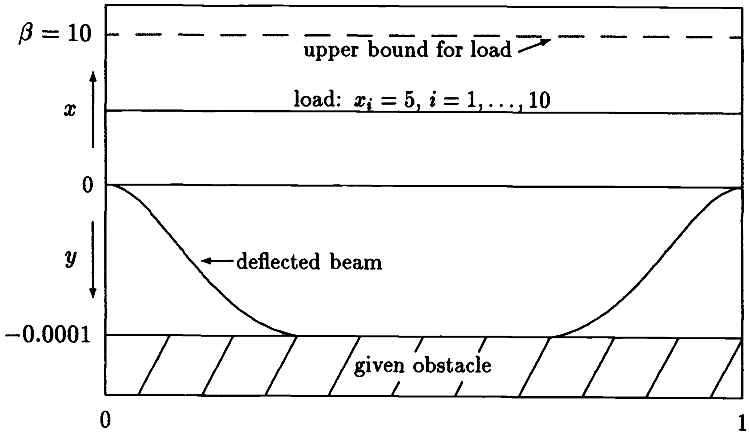


FIG. 4.1. *Beam (constant load).*

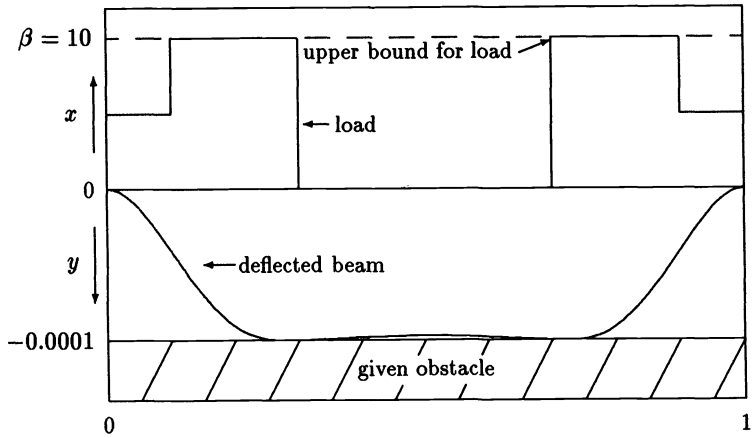


FIG. 4.2. *Beam (optimized load).*

problem becomes

$$(4.5)\quad \text{minimize} \int_0^1 \sum_{i=1}^{2n-2} (y_i\varphi_i(s) - g(s))^2 ds \quad \text{subject to}$$

$$(4.6)\quad y \in \arg\min_{z\in K} \frac{1}{2} z^T H z - b(x)^T z, \qquad x \in X_{ad}, \qquad (x \in \mathbb{R}^n,\ y \in \mathbb{R}^{2n-2}).$$

Now let $y(x) \in \mathbb{R}^{2n-2}$ denote the unique solution of (4.6) for given $x \in \mathbb{R}^n$ and write $l(y)$ for the integral in (4.5). Then our problem becomes

$$(4.7)\qquad \text{minimize } f(x) := l(y(x)) \quad \text{subject to } x \in X_{ad}.$$

Obviously f is not convex; further, because of the constraint in (4.6), the function $y(\cdot)$ (and thus also f) depends in a nonsmooth way on x. To compute $f(x)$ we have

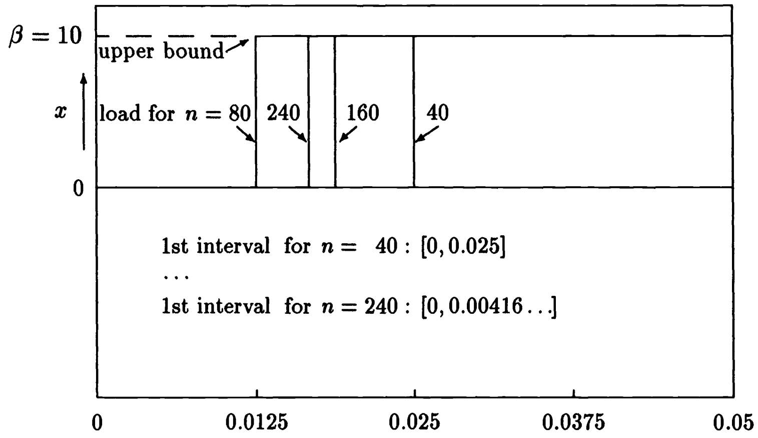

$\beta = 10$
upper bound
x
load for $n = 80$
240
160
40
0
1st interval for $n =$ 40 : [0, 0.025]
...
1st interval for $n = 240$: [0, 0.00416...]
0
0.0125
0.025
0.0375
0.05

FIG. 4.3. *Load for different discretizations.*

to solve the quadratic programming problem (4.6). With its solution $y = y(x)$ and a Lagrange multiplier λ put

$$\tilde{K} := \left\{ z \in \mathbb{R}^{2n-2} \,|\, z_{2i-1} \text{ satisfies (4.8) for } i = 1, 2, \cdots, n-1 \right\} ,$$

$$z_{2i-1} \begin{cases} = 0, & \text{if } y_{2i-1} = g(\frac{i}{n}) \text{ and } \lambda_{2n-i} > 0 , \\ \geq 0, & \text{if } y_{2i-1} = g(\frac{i}{n}) \text{ and } \lambda_{2n-i} = 0 , \end{cases} \tag{4.8}$$

and solve the derived quadratic programming problem

$$\min_{z \in \tilde{K}} \tfrac{1}{2} z^T H z - \nabla_y l(y(x))^T z . \tag{4.9}$$

Then, under some technical assumptions (which we skip here), the following is proved in [33]: If p solves (4.9), then $\nabla b(x)^T p$ is a subgradient of f at x. Hence at every iteration we must solve two quadratic programming problems of dimension $2n - 2$ to compute $f(x)$ and one $g \in \partial f(x)$. For these quadratic subproblems we use a code due to Powell. In our experiments we have set $\beta = 10$, $M = 5$, $g = -0.001$, and $n = 10$, 20, 40, 80, 160, 240. Further, we have incorporated the simple linear constraint $(1/n) \sum_{i=1}^{n} x_i = 5$ into the BTNC code itself. Figure 4.1 shows the deflection of the beam for $n = 10$ and for given $x = (5, 5, \cdots, 5)^T$ without optimizing; Fig. 4.2 gives the result for the "optimal" $x = (5, 10, 10, 0, \cdots, 0, 10, 10, 5)^T$ provided by BTNC.

We mention that these results differ substantially from those shown in [16] and [33]; a restart with BTNC from the data in that place proves that the results in [16] and [33] are not yet optimal. The "correctness" of our outcome is further confirmed by the BTNC solution for $n = 40$, 80, 160, and 240 (compare Fig. 4.3):

$$\begin{aligned} x &= (x_1, x_2, \cdots, x_{11}, x_{12}, \cdots)^T = (0, 10, \cdots, 10, 0, \cdots)^T \in \mathbb{R}^{40}, \\ x &= (x_1, x_2, \cdots\cdot, x_{21}, x_{22}, \cdots)^T = (0, 10, \cdots\cdot, 10, 0, \cdots)^T \in \mathbb{R}^{80}, \\ x &= (x_1, x_2, x_3, x_4, \cdots\cdots, x_{43}, x_{44}, \cdots)^T = (0, 0, 0, 10, \cdots\cdots, 10, 0, \cdots)^T \in \mathbb{R}^{160}, \\ x &= (x_1, x_2, x_3, x_4, x_5, \cdots\cdots, x_{64}, x_{65}, \cdots)^T = (0, 0, 0, 0, 10, \cdots\cdots, 10, 0, \cdots)^T \in \mathbb{R}^{240}. \end{aligned}$$

The stopping criterion was always $\varepsilon = 10^{-9}$.

Acknowledgment. We are indebted to the referees for their extremely careful reading of this paper and their valuable suggestions for improving the paper. We are also grateful to M. Grötschel, of the University of Augsburg, and U. Derigs, of the University of Bayreuth, for providing us with the data used in §4.2.

REFERENCES

[1] W. Achtziger, *Nichtglatte Optimierung—Ein spezielles Problem bei der Durchführung von Versuchen,* Diplomarbeit, Universität Bayreuth, Bayreuth, Germany, 1989.

[2] B.M. Bell, *Global convergence of a semi-infinite optimization method*, Appl. Math. Optim., 21 (1990), pp. 69–88.

[3] J. Charalambous and A.R. Conn, *An efficient method to solve the minimax problem directly*, SIAM J. Numer. Anal., 15 (1978), pp. 162–187.

[4] F.H. Clarke, *Nonsmooth Analysis and Optimization*, Wiley–Interscience, New York, 1983.

[5] J. Cullum, W.E. Donath, and P. Wolfe, *The minimization of certain nondifferentiable sums of eigenvalues of symmetric matrices*, Math. Programming Stud., 3 (1975), pp. 35–55.

[6] V.F. Demyanov and V.N. Malozemov, *Introduction to Minimax*, John Wiley, New York, Toronto, 1974.

[7] U. Derigs, *Another composite heuristic for solving Euclidean traveling salesman problems*, Preprint, Department of Mathematics, University of Maryland, College Park, MD, 1982.

[8] R.A. El-Attar, M. Vidyasagar, and S.R.K. Dutta, *An algorithm for ℓ_1-norm minimization with application to nonlinear ℓ_1-approximation*, SIAM J. Numer. Anal., 16 (1979), pp. 70–86.

[9] Yu.Y. Ermoliev, *Stochastic Programming Methods*, Nauka, Moscow, U.S.S.R., 1976.

[10] N. Gaffke and R. Mathar, *Linear minimax estimation under ellipsoidal parameter space and related Bayes* L-*optimal design*, DFG-Schwerpunktprogramm "Anwendungsbezogene Optimierung und Steuerung," Report 42, 1988.

[11] J.L. Goffin, *Affine methods in nondifferentiable optimization*, Tech. Report, McGill University, Montreal, Québec, Canada, 1987.

[12] ———, Private communication.

[13] J.L. Goffin, A. Haurie, and J.P. Vial, *Decomposition and Nondifferentiable Optimization with the Projective Algorithm*, GERAD, Faculty of Management, McGill University, Montreal, Québec, Canada, 1989.

[14] M. Grötschel, L. Lovász, and A. Schrijver, *Geometric Algorithms and Combinatorial Optimization*, Springer-Verlag, Berlin, 1988.

[15] N. Gupta, *A Higher than First Order Algorithm for Nonsmooth Constrained Optimization*, Ph.D. thesis, Department of Philosophy, Washington State University, Pullman, WA, 1985.

[16] J. Haslinger and P. Neittaanmäki, *Finite Element Approximation for Optimal Shape Design*, John Wiley, Chichester, U.K., 1988.

[17] M. Held and R.M. Karp, *The traveling-salesman problem and minimum spanning trees: Part* 2, Math. Programming, 1 (1971), pp. 6–25.

[18] K.C. Kiwiel, *Methods of Descent for Nondifferentiable Optimization*, Springer-Verlag, Berlin, New York, 1985.

[19] ———, *A method for solving certain quadratic programming problems arising in nonsmooth optimization*, IMA J. Numer. Anal., 6 (1986), pp. 137–152.

[20] ———, *An ellipsoid trust region bundle method for nonsmooth convex optimization*, SIAM J. Control Optim., 27 (1989), pp. 737–757.

[21] ———, *Proximity control in bundle methods for convex nondifferentiable optimization*, Math. Programming, 46 (1990), pp. 105–122.

[22] C. Lemaréchal, *A view of line searches*, in Optimization and Optimal Control, Lecture Notes in Control and Information Sciences, W. Oettli and J. Stoer, eds., Springer-Verlag, Berlin, Heidelberg, 1981.

[23] ———, *Extensions diverses des méthodes de gradient et applications*, Thèse d'Etat, Paris, 1980.

[24] ———, *Nondifferentiable optimization*, in Handbooks in Operations Research and Management Science, Vol. 1, Optimization, G.L. Nemhauser, A.H.G. Rinnooy Kan, and M.J. Todd, eds., North-Holland, Amsterdam, 1989.

[25] C. Lemaréchal and R. Mifflin, *A Set of Nonsmooth Optimization Test Problems*, in Nonsmooth Optimization, C. Lemaréchal and R. Mifflin, eds., Pergamon Press, Oxford, 1978, pp. 151–165.

[26] C. Lemaréchal, J.-J. Strodiot, and A. Bihain, *On a bundle algorithm for nonsmooth opti-*

mization, in Nonlinear Programming 4, O.L. Mangasarian, R.R. Meyer, and S.M. Robinson, eds., Academic Press, New York, 1981.

[27] C. Lemaréchal and M. Bancora Imbert, *Le Module* M1FC1, Tech. Report, Institut de Recherche d'Informatique et d'Automatique, Le Chesnay, 1985.

[28] K. Madsen and H. Schjaer–Jacobsen, *Linearly constrained minimax optimization*, Math. Programming, 14 (1978), pp. 208–223.

[29] M.M. Mäkelä, *Methods and algorithms for nonsmooth optimization*, Reports on Applied Mathematics and Computing, No. 2, Department of Mathematics, University of Jyväskylä, Jyväskylä, Finland, 1989.

[30] ———, *On the Methods of Nonsmooth Optimization*, Proc. 14th International Federation on Information Processing Conference on System Modelling and Optimization, Leipzig, Germany, 1989, to appear.

[31] R. Mifflin, *A modification and an extension of Lemaréchal's algorithm for nonsmooth optimization*, Math. Programming Stud., 17 (1982), pp. 77–90.

[32] ———, Private communication.

[33] J.V. Outrata, *On the numerical solution of a class of Stackelberg problems*, Zeitschrift für Operations Research, 4 (1990), pp. 255–278.

[34] B.T. Poljak, *Subgradient methods: A survey of Soviet research*, in Nonsmooth Optimization, C. Lemaréchal and R. Mifflin, eds., Pergamon Press, Oxford, U.K., 1978, pp. 5–29.

[35] R.T. Rockafellar, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1972.

[36] K. Schittkowski, *Heuristic reasoning in mathematical programming*, DFG-Schwerpunktprogramm "Anwendungsbezogene Optimierung und Steuerung", Report 209, 1990.

[37] H. Schramm, *Eine Kombination von Bundle- und Trust-Region-Verfahren zur Lösung nichtdifferenzierbarer Optimierungsprobleme*, Bayreuther Mathematische Schriften, Heft 30, Bayreuth, Germany, 1989.

[38] H. Schramm and J. Zowe, *A combination of the bundle approach and the trust region concept*, in Mathematical Research, Vol. 45, Advances in Mathematical Optimization, J. Guddat et al., eds., Akademie-Verlag, Berlin, 1988, pp. 196–209.

[39] N.Z. Shor, *Minimization Methods for Nondifferentiable Functions*, Springer-Verlag, Berlin, Heidelberg, 1985.

[40] G. Sonnevend and J. Stoer, *Global ellipsoidal approximations and homotopy methods for solving convex analytic programs*, Appl. Math. Optim., 21 (1990), pp. 139–165.

[41] Ph. Wolfe, *A method of conjugate subgradients for minimizing nondifferentiable convex functions*, Math. Programming Stud., 3 (1975), pp. 145–173.

[42] J. Womersley, *Numerical methods for structured problems in nonsmooth optimization*, Ph. D. thesis, Mathematics Department, University of Dundee, Dundee, Scotland, 1981.

[43] J. Zowe, *Nondifferentiable optimization—A motivation and a short introduction into the subgradient- and the bundle-concept*, in NATO ASI Series, Vol. F15, Computational Mathematical Programming, K. Schittkowski, ed., Springer-Verlag, Berlin, Heidelberg, 1985.

[44] ———, *The* BT-*algorithm for minimizing a nonsmooth functional subject to linear constraints*, in Nonsmooth Optimization and Related Topics, F.H. Clarke, V.F. Demyanov, and F. Gianessi, eds., Plenum Press, New York, 1989, pp. 459–480.

SIAM J. OPTIMIZATION
Vol. 2, No. 1, pp. 153–171, February 1992

NECESSARY OPTIMALITY CONDITIONS FOR NONSMOOTH MULTICRITERIAL OPTIMIZATION PROBLEMS*

TILO STAIB†

Abstract. Necessary optimality conditions of the first order are derived for nonsmooth nonconvex constrained optimization problems where the cost mapping is vector-valued and all occurring spaces are infinite-dimensional. These necessary conditions are given in the Karush–Kuhn–Tucker formulation and hold for various optimality concepts as proper and weak efficiency. An investigation and comparison of different constraint qualifications is also included.

Key words. multicriterial optimization, nonsmooth optimization, constraint qualifications

AMS(MOS) subject classifications. 90C31, 49B27

1. Introduction. The theory of nonconvex and nonsmooth optimization problems has made rapid progress in the last 20 years and problems with a single objective (cost) function have been fairly well analyzed. This is evidenced by the works of, e.g., Clarke [9], [7]; Demjanov [10]; Ioffe [16], [17]; Robinson [31]–[34]; Rockafellar [35], [36]; and Warga [43] and many others.

The general intention is to introduce a differentiability concept for a nonsmooth (i.e., non-Fréchet-differentiable) and nonconvex functional $f : X \rightarrow I\!R$ where X is a normed space, and to find a set of linear approximations for its derivative that contains enough information about the local behavior of the problem to provide, e.g., necessary or sufficient optimality conditions that are applicable to control problems or can be used to study the stability of the problem.

It has turned out that a useful notion is Clarke's derivative:

$$f^\circ(x,h) := \lim_{\epsilon,\lambda \to 0_+} \sup \left\{ \theta^{-1}(f(v+\theta h) - f(v)) \;\middle|\; 0 < \theta \leq \lambda, \|v - x\| \leq \epsilon \right\},$$

which exists if f is locally Lipschitz. The mapping $f^\circ(x,\cdot)$ is always sublinear and hence has a convex subdifferential. Moreover, f° coincides with the directional derivative if f is convex and with the gradient if f is strictly differentiable. This notion can readily be generalized for $f : X \rightarrow I\!R^n$ and the derivative can be approximated by gradients if $X = I\!R^m$. However, there is no general agreement as to how to define a similar derivative if $f : X \rightarrow Y$ and Y is an infinite-dimensional vector space and it seems impossible to recover all the useful properties mentioned above.

In the past decade there have been several attempts to define such an extension of Clarke's derivative. One of the most general notions (and perhaps the first) was given by Kusraev [23]; others were given by Thibault [40], Papageorgiou [26], and Reiland [30]. Such an extension is desirable to derive Karush–Kuhn–Tucker conditions for an extended class of constrained optimization problems, where the constraints and eventually the objective mapping are nonsmooth operators with infinite-dimensional range. Here it is essential to avoid strong assumptions like convexity or Fréchet-differentiability.

If one applies the necessary conditions of [23], [40], [26], and [30] to nonsmooth optimization problems over function spaces (e.g., optimal control problems), one realizes that their notions impose assumptions that are often hard to verify even for

* Received by the editors November 30, 1989; accepted for publication (in revised form) February 26, 1991.

† Institut für Angewandte Mathematik, Universität Erlangen–Nürnberg, Martensstrasse 3, 8520 Erlangen, West Germany.

simple mappings in L^p-spaces. Moreover, their derivatives generally do not coincide with the directional derivative, the Gâteaux derivative or the Fréchet derivative, even if those exist. We will demonstrate this by simple examples.

In order to improve this, we introduce in §2 another notion of generalized differentiability, being in some aspects similar to the mentioned ones, but being more flexible and covering the smooth as well as the nonsmooth cases. Using this notion we will derive necessary opimality conditions (i.e., Karush–Kuhn–Tucker conditions) for multicriterial optimization problems with nonsmooth constraints, where all occurring spaces may be infinite-dimensional (§3). It will turn out that an appropriate approach to the constrained problem is that of Guignard [12], introduced for Fréchet-differentiable problems and also used in [4] and [30]. This approach is slightly different from the Dubovitskii–Miljutin theory (see, e.g., [11] or [21]).

Further, we will investigate how the constraint qualifications used in this approach relate to others given in the literature, e.g., Slater's, and give some examples of how to verify them.

2. Generalized derivatives. Let X be a topological vector space (tvs) with dual X^* and $(Y, \mathcal{T})$ be an order-complete topological vector lattice with a given topolgy $\mathcal{T}$. Topological notions, such as general limits, will be prefixed with $\mathcal{T}$ only when $\mathcal{T}$ denotes the strong topology; then the prefix will be omitted. The order structure of Y is assumed to be generated by a convex pointed cone C_Y that is $\mathcal{T}$- normal, i.e., $\mathcal{T}$ has a base $\mathcal{U}$ of zero-neighborhoods with the property $U = (U + C_Y) \cap (U - C_Y)$ for all $U \in \mathcal{U}$ (see, e.g., [6]). The operations sup, inf, and $|\cdot|$ and $\leq_{C_Y}$ should be understood with respect to this ordering, as well as monotonicity, normality, increasing, order-bounded, and so on (see, e.g., [27] or [6]). The dual cone $C_{Y^*} \subseteq Y^*$ is defined by $C_{Y^*} := \{ y^* \in Y^* \mid y^*(c) \geq 0 \ \forall c \in C \}$ and its quasi interior by $C^i_{Y^*} := \{ y^* \in Y^* \mid y^*(c) > 0 \ \forall c \in C, c \neq 0 \}$. Order intervals will be the sets $[y_1, y_2]_{C_Y} := \{ u \in Y \mid y_1 \leq_{C_Y} u \leq_{C_Y} y_2 \}$. Finally, $f : X \to Y$ will be a given mapping.

DEFINITION 2.1. Let $x \in X$. A filter $\mathcal{U}(x) = \{U \subseteq X\}$ is called *contraction system* around x, if $\{x\} = \bigcap \{ U \mid U \in \mathcal{U}(x) \}$ holds and *monotonic contraction system* (MCS), if additionally the sets U depend on a real parameter in the following monotonic manner:

$$\mathcal{U}(x) := \{ U(x,\epsilon) \subseteq X \mid \epsilon \in I\!R, \ \epsilon > 0 \}, \quad \epsilon_1 \leq \epsilon_2 \Rightarrow U(x, \epsilon_1) \subseteq U(x, \epsilon_2).$$

If $V \subseteq X$ is a cone and if the sets $U(x, \epsilon)$ have the additional property that for every $k \in V$, $\epsilon > 0$, there exist real numbers $\delta(\epsilon, k), \gamma(\epsilon, k) > 0$ satisfying $\lim_{\epsilon \to 0} \delta(\epsilon, k) = \lim_{\epsilon \to 0} \gamma(\epsilon, k) = 0$ and $w := u + \theta k \in U(x, \gamma(\epsilon, k))$ for all $u \in U(x, \epsilon), 0 < \theta < \delta(\epsilon, k)$, then we call $\mathcal{U}(x)$ V-*stable.*

Of course, the family of all ϵ-balls in a normed space is a monotonic contraction system, but this is by no means the only example.

Example 2.1. Every system $\mathcal{U}(x)$ constructed of one of the family of sets $U(x, \epsilon)$ given below, is a MCS. The one of (a) is L^∞-stable, (c) is X-stable, (b) is C_X-stable and X-stable if $\operatorname{int} C_X \neq \emptyset$, (d) is 0-stable.

(a) Let $\Omega \subseteq I\!R^n$ be nonempty, μ the Lebesgue measure on Ω, and let X be a linear space of functions $\Omega \to I\!R$.

$$U(x,\epsilon) := \{ u \in X \mid |x(\omega) - u(\omega)| \leq \epsilon \ \ \forall \omega \in \Omega \} \quad \text{or} \tag{1}$$

$$U(x,\epsilon) := \{ u \in X \mid |x(\omega) - u(\omega)| \leq \epsilon \ \mu \text{ a.e. } \omega \in \Omega \}. \tag{2}$$

(b) Let (X, C_X) be a partially ordered tvs and let $c \in C_X$:

$$U(x,\epsilon) := \{ u \in X \mid x - \epsilon c \leq u \leq x + \epsilon c \}.$$

(c) Let X be normed. $U(x,\epsilon) := \{ u \in X \mid \|x-u\| \leq \epsilon \}$.

(d) $U(x,\epsilon) := \{x\}$.

(e) For every subspace $V \subseteq X$ the sets $V \cap U(x,\epsilon)$, where the sets $U(x,\epsilon)$ are taken as in (a), (b), (c), or (d), form a monotonic contraction system.

With the aid of a MCS we can now define a general continuity property of Lipschitz type and a generalized derivative of f, which are related to the topological or the order structure (or both) depending on the choice of the contraction system.

DEFINITION 2.2. The mapping $f : X \to Y$ is called *$\mathcal{U}$-Lipschitz-continuous* at $x \in X$, if there exist a MCS $\mathcal{U}(x)$, real numbers $\overline{\lambda}, \overline{\epsilon} > 0$, and a mapping $P : X \times \mathbb{R}^2_+ \to Y$ which has for all $0 < \lambda < \overline{\lambda}$ and $0 < \epsilon < \overline{\epsilon}$ the following properties:

$$\left| \frac{f(u+\theta h) - f(u)}{\theta} \right| \leq P(h,\epsilon,\lambda) \quad \forall\, u \in U(x,\epsilon), \theta \in (0,\lambda) \ ,$$

$$P \text{ is continuous in } h \quad \text{and} \quad \mathcal{T}\text{-}\lim_{h\to 0} P(h,\epsilon,\lambda) = 0 \quad \forall\, \lambda,\epsilon > 0.$$

If f is $\mathcal{U}$-Lipschitz-continuous in the sense of Definition 2.2, we obtain the existence of a generalized derivative in the sense of the following theorem.

THEOREM 2.3. *Assume that $\mathcal{U}(x)$ is a monotonic contraction system, that f is $\mathcal{U}$-Lipschitz-continuous, and that $(Y, \mathcal{T})$ possesses additionally the Daniell property either for sequences or for nets (i.e., every monotonically nonincreasing sequence (or net) that is order-bounded from below converges with respect to $\mathcal{T}$). Then the limit*

$$\mathcal{T}\text{-}D^s f(x,h) := \mathcal{T}\text{-}\lim_{\epsilon,\lambda\to 0} \sup_{\substack{u\in U(x,\epsilon)\\ 0<\theta\leq\lambda}} \frac{f(u+\theta h) - f(u)}{\theta} \tag{3}$$

exists and has the following properties:

(a) $\mathcal{T}\text{-}D^s f(x,\delta h) = \delta \mathcal{T}\text{-}D^s f(x,h) \quad \forall\, \delta \geq 0.$

(b) $\mathcal{T}\text{-}\lim_{h\to 0} \mathcal{T}\text{-}D^s f(x,h) = 0.$

(c) *If for a convex cone $V \subseteq X$ the* MCS *$\mathcal{U}(x)$ is V-stable then for all $h,k \in V$ we have $\mathcal{T}\text{-}D^s f(x,h+k) \leq \mathcal{T}\text{-}D^s f(x,h) + \mathcal{T}\text{-}D^s f(x,k)$. If $X_0 \subseteq X$ is a subspace and this inequality holds for every $h,k \in X_0$, then $\mathcal{T}\text{-}D^s f(x,\cdot)$ is sublinear on X_0. If $V = X$, then the mapping $\mathcal{T}\text{-}D^s f(x,\cdot) : X \to Y$ is continuous.*

(d) *If $M \subseteq Y$ is a set with the properties $\mathcal{T}\text{-}\operatorname{int} M \neq \emptyset$ and $\mathcal{T}\text{-}\operatorname{int} M - C_Y \subseteq \mathcal{T}\text{-}\operatorname{int} M$ and if $\mathcal{T}\text{-}D^s f(x,h) \in \mathcal{T}\text{-}\operatorname{int} M$ holds for a given $h \in X$, then there are $\epsilon, \lambda > 0$ with*

$$\frac{f(u+\theta h) - f(u)}{\theta} \in \mathcal{T}\text{-}\operatorname{int} M \quad \forall\, u \in U(x,\epsilon),\ 0 < \theta < \lambda. \tag{4}$$

Proof. Existence. If ϵ, λ are sufficiently small, the order-completeness of Y and the fact that f is $\mathcal{U}$-Lipschitz-continuous guarantee the existence of

$$\begin{aligned} S(\epsilon,\lambda) &:= \sup \{ \theta^{-1}(f(u+\theta h) - f(u)) \mid u \in U(x,\epsilon),\ 0 < \theta \leq \lambda \} \in Y, \\ S &:= \inf_{\epsilon,\lambda > 0} S(\epsilon,\lambda) \in Y. \end{aligned}$$

For all monotonically decreasing positive real sequences (or nets) $\epsilon \to 0, \lambda \to 0$, by definition, the elements $S(\epsilon, \lambda)$ form a monotonically nonincreasing sequence (or net) in Y that is bounded from below by S. Hence the τ-Daniell property of Y implies the existence of $\tau\text{-}Df(x,h) = \tau\text{-}\lim_{\epsilon,\lambda\to 0} S(\epsilon,\lambda)$. Obviously the monotonicity of $S(\epsilon,\lambda)$ implies that this limit is independent of the choice of the sequences ϵ, λ.

The proof of assertion (a) and of sublinearity in (c) is the same as for a real-valued f.

To prove (b), note that for ϵ, λ sufficiently small, we have $|Df(x,h)| \leq P(x,h,\epsilon,\lambda)$, hence for $h \to 0$ from $P(x,h,\epsilon,\lambda) \to 0$ and the normality of the ordering, $Df(x,h) \to 0$ follows. Now we will see that this continuity at 0, together with sublinearity in $V = X$, implies continuity everywhere, i.e., the last assertion in (c). Take a net $\{h_i\}_{i\in I} \subseteq X$ with $k := \lim_{i\in I} h_i$. If Df is sublinear on $V = X$ we obtain

$$Df(x,k) = Df(x,k-h_i+h_i) \leq Df(x,k-h_i) + Df(x,h_i), \tag{5}$$

$$Df(x,h_i) = Df(x,h_i-k+k) \leq Df(x,h_i-k) + Df(x,k), \tag{6}$$

and thus

$$-Df(x,h_i-k) \leq -Df(x,h_i) + Df(x,k) \leq Df(x,k-h_i).$$

Since from (b) $\lim_{i\in I} Df(x,k-h_i) = \lim_{i\in I} Df(x,h_i-k) = 0$ follows, and since C_Y is normal, we can conclude that $Df(x,k) = \lim_{i\in I} Df(x,h_i)$; hence $Df(x,\cdot)$ is continuous at $k \in X$.

To prove (d), note that $Df(x,h) \in \operatorname{int} S$ implies the existence of a neighborhood V of 0 in Y with $Df(x,h) + V \subseteq \operatorname{int} S$. Hence there are $\epsilon, \lambda > 0$ with $S(\epsilon,\lambda) \in Df(x,h) + V$ and together with

$$\frac{f(u+\theta h) - f(u)}{\theta} \in S(\epsilon,\lambda) - C_Y \subseteq Df(x,h) + V \ - \ C_Y \subseteq \operatorname{int} S,$$

we conclude (d). □

DEFINITION 2.4. We define the *directional derivative* of f, if it exists, by the limit

$$\tau\text{-}D^r f(x,h) := \tau\text{-}\lim_{\theta\to 0} \theta^{-1}\big(f(x+\theta h) - f(x)\big).$$

If this mapping is sublinear on X, i.e. for all $h, k \in X$ $\tau\text{-}D^r f(x,h+k) \leq \tau\text{-}D^r f(x,h) + \tau\text{-}D^r f(x,k)$, we will refer to this as CASE(D). We define the *generalized derivative* of f by

$$\tau\text{-}Df(x,\cdot) := \begin{cases} \tau\text{-}D^r f(x,\cdot) & \text{in CASE(D),} \\ \tau\text{-}D^s f(x,\cdot) & \text{else.} \end{cases}$$

If τ is the strong topology, then we will denote this mapping simply by $Df(x,\cdot)$. If, in particular, $\tau\text{-}D^r f(x,\cdot) : X \to Y$ is linear, then it is called the *Gâteaux derivative* of f.

Remark 2.2. Obviously, by definition and by Theorem 2.3, the generalized derivative is always sublinear, at least on the convex cone V where $\mathcal{U}(x)$ is V-stable; but stability is needed only if $D^r f$ is not sublinear. We could define $D^s f$ even with a nonmonotonic contraction system, but this would require an additional sup-operator (see (a) below) and is of unnecessary generality. If f is a convex operator, then by [6],

$D^r f$ and hence Df are sublinear. If f is Gâteaux-differentiable, then Df coincides with the Gâteaux derivative and thus the smooth cases are covered. Under certain circumstances $D^r f$ and $D^s f$ coincide. If $Y = I\!R^n$ and $f = (f_1, \cdots, f_n)$, then Df is the vector of Clarke's derivative of the functions f_i if one chooses a MCS as in Example 2.1 (c). Since each f_i is real-valued, one can deduce with [9, Chap. 2.3], that these derivatives coincide with the directional derivative in the convex case and with the gradient in the strict differentiable case. If X and Y are function spaces one can easily deduce similar results for the topology of *pointwise* convergence in Y, but to extend them to norm-topologies, one has to impose more assumptions (see [39]). The statement of Theorem 2.3 (d) can be strengthened in CASE(D) to hold for *every* set S with nonempty interior and leads to the definition of the Gâteaux variation as in [21] or [38]. However, the weaker form will be sufficient to play a crucial role in the derivation of necessary optimality conditions for multicriteria optimization problems (see next theorem). We now discuss the relationships of the introduced differentiability notion to similar concepts.

(a) Kusraev [23] calls $f : X \longrightarrow Y$ *locally Lipschitzian* if X, Y are topological vector lattices (i.e., ordered), order convergence in Y implies topological convergence, and if for every $x \in X$ there is a neighborhood V of x and a sublinear, monotone, and continuous operator $P : X \longrightarrow Y$ with

$$|f(x_1) - f(x_2)| \leq P(x_1 - x_2) \quad \forall\, x_1, x_2 \in V. \tag{7}$$

For those mappings he obtains existence of the mapping (which he calls Clarke's derivative of f)

$$f^\circ(x,h) = \sup_{\mathcal{F} \to x} \inf_{0<\epsilon, U \in \mathcal{F}} \sup_{u \in U, 0 < \theta \leq \epsilon} \frac{1}{\theta} \left(f(u + \theta h) - f(u) \right), \tag{8}$$

where $\mathcal{F} \longrightarrow x$ is a filter converging to x and the first sup is taken over all those filters. The difference between this definition and Definition 2.2 and Theorem 2.3 lies in the use of arbitrary filters instead of a MCS, which requires the additional sup-operator, and in the assumption that P is *monotone and sublinear*. This assumption is very strong (see the examples below) and can easily be dropped without losing the essential properties of the derivative, as Definition 2.2 shows.

(b) Reiland [30] calls f *order Lipschitz* if there is a neighborhood $U(x)$ of x, a neighborhood $W \subseteq X$ of 0, a real $\epsilon > 0$, and order bounds $y_1, y_2 \in Y$ with

$$y_1 \leq_{C_Y} \theta^{-1} \left(f(u + \theta w) - f(u) \right) \leq_{C_Y} y_2 \quad \forall\, 0 < \theta \leq \epsilon, u \in U(x), w \in W. \tag{9}$$

(c) Papageorgiou [26] calls f *order Lipschitz* if there is a neighborhood $U(x)$ of x, a neighborhood $W \subseteq X$ of 0, and an element $k \in C_Y$ with

$$|f(u + w) - f(u)| \leq k\|w\| \quad \forall\, u \in U(x), w \in W. \tag{10}$$

(d) Topological definition. f is Lipschitz-continuous if and only if there is a constant $\lambda > 0$ with

$$\|f(u + w) - f(u)\| \leq \lambda\|w\| \quad \forall\, u \in U(x), w \in W. \tag{11}$$

It is easy to see that (a), (b), and (c) imply the generalized Lipschitz property defined in Definition 2.2 as well as the usual topological Lipschitz continuity (d) if C_Y is normal. If $\operatorname{int} C_Y \neq \emptyset$ then Remark 2.2 (d) implies (b) and (c) (but not (a)!). Thibault's notion [40] is different and consists of a definition of a (sublinear) subderivative via the contingent cone of the epigraph of f. If f is "compactly Lipschitzian," one can approximate this subderivative by sequences of difference quotients. In [40], an application to optimization problems is not given.

By the following example we will see that Definition 2.2 is weaker than the other definitions, which also connect the topological and the order structure.

Example 2.3. (a) Let $I \subseteq \mathbb{R}$ be a bounded interval and consider

$$f : X := L^2(I) \to L^1(I) =: Y, \quad f(x) = x(t)^2, \quad x \in X.$$

Obviously $\theta^{-1}(|f(u+\theta h) - f(u)|)(t) = |2u(t)h(t) + \theta h(t)^2|, \quad t \in I$. The mapping f is not Lipschitz in the sense of Remark 2.2 (a), (b), or (c) because of

$$\sup \{ \, |2u(t)h(t)| \, \bigm| \, \|x-u\|_{L^2(I)} \le \epsilon \, \} = \infty \not\in Y.$$

There is even no sublinear continuous operator P with $|2uh + \theta h^2| \le P(h)$ for all $u : \ \|u-x\|_{L^2} \le \epsilon$. But by taking the monotonic contraction system $U(x,\epsilon) = \{ \, u \in X \, \bigm| \, |u(t) - x(t)| \le \epsilon \quad \mu \text{ a.e} \, \}$ we can conclude

$$|2uh + \theta h^2| \le 2(|x| + \epsilon)|h| + \lambda h^2 \ =: \ P(x,h,\lambda,\epsilon) \qquad \forall\, 0 < \theta < \lambda, u \in U(x,\epsilon)$$

and this mapping P (which is not sublinear) meets the assumptions of Definition 2.2. One could choose even the monotonic contraction system of Example 2.1 (d), since f is Gâteaux-differentiable.

(b) Let $I \subseteq \mathbb{R}$ be a bounded interval and consider the mapping

$$f : X := L^\infty(I) \to L^\infty(I) =: Y, \qquad f(x(t)) = \exp(|x(t)|).$$

Since for all $u \in X, \delta > 0$, the implication $\|u\|_{L^\infty} \ \le \ \delta \Rightarrow |u(t)| \ \le \ \delta, \mu$ almost everywhere holds, we conclude

$$\theta^{-1}|f(u+\theta h) - f(u)| \le |h| \exp(|x| + \epsilon + \lambda|h|) \qquad \forall\, \theta \in (0,\lambda), u \in U(x,\epsilon),$$

μ almost everywhere in I. This inequality is true for every monotonic contraction system $\mathcal{U}(x)$ of Example 2.1. The operator

$$P(x,h,\epsilon,\lambda) := |h| \exp(|x| + \epsilon + \lambda|h|)$$

is neither sublinear nor can we find a sublinear operator that majorizes P for all $h \in X$. Hence f is not Lipschitz in the sense of Remark 2.2 (a) but P does meet the assumptions of Definition 2.2 and f is therefore $\mathcal{U}$-Lipschitz-continuous. From $P(x,h,\epsilon,\lambda) \le |h| \exp(|x| + \epsilon + \lambda)$ for all $h : \|h\|_{L^\infty} \le 1$ we deduce that f is Lipschitz-continuous in the sense of Remark 2.2 (b), (c), and (d), which illustrates the equivalence of these definitions if the ordering cone has nonempty interior (as is true in $L^\infty(I)$). This example shows that Kusraev's definition is too strong.

For the generalized derivative we now obtain a set of linear approximations, i.e., a subdifferential, and investigate its properties. This will be done by applying the theory for convex operators (see, e.g., [42], [6]) to the sublinear generalized derivative of f (see, e.g., [30]).

THEOREM 2.5. *Let f be $\mathcal{U}$-Lipschitz-continuous and assume that the* MCS $\mathcal{U}(x)$ *is X-stable. $\mathcal{L}(X,Y)$ denotes the set of linear continuous operators from X to Y. Then the following assertions hold.*

(a) *The subdifferential of f at x, i.e.,*

$$\partial f(x) := \{\, L \in \mathcal{L}(X,Y) \mid L(h) \leq Df(x,h) \quad \forall\, h \in X \,\}$$

is nonempty and convex.

(b) *If X is a Fréchet space, then the set $\partial f(x)$ is equicontinuous.*

(c) *If the order intervals of Y are weakly compact, then the set $\partial f(x)$ is compact in $\mathcal{L}_s(X,Y_w)$ (i.e., the set of linear operators from X to Y that are continuous with respect to the weak topology of Y), if $\mathcal{L}_s(X,Y_w)$ is equipped with the topology of pointwise convergence.*

(d) *If C_Y is normal, then for all $h \in X$: $Df(x,h) = \max_{L \in \partial f(x)} L(h)$.*

For a composition of linear positive operators and Lipschitz-continuous operators, one can easily derive the following chain-rules.

THEOREM 2.6. (a) *Let (Z, C_Z) be an order-complete $\mathcal{T}$-Daniell topological vector lattice. Then for every linear and continuous operator $L : Y \to Z$ that is positive (i.e., $L(C_Y) \subseteq C_Z$) the following holds:*

$$D(L \circ f)(x,h) \leq L \circ Df(x,h) \qquad \forall\, h \in X,$$

$$\partial(y^* \circ f)(x) \subseteq \overline{y^* \circ \partial f(x)}^{w^*} \qquad \forall\, y^* \in C_{Y^*}.$$

(b) *If the order intervals of Y are weakly compact, we even have*

$$\partial(y^* \circ f)(x) \subseteq y^* \circ \partial f(x) \qquad \forall\, y^* \in C_{Y^*}.$$

Proof. See, e.g., [39] or [30]. □

As a special case of this theorem we note that for positive linear functionals (i.e., $y^* \in C_{Y^*}$) the following relation holds:

$$D(y^* \circ f)(x,h) \leq y^* \circ Df(x,h) \qquad \forall\, h \in X,\ y^* \in C_{Y^*}.$$

3. Necessary optimality conditions. In this section we use the generalized derivative to derive necessary optimality conditions for nonsmooth constrained multicriterial optimization problems. The main tool to approximate the feasible set will be the well-known contingent cone (also Bouligand's tangent cone) defined below. This cone can also be used to define the contingent derivative of set-valued maps (see Aubin [1], [2]; Robinson [34]; and Klose [22] where it is used to study stability of nonlinear programming problems). With this contingent derivative Luc [24] obtains even necessary conditions for some vector optimization problems, but this approach is different from ours.

DEFINITION 3.1. Let X be a tvs with a given topology $\mathcal{T}$, $S \subseteq X$ be a nonempty subset, and $x \in X$ be given. Then the *contingent cone* or *tangent cone of Bouligand* of S at x is defined as

$$T_{\mathcal{T}}(S,x) = \left\{ h \in X \;\middle|\; \begin{matrix} \exists \{h_n\}_{n\in\mathbb{N}} \subseteq X, & h = \mathcal{T}\text{-}\lim\limits_{n\to\infty} h_n, \\ \exists \{t_n\}_{n\in\mathbb{N}} \subset \mathbb{R}_+, & t_n \to 0, \end{matrix} \quad x + t_n h_n \in S \ \ \forall\, n \in \mathbb{N} \right\}.$$

If $\mathcal{T}$ is the strong topology of X we will simply write $T(S,x)$ and $T_w(S,x)$ if $\mathcal{T}$ is the weak topology.

DEFINITION 3.2. (a) $f(x)$ is a (*Pareto-*) *minimal* element of the set $f(S)$ if $(f(x) - C_Y \setminus \{0\}) \cap f(S) = \emptyset$.

(b) Let $\operatorname{int} C_Y \neq \emptyset$. Then $f(x)$ is a *weakly minimal* element of the set $f(S)$ if $(f(x) - \operatorname{int} C_Y) \cap f(S) = \emptyset$.

(c) $f(x)$ is a *properly minimal* element of the set $f(S)$, if $f(x)$ is minimal in the sense of (a) and if additionally $\overline{T(f(S) + C_Y, f(x))} \cap -C_Y = \{0\}$ holds.

(d) $f(x)$ is a *weakly properly minimal* element of the set $f(S)$, if $f(x)$ is minimal in the sense of (a), and if additionally $\overline{T_w(f(S) + C_Y, f(x))} \cap -C_Y = \{0\}$ holds.

The set of all minimal (weakly minimal, properly minimal, weakly properly minimal) elements of $f(S)$ is denoted by MIN($f(S)$) (WMIN($f(S)$), PMIN($f(S)$), WPMIN($f(S)$), respectively) and obviously

$$\text{WPMIN}(f(S)) \subseteq \text{PMIN}(f(S)) \subseteq \text{MIN}(f(S)) \subseteq \text{WMIN}(f(S)) \tag{12}$$

and the first two sets coincide if $Y = \mathbb{R}^n$.

For weakly minimal and weakly properly minimal elements, one can derive the following necessary optimality conditions of the first order.

THEOREM 3.3. *Assume that f satisfies the hypotheses of Theorem* 2.3, *i.e., let f be $\mathcal{U}$-Lipschitz-continuous and assume that the* MCS $\mathcal{U}(x)$ *is X-stable. Let $f(x)$ be a weakly minimal element of the set $f(S)$. Then the following statements hold:*

(a) *$Df(x,h) \notin -\operatorname{int} C_Y$ for all $h \in T(S,x)$.*

(b) *For every convex cone $K \subseteq T(S,x)$ there is some $y^* \in C_{Y^*}$, $y^* \neq 0_{Y^*}$ with*

$$y^* \circ Df(x,h) \geq 0 \quad \forall\, h \in K, \tag{13}$$

$$0 \in \partial\big(y^* \circ f\big)(x) - K^*. \tag{14}$$

(c) *If the order intervals in Y are weakly compact, then*

$$0 \in y^* \circ \partial f(x) - K^*. \tag{15}$$

Proof. (a) Suppose that there is some $h \in T(S,x)$ with the property $Df(x,h) \in -\operatorname{int} C_Y$. Then there are sequences $\{h_n\}_{n\in\mathbb{N}} \subseteq X$ and $\{\theta_n\}_{n\in\mathbb{N}} \subseteq \mathbb{R}_+$ and a number $n_0 \in \mathbb{N}$ for which

$$0 = \lim_{n\to\infty} \theta_n, \quad h = \lim_{n\to\infty} h_n \quad \text{and} \quad x + \theta_n h_n \in S \quad \forall\, n \in \mathbb{N}.$$

Continuity of the mapping $Df(x,\cdot)$ then yields

$$Df(x,h_n) \in -\operatorname{int} C_Y \quad \forall\, n > n_0.$$

Since the set $M := -\operatorname{int} C_Y$ fulfills the assumptions of Theorem 2.3 (d) we obtain

$$f(x + \theta_n h_n) - f(x) \subseteq -\theta_n \operatorname{int} C_Y \subseteq -\operatorname{int} C_Y$$

for all numbers $n \in \mathbb{N}, n > n_0$ for which θ_n is sufficiently small. But the last inclusion is a contradiction to $f(x)$ being weakly minimal.

(b) From (a) we conclude that $Df(x,K) + C_Y \cap -\operatorname{int} C_Y = \emptyset$, sublinearity of $Df(x,\cdot)$ implies convexity of the set $Df(x,K) + C_Y$ and hence we can apply a separation argument to obtain existence of a positive linear functional $y^* \in C_{Y^*}$, $y^* \neq 0_{Y^*}$, satisfying (13). To prove (14) set $p(h) := y^* \circ Df(x,h)$. Observe that $p : X \to \mathbb{R}$ is sublinear and continuous. Hence the set

$$epi(p) := \big\{\, (h,r) \in X \times \mathbb{R} \;\big|\; p(h) \leq r \,\big\}$$

is convex and has nonempty interior in $X \times \mathbb{R}$. Because (15) implies $p(k) \geq 0$ for all $k \in K$ and hence

$$\operatorname{int} epi(p) \;\cap\; \{\, (k,\alpha) \in X \times \mathbb{R} \;\mid\; k \in K, \alpha \leq 0 \,\} = \emptyset. \tag{16}$$

By application of a separation theorem to this intersection we obtain the existence of a linear continuous functional $(x_1^*, \beta) \in (X \times \mathbb{R})^* = X^* \times \mathbb{R}$, $(x_1^*, \beta) \neq 0$, satisfying

$$x_1^*(h) + \beta r \leq 0 \leq x_1^*(k) + \beta\alpha \quad \forall\, h \in X, k \in K, r \geq p(h), \alpha \leq 0. \tag{17}$$

These inequalities imply $\beta \leq 0$. If $\beta = 0$, then (17) yields $x_1^* = 0$ in contradiction to $(x_1^*, \beta) \neq 0$. Hence we have $\beta < 0$ and thus $-\beta^{-1} > 0$. Setting $r = p(h)$ in the first inequality of (17) we obtain $x_1^*(h) \leq -\beta p(h)$ and, equivalently, by $-\beta^{-1} > 0$ $x^*(h) := -\beta^{-1} x_1^*(h) \leq p(h) = y^* \circ Df(x,h)$, hence $x^* \in \partial\big(y^* \circ f\big)(x)$. From the second inequality in (17) we deduce $x_1^*(k) \geq 0$ for all $k \in K$, hence $x_1^* \in K^*$ and $x^* \in K^*$. Both properties of x^* together imply $0 \in \partial\big(y^* \circ f\big)(x) - K^*$.

(c) This statement follows from Theorem 2.5. □

There always exists a convex cone K meeting the assumption $K \subseteq T(S,x)$ of the preceding theorem, for example, $K = \{0\}$ or Clarke's tangent cone (which is always convex; see, e.g., [9]). But it can be shown by the following simple example that it is generally possible to choose the cone K bigger than Clarke's, thus obtaining sharper necessary optimality conditions (because of $K_1 \subseteq K_2 \Rightarrow K_2^* \subseteq K_1^*$). Let $S := \{\, (t,s) \in \mathbb{R}^2 \mid s \leq |t| \,\}$. Obviously S is closed but not convex and we have $T(S,0) = S$. Clarke's tangent cone at S at the point $x = 0$ is $T_{Cl}(S,0) = \{\, (t,s) \in \mathbb{R}^2 \mid s \leq -|t| \,\}$. This cone is convex but also every cone $K_\alpha := \{\, (t,s) \in \mathbb{R}^2 \mid s \leq \alpha t \,\}$ with $|\alpha| \leq 1$ is convex, but much larger than the set $T_{Cl}(S,0)$. It satisfies

$$T_{Cl}(S,0) \;\subset\; K_\alpha \;\subset\; T(S,0),$$

where all inclusions are strict. Hence for the dual cones appearing in the necessary conditions we have the strict inverse inclusions, i.e., $K_\alpha^* \subset N(S,0) = T_{Cl}^*(S,0)$ where N is the cone of normals of $T_{Cl}^*(S,0)$ (see [9]). In fact $K_\alpha^* = \{\, \beta(1,-\alpha) \mid \beta \geq 0 \,\}$ is a very small cone consisting of only one direction whereas the cone $N(S,0) = \{\, (t,s) \in \mathbb{R}^2 \mid s \geq |t| \,\}$ is quite big. Hence the condition $0 \in \partial f(x) - K^*$ does give us much more information about the minimum than $0 \in \partial f(x) - N(S,0)$.

Note also that Theorem 3.3 (a) does not require any convexity assumption on K. It states that $0 \in \mathrm{WMIN}(Df(x,T(S,x)))$, thus being a generalization of the classical necessary optimality condition $\langle\, \nabla f(x) | h \,\rangle \geq 0$ for all $h \in T(S,x)$ for a real-valued, differentiable mapping f. The same holds for Theorem 3.4(b), which gives necessary optimality conditions for weakly properly minimal elements.

THEOREM 3.4. *Assume that f satisfies the hypotheses of Theorem 2.3, that is, let f be $\mathcal{U}$-Lipschitz-continuous and assume that the* MCS *$\mathcal{U}(x)$ is X-stable. If the order intervals of Y are weakly compact it follows that*

(a) $Df(x,T(S,x)) \subseteq T_w(f(S) + C_Y, f(x))$.

(b) *If $f(x)$ is now a weakly properly minimal element of the set $f(S)$ it follows that $\overline{Df(x,T(S,x)+C_Y)}^{w} \cap -C_Y = O_Y$.*

(c) *Assume that C_Y has a weakly compact base. Then for every convex cone $K \subseteq T(S,x)$ there is a strictly positive linear functional $y^* \in C_{Y^*}^i$ with*

$$y^* \circ Df(x,h) \geq 0 \quad \forall\, h \in K, \tag{18}$$

$$0 \in \partial\big(y^* \circ f\big)(x) - K^*, \tag{19}$$

$$0 \in y^* \circ \partial f(x) - K^*. \tag{20}$$

Proof. (a) Let $h \in T(S,x)$ be given. Then there are sequences $\{h_n\}_{n\in\mathbb{N}} \subseteq X$ and $\{\theta_n\}_{n\in\mathbb{N}} \subseteq \mathbb{R}_+$ with

$$0 = \lim_{n\to\infty} \theta_n, \quad h = \lim_{n\to\infty} h_n \quad \text{and} \quad x + \theta_n h_n \in S \quad \forall\, n \in \mathbb{N}.$$

Set $d_n := \theta_n^{-1}(f(x+\theta_n h_n) - f(x))$ and $e_n := \theta_n^{-1}(f(x+\theta_n h) - f(x))$. We show that there are subsequences of $\{d_n\}_{n\in\mathbb{N}}$ and $\{e_n\}_{n\in\mathbb{N}}$ converging weakly (in Y) to the same limit. Because f is $\mathcal{U}$-Lipschitz-continuous there is an operator P and a number $n_0 \in \mathbb{N}$ with $\lim_{h\to 0} P(x,h,\epsilon,\lambda) = 0$ if $0 < \epsilon, \lambda$ is sufficiently small, for which

$$|e_n| \leq P(x,h,\epsilon,\lambda) \quad \forall\, n > n_0$$

holds, i.e., the sequence $\{e_n\}_{n\in\mathbb{N}}$, $n > n_0$, is contained in an appropriate order interval. Now the weak compactness of order intervals implies the weak convergence of a subsequence $\{e_{n_k}\}_{k\in\mathbb{N}}$, against a limit $d \in Y$. Again we deduce from f being $\mathcal{U}$-Lipschitz-continuous the inequality

$$0 \leq |d_{n_k} - e_{n_k}| \leq P(h_{n_k} - h).$$

Together with

$$-y^*(|d_{n_k} - e_{n_k}|) \leq y^*(d_{n_k} - e_{n_k}) \leq y^*(|d_{n_k} - e_{n_k}|) \quad \forall\, y^* \in C_{Y^*}$$

and $\lim_k P(h_{n_k} - h) = 0$ we can conclude $\lim_{k\to\infty} y^*(d_{n_k} - e_{n_k}) = 0$ for all $y^* \in C_{Y^*}$. Since C_Y is normal, we have $Y^* = C_{Y^*} - C_{Y^*}$ and thus $\lim_k y^*(d_{n_k} - e_{n_k}) = 0$ for all $y^* \in Y^*$. From the equality $y^*(d_{n_k}) = y^*(e_{n_k}) + y^*(d_{n_k} - e_{n_k})$, we finally deduce the weak convergence of the subsequence $\{d_{n_k}\}_{k\in\mathbb{N}}$ with the limit d. We now claim $d \in T_w(f(S), f(x))$. Since $h \in T(S,x)$ we have $f(x+\theta_n h_n) \in f(S)$ for all $n \in \mathbb{N}$. Thus the sequence

$$y_k := d_{n_k} = \theta_{n_k}^{-1}(f(x + \theta_{n_k} h_{n_k}) - f(x))$$

has the weak limit d and hence $d \in T_w(f(S), f(x))$. The definitions of $Df(x,h)$ and e_n now imply that $Df(x,h) \in d + C_Y$ holds. With a standard argument we can now conclude that

$$d + C_Y \subseteq T_w(f(S), f(x)) + C_Y \subseteq T_w(f(S) + C_Y, f(x))$$

and hence we obtain the statement (a). Since convexity of C_Y implies $C_Y + C_Y \subseteq C_Y$ we can even deduce for every $h \in T(S,x)$:

$$Df(x,h) + C_Y \subseteq d + C_Y + C_Y \subseteq d + C_Y \subseteq T_w(f(S) + C_Y, f(x)),$$

and thus

$$Df(x, T(S,x)) + C_Y \subseteq T_w(f(S) + C_Y, f(x)), \tag{21}$$

which we will need for the next step.

(b) The proper weak minimality of $f(x)$ together with (a) and the fact shown in (21) which implies that

$$\overline{Df(x,T(S,x))+C_Y}^{\,w} \cap -C_Y \subseteq \overline{T_w(f(S)+C_Y,f(x))}^{\,w} \cap -C_Y = \{0\}$$

immediately yields (b).

(c) From (b) we deduce $\overline{Df(x,K)+C_Y} \cap -C_Y = \{0\}$. Since both sets are convex and closed cones and C_Y has a weakly compact base we can apply a separation argument (see, e.g., [4]) to show the existence of $y^* \in C^i_{Y^*}$ with $y^* \circ Df(x,h) \geq 0$ for all $h \in K$. The subdifferential condition (19) can be derived as in the preceding theorem and (20) follows from (19) together with Theorem 2.6, since the order intervals were assumed to be weakly compact. □

We remark that for a directional derivative Df we even have

$$Df(x,T(S,x)) \subseteq T(f(S),f(x)).$$

For the case in which the constraint set S is given by explicit operator constraints we will now derive an abstract multiplier rule by "computing" K in terms of the constraint mappings. This theorem will be an extension of a corresponding one of [30] to vector-valued objective mappings, i.e., multicriterial optimization problems. [9] and [25] also derive necessary conditions for those problems, but only for objective functions and constraints that have a finite-dimensional range. In the infinite-dimensional case Borwein [4] gives a general Kuhn–Tucker theorem that characterizes properly minimal elements under the assumption that objective and constraints are Fréchet-differentiable. Hence the following theorem is an extension of Borwein's to Lipschitz mappings.

Let Z_1,Z_2 be partially ordered and $g_1\colon X \to Z_1$, $g_2\colon X \to Z_2$ be given mappings, and suppose that $S_0 \subseteq X$, $D_1 \subseteq Z_1$, and $D_2 \subseteq Z_2$ are nonempty sets. For $i = 1,2$ we define

$$S_i := \{\, x \in X \mid g_i(x) \in D_i \,\}, \tag{22}$$

$$S := S_0 \cap S_1 \cap S_2, \tag{23}$$

$$P_i := \overline{\mathrm{co}\,T}(D_i, g_i(x)), \tag{24}$$

$$P_i^* := \{\, z^* \in Z_i^* \mid z^*(p) \geq 0,\ \forall p \in P_i \,\}, \tag{25}$$

$$J_i := \{\, h \in X \mid L(h) \in P_i,\ \forall L \in \partial g_i(x) \,\}, \tag{26}$$

$$H_i := \{\, x^* \in X^* \mid \exists p^* \in P_i^*, L \in \partial g_i(x) \ :\ x^* = p^* \circ L \,\}, \tag{27}$$

$$J_0 := \overline{\mathrm{co}\,T}(S_0, x). \tag{28}$$

The sets P_i^* are multiplier sets and the sets H_i are sets of compositions of multipliers and subdifferentials. If in particular inequality and equality constraints are given, take $D_1 := -C$ and $D_2 = 0$, where $C \subseteq Z_1$ is a convex cone representing the inequality constraints. The following theorem then gives necessary optimality conditions, which can be considered as abstract Karush–Kuhn–Tucker conditions. Note that the multiplier y^* is always nonzero.

THEOREM 3.5. *We assume that at a point $x \in X$ with $f(x) \in$* WMIN$(f(S))$ *or $f(x) \in$* WPMIN$(f(S))$ *the mappings f, g_1, g_2 satisfy the hypotheses of Theorem 2.3, i.e., let f, g_1, g_2 be $\mathcal{U}$-Lipschitz-continuous and assume that the* MCS $\mathcal{U}(x)$ *is*

X-stable. The subdifferentials will be denoted by $\partial f(x)$, $\partial g_1(x)$, $\partial g_2(x)$. Let $K \subseteq X$ be a convex cone satisfying

$$K \subseteq T(S,x) \quad \textit{and} \quad K^* = (J_0 \cap J_1 \cap J_2)^*. \tag{29}$$

Then there is some $y^* \in C_{Y^*}$, $y^* \neq 0$, if $f(x) \in \mathrm{WMIN}(f(S))$ (or $y^* \in C^i_{Y^*}$, if $f(x) \in \mathrm{WPMIN}(f(S))$), satisfying the relation

$$0_{X^*} \in \partial\big(y^* \circ f\big)(x) - \overline{\operatorname{co} H_1 + \operatorname{co} H_2 + J_0^*}^{w^*}. \tag{30}$$

If additionally the relation $K^* = J_0^* + J_1^* + J_2^*$ holds, then

$$0_{X^*} \in \partial\big(y^* \circ f\big)(x) - \Big(\overline{\operatorname{co} H_1}^{w^*} + \overline{\operatorname{co} H_2}^{w^*} + J_0^*\Big). \tag{31}$$

If, moreover, the sets H_1 and H_2 are w^*-closed, then

$$0_{X^*} \in \partial\big(y^* \circ f\big)(x) - \Big(\operatorname{co} H_1 + \operatorname{co} H_2 + J_0^*\Big). \tag{32}$$

Proof. If $f(x)$ is weakly minimal it follows from Theorem 3.3 (from Theorem 3.4 if $f(x)$ is weakly properly minimal) and from (29) that

$$0_{X^*} \in \partial\big(y^* \circ f\big)(x) - K^* = \partial\big(y^* \circ f\big)(x) - (J_0 \cap J_1 \cap J_2)^*.$$

We will show $K^* = (J_0 \cap J_1 \cap J_2)^* \subseteq B := \overline{\operatorname{co} H_1 + \operatorname{co} H_2 + J_0^*}^{w^*}$. To get a contradiction, assume that there is some $x^* \in K^*$, $x^* \notin B$. By applying a separation theorem to the w^*-compact set $\{x^*\}$ and the w^*-closed set B we conclude the existence of $h \in X$ and $\alpha \in I\!R$ satisfying the inequality

$$x^*(h) < \alpha \leq b^*(h) \qquad \forall\, b^* \in B.$$

Because of $0 \in B$ and the definition of B we have $\alpha = 0$ and thus

$$b^*(h) \geq 0 \qquad \forall\, b^* = p^* \circ P_i^*, \quad L \in \partial g_i(x),\ i = 1,2 \quad \text{and} \quad b^* \in J_0^*.$$

From $p^* \circ L(h) \geq 0$ for all $p^* \in P_i^*$ it follows that $L(h) \in P_i^{**} = P_i$, for all $L \in \partial g_i(x)$, since P_i is a closed convex cone. This implies $h \in J_1 \cap J_2$ and with $J_0^{**} = J_0$, even $h \in J_0$. Because of $x^* \in K^*$, then $x^*(h) \geq 0$ in contradiction to the inequality $x^*(h) < \alpha = 0$.

In order to prove (31) we show $J_i^* = \overline{\operatorname{co} H_i}^{w^*}$. Suppose that $x^* \in J_i^*$ with $x^* \notin \overline{\operatorname{co} H_i}^{w^*}$. Using a separation argument again, we obtain $h \in X$ with

$$x^*(h) < 0 \leq v^*(h) \qquad \forall\, v^* \in \operatorname{co} H_i.$$

This implies $L(h) \in P_i^{**} = P_i$ for all $L \in \partial g_i(x)$ since P is a closed and convex cone and hence $h \in J_i$ for $i = 1,2$. But this implies $x^*(h) \geq 0$ in contradiction to $x^*(h) < 0$ from the inequality above. Now let $x^* \in \operatorname{co} H_i$ and $x^* \notin J_i^*$. Linear separation then yields the existence of some $h \in X$ with the property

$$x^*(h) < 0 \leq v^*(h) \qquad \forall\, v^* \in J_i^*.$$

Because of $J_i^{**} = J_i$ this implies $h \in J_i$, which means $L(h) \geq 0$ for all $L \in \partial g_i(x)$ and hence $z^* \circ L(h) \geq 0$ for all $z^* \circ L \in H_i$. Since $x^* \in \operatorname{co} H_i$ this inequality is a contradiction to the separation inequality. The last statement of the theorem follows from $\operatorname{co} H_i$ being w^*-closed. □

We now draw conclusions from this theorem. Regarding the definition of the sets H_1, H_2 the statement (32) means that for $j = 1, \cdots, n$ and $k = 1, \cdots, n$ there are linear operators

$$L^0 \in \partial f(x), \quad L_j^0 \in \partial g_1(x), \quad L_k^0 \in \partial g_2(x),$$

real numbers

$$\alpha_j \geq 0, \quad \beta_k \geq 0 \quad \text{with} \quad \sum_j \alpha_j = \sum_k \beta_k = 1,$$

and linear functionals (i.e., multipliers)

$$y^* \in C_{Y^*}, \quad y^* \neq 0_{Y^*} \quad (\text{respectively,} \quad y^* \in C_{Y^*}^i), \quad z_{1,j}^* \in P_1^*, \quad z_{2,k}^* \in P_2^*,$$

with the property

$$x^* := y^* \circ L^0 - \sum_j \alpha_j z_{1,j}^* \circ L_j^1 - \sum_k \beta_k z_{2,k}^* \circ L_k^2 \in J_0^*, \tag{33}$$

and hence $x^*(h) \geq 0$ for all $h \in J_0$. If the subdifferentials of g_1, g_2 consist of single elements (e.g., if g_1, g_2 are Gâteaux-differentiable) then the sets H_1 and H_2 are convex, and one has the classical Karush–Kuhn–Tucker condition

$$y^* \circ L^0(h) - z_1^* \circ Dg_1(x,h) - z_2^* \circ Dg_2(x,h) \geq 0 \quad \forall\, h \in J_0. \tag{34}$$

If we consider the particular case that $D_1 := C$ is a closed convex cone and $D_2 := 0$ then the set S_1 represents inequality constraints and S_2 represents equality constraints and we obtain the relations

$$P_1 = \overline{K}(-C, g_1(x)) \quad \text{and} \quad P_2 = 0. \tag{35}$$

This implies the *complementary slackness* condition $z_1^*(g_1(x)) = 0$ and the multiplier sets are $P_1^* = -C^*$ and $P_2^* = Z_2^*$. If $C = C_{Z_1}$, then we obtain (34) even without assuming that the mapping g_1 is Gâteaux-differentiable, that is, even in a nonsmooth case. For the generalized derivative Dg_1,

$$L_j^1(h) \leq_C Dg_1(x,h) \quad \forall\, h \in X, \ L_j^1 \in \partial g_1(x)$$

holds. This inequality obviously extends to convex combinations $L^1 = \sum_j \alpha_j L_j^1$, which, together with $z_1^* \in -C^*$, implies the inequality

$$-z_1^* \circ L^1(h) \leq -z_1^* \circ Dg_1(x,h) \quad \forall\, h \in X.$$

The same argument applies to $y^* \in C_{Y^*}$ and hence (33) implies

$$y^* \circ Df(x,h) - z_1^* \circ Dg_1(x,h) - z_2^* \circ Dg_2(x,h) \geq 0 \quad \forall\, h \in J_0, \tag{36}$$

if g_2 is Gâteaux-differentiable and f and g_1 are generalized differentiable. If g_2 is only generalized differentiable one has the full convex combination from (33), which seems to be typical in the nonsmooth case for control problems as well; see, e.g., [9] (Hamiltonian multipliers). If the spaces X, Y, Z_1, Z_2 are finite-dimensional (even if X is a Banach space) Rockafellar [36], Clarke [9], and Minami [25] obtain, with different methods, somewhat sharper necessary conditions than those of Theorem 3.5. The optimality conditions (30) and (31) are asymptotic conditions (see, e.g., [4]), which are generally difficult to apply. Hence we will now investigate the assumptions for (32) as well as (29), which can be understood as abstract constraint qualifications (CQ) of a type introduced by Guignard in [12]. It will turn out that they are implied by the classical Slater conditions but that they are more general than those. In fact, as has been demonstrated by Bazaraa and Shetty [3] in finite-dimensional spaces, Guignard's CQ is the weakest among several known CQs, moreover, Guignard's CQ does not assume the existence of interior points of C and S_0 required by Slater's CQ (see also Penot [28]). We will use the following lemma.

LEMMA 3.6. (a) *For every convex cone I with $\operatorname{int} I \neq \emptyset$ the dual cone I^* is w^*-locally compact* (*or, equivalently, has a w^*-compact base*).

(b) *If I^* and J^* are convex closed cones, I^* is w^*-locally compact and if $I^* \cap - J^* = \{0\}$ then the set $I^* + J^*$ is w^*-closed.*

Proof. The statement (a) is due to Ky Fan (see, e.g., [44]). The statement (b) is due to Dieudonnée (see, e.g., [20] or [14, Lemma 15(d)]). □

THEOREM 3.7. *Assume that $\operatorname{int} S_0 \neq \emptyset$ and $\operatorname{int} C \neq \emptyset$. Define $I_0 := \operatorname{int} J_0$ and $I_1 := \{\, h \in X \mid Dg_1(x,h) \in \operatorname{int} P_1 \,\}$ and $K := I_0 \cap I_1 \cap J_2$. If the cone K satisfies the Slater condition $K \neq \emptyset$ and if additionally the tangential inclusion*

$$J_2 \subseteq T(S_2, x) \tag{37}$$

holds, then K satisfies Guignard's CQ, *i.e., K is a convex cone with the properties*

$$K \subseteq T(S,x), \quad K^* = (J_0 \cap J_1 \cap J_2)^*, \quad K^* = J_0^* + J_1^* + J_2^*. \tag{38}$$

Proof. It is standard (convexity arguments) to prove the following statements:

$$\overline{I_0} = J_0, \qquad \overline{I_1} = J_1, \tag{39}$$

$$K^* = (J_0 \cap J_1 \cap J_2)^*, \qquad \overline{K} = J_0 \cap J_1 \cap J_2, \tag{40}$$

$$I_0 \cap I_1 \cap J_2 \subseteq T(S,x). \tag{41}$$

We will now deduce the equality $K^* = J_0^* + J_1^* + J_2^*$. Since

$$J_0^* + J_1^* + J_2^* \subseteq K = (J_0 \cap J_1 \cap J_2)^* \subseteq \overline{J_0^* + J_1^* + J_2^*}^{w^*}, \tag{42}$$

we only have to show that the set $J_0^* + J_1^* + J_2^*$ is w^*-closed, which will be done by Lemma 3.6. We now observe that $K \neq \emptyset$ implies $I_k^* \cap -J_2^* = \{0\}$ $(k = 0,1)$ and that I_0^*, I_1^* are w^*-locally compact. Hence repeated application of (a) and (b) of Lemma 3.6 gives w^*-closedness of the set $(J_0 \cap J_1 \cap J_2)^*$. □

We remark that the Slater CQ is sometimes stated in the equivalent form

$$\exists \ h \in \operatorname{int} S_0 : Dg_1(x,h) \in \operatorname{int} K(-C, g_1(x)), \qquad Dg_2(x,h) = 0.$$

LEMMA 3.8. *For $i = 1, 2$:*

(a) $H_i = \bigcup \{\, L^*(P_i^*) \mid L^* \text{ adjoint of } L \in \partial g_i(x) \,\}$.

(b) *If $\partial g_i(x) = \{L\}$, i.e., if the subdifferential consists of a single element, then the set H_i is always convex.*

Proof. The proof is immediate. □

We now discuss various sufficient conditions for H_1, H_2 being w^*-closed.

THEOREM 3.9. *Let X, Z_1, Z_2 be Banach spaces.*

(a) *If $H_2 = \bigcup \{\, L^*(P_2^*) \mid L^*$ adjoint of $L \in \partial g_2(x) \,\} = X^*$, then H_2 is convex and w^*-closed. If one $L \in \partial g_2(x)$ is a closed operator (i.e., $L(X)$ is closed) and injective, then $H_2 = X^*$, and this set is convex and w^*-closed.*

(b) *Now let $\partial g_2(x) = \{L\}$. Then H_2 is always convex and the following holds.*
 (i) *If Z_2 is finite-dimensional, then the set H_2 is always w^*-closed and locally w^*-compact.*
 (ii) *If L is a closed operator then H_2 is w^*-closed and norm-closed. Moreover, for $N := \{\, h \in X \mid L(h) = 0 \,\}$ we have $H_2 = N^* = \{\, x^* \in X^* \mid x^*(x) = 0 \;\forall\, x \in N \,\}$. Hence, if L is injective, then $H_2 = X^*$, and otherwise H_2 is a proper w^*-closed subspace of X^*.*

(c) *If $\partial g_1(x) = \{L\}$ and $\operatorname{int} C \neq \emptyset$, then the set $H_1 = L^*(-C^*)$ is a convex cone with w^*-compact base.*

Proof. (a) Because L is injective it follows that $\overline{L^*(Z_2^*)}^{w^*} = X^*$ (see, e.g., [37, Thm. 4.12 and Corollaries]) and closedness of L is equivalent to closedness of L^* in the norm-topology and in the w^*-topology (see, e.g., [37, Thm. 4.14]), hence $L^*(Z_2^*) = X^* = H_2$ and this last set is obviously convex and w^*-closed.

(b) (i) The set $H_2 = L^*(Z_2^*)$ is a finite-dimensional subspace and hence w^*-closed and locally w^*-compact.
 (ii) This follows the same argument as in (a). The characterization of H_2 follows from considering the quotient space $X_N := X/N$ on which L is injective and using the fact that $X_N^* = N^*$; hence by closedness of L one can deduce $L^*(Z_2^*) = X_N^* = N^*$.

(c) See [44]. □

The preceding theorem can be regarded as an extension of the surjectivity assumption for the Fréchet derivative of the equality constraints, which implies, by the famous Ljusternik's theorem, the tangential inclusion $J_2 \subseteq T(S_2, x)$. For an extension of this to Gâteaux-differentiable mappings, see [29] or [21].

Apart from the Slater and Guignard CQ, there are many other CQs in the literature. Those that apply to infinite-dimensional problems will be dicussed now. The original CQ of Guignard reads as follows. For $g := (g_1, g_2)$, $Z := Z_1 \times Z_2$, $C := C_{Z_1} \times \{0\}$, $J := J_0 \cap J_1 \cap J_2$, $H := \{\, x^* \in X^* \mid \exists z^* \in C^*, x^*(h) = z^* \circ Dg(x,h) \,\}$ there is a closed, convex cone $K_0 \subseteq X$ with the properties that $K_0 \cap J \subseteq T(S, x)$, and that the sets $K_0^* + J^*$ and H are w^*-closed. This CQ obviously does not assume the existence of interior points of the sets C and S_0, hence the equality and the inequality constraints can be treated together. The same holds for the CQ of Zowe and Kurcyusz [45] and Robinson [31], which can be formulated as follows. Assume that the mappings f and $g := (g_1, g_2)$ are Fréchet-differentiable and $C := C_{Z_1} \times \{0\}$, $Z := Z_1 \times Z_2$.

$$Z = Dg(x, K(S_0, x)) - K(-C, g(x)). \tag{43}$$

Zowe and Kurcyusz [45] show that this CQ implies

$$L(S, x) := \{\, h \in X \mid h \in K(S_0, x),\; Dg(x,h) \in -C \,\} \subseteq T(S, x),$$

which means that the CQ (29) of Theorem 3.5 holds because of $J_0 \cap J_1 \cap J_2 = L(S, x)$,

and

$$K := J_0 \cap J_1 \cap J_2 := L(S,x) \subseteq T(S,x).$$

If one assumes existence of interior points of the sets C_{Z_1} and S_0, then the CQ (43) is equivalent to the following CQ (see [41]):

$$\text{(44)} \quad \exists\, s \in \operatorname{int} S_0 : \; Dg(x, s-x) \in K(-C, g(x)), \qquad Dg(x,X) - K(-C,g(x)) = Z.$$

[41] makes use of this CQ to apply a multiplier rule of [45] to optimal control problems.

In [21] a necessary condition is proved (for Fréchet-differentiable constraints) under the following constraint qualification:

$$\text{(45)} \qquad \exists\, h \in I_0 : g_1(x) + Dg_1(x,h) \in -\operatorname{int} C_{Z_1}, \qquad Dg_2(x,h) = 0,$$

(46) $0 \in \operatorname{int} Dg_2(x, I_0)$, f, g_1, g_2 are Fréchet-differentiable and Dg_2 is surjective.

We see that by Theorems 3.7 and 3.9 these assumptions imply (29).

Now we demonstrate by an example that there are situations where the modified Guignard CQ (or the Robinson–Zowe–Kurcyusz CQ) are advantageous because the Slater CQ cannot be applied. This example may appear simple but in fact covers the essential structure of more complicated examples as optimal control problems of nonlinear partial differential equations or variational inequalities. The application of the Karush–Kuhn–Tucker conditions derived above to problems of this type will be the subject of a future paper.

Example 3.1 (pointwise constraints). Let $X = L^p(I)$ be the space of real functions on the real interval $I = [a,b]$ whose pth power is integrable with respect to the Lebesgue measure μ. As is well known, the natural ordering cone

$$\text{(47)} \qquad C = \{\, x \in X \;\mid\; x(t) \geq 0, \ \mu \text{ a.e. on } I \,\}$$

has an empty interior for $1 \leq p < \infty$. Hence the Slater CQ cannot be applied to the problem to minimize a given functional or mapping with respect to the constraint set $S := \{\, x \in L^2(I) \mid g(x) := -x(t) \leq 0 \ \mu \text{ a.e.} \,\}$. But because of the linearity of g we can easily see that $T(S,x)$ is a convex closed cone with $T(S,x) = J := \{\, h \in X \mid Dg(x,h) \in -C \,\}$ and thus $J^* = T^*$. Because the mapping g is Fréchet-differentiable, we can verify (43) and also (29).

We now consider the constraint set

$$\text{(48)} \qquad S := \{\, x \in X \;\mid\; g(x) := |x(t)| - \epsilon \leq 0 \ \mu \text{ a.e} \,\}$$

defined by the nonsmooth constraint $g : X := L^2(I) \to Z := L^2(I)$ where $\epsilon > 0$ is a given real number. At first we determine $Dg(x,h)$. Since g is convex it is sufficient to compute the directional derivative. We find

$$\text{(49)} \qquad Dg(x(t),h(t)) = \begin{cases} h(t)\operatorname{sgn}(x(t)), & t \in I : x(t) \neq 0, \\ |h(t)|, & t \in I : x(t) = 0. \end{cases}$$

This derivative is sublinear in h but not linear if $x = 0$ on a subset of I of positive measure. We will now show

$$\text{(50)} \qquad J := \{\, h \in X \;\mid\; L(h) \in K(-C, g(x)) \quad \forall\, L \in \partial g(x) \,\} = T(S,x).$$

We will show that this identity even holds for every $x \in S$. Because C is normal, from Theorem 2.5 it follows that $J = \{ h \in X \mid Dg(x,h) \in K(-C, g(x)) \} = \{ h \in X \mid g(x) + Dg(x,h) \in -C \}$, where the last identity follows from $K(-C, g(x)) = \cup_{\lambda \geq 0} \lambda(-C - g(x))$. The inclusion $T(S,x) \subseteq J$ is easy to see, hence we show the inverse. Let $h \in J$ be given, i.e., $g(x) + Dg(x,h) \leq 0$ μ almost everywhere. Then we have (neglecting sets with measure zero)

$$|x(t)| - \epsilon + h(t)\operatorname{sgn}(x(t)) \leq 0, \qquad t \in M := \{ t \in I \mid x(t) \neq 0 \}, \tag{51}$$

$$-\epsilon + |h(t)| \leq 0, \qquad t \in N := \{ t \in I \mid x(t) = 0 \}. \tag{52}$$

We now construct a sequence $\{h_n\}_{n \in \mathbb{N}} \subseteq X$ converging to h and a sequence $\{\lambda_n\}_{n \in \mathbb{N}}$ of positive real numbers with the properties $\lambda_n \to 0$, with $x + \lambda_n h_n \in S$, i.e., $|x + \lambda_n h_n| - \epsilon \leq 0$. This last statement then implies $h \in T(S,x)$. Define for $n \geq 1$ the sets $M^1 := \{ t \in M \mid |x(t)| = \epsilon \}$, $M^2 := \{ t \in M \mid |x(t)| < \epsilon \}$, $B_n := \{ t \in M \mid |h(t)| \leq n \}$, $M_n^1 := M^1 \cap B_n$, $M_n^2 := M^2 \cap \{ t \in M \mid |x(t)| \leq \epsilon - \frac{\epsilon}{2n} \}$. Obviously we have $M = M^1 \cup M^2$. The indicator function of a subset $I_0 \subseteq I$ is denoted by $\Phi(I_0)$, i.e., $\Phi(I_0)(t) = 1$ if $t \in I_0$ and $\Phi(I_0)(t) = 0$ if $t \in \mathbb{R} \setminus I_0$. Then we define

$$h_n(t) := h(t)(\Phi(N)(t) + \Phi(M_n^1)(t) + \Phi(M_n^2)(t)), \qquad \lambda_n := \min\left(1, \frac{\epsilon}{2n^2}\right).$$

Since $\Phi(M_n^1) \to \Phi(M^1)$ and $\Phi(M_n^2) \to \Phi(M^2)$ pointwise and, by Lebesgue's theorem of dominated convergence, in $L^2(I)$ as well, we can deduce $h_n \to h$ in $X = L^2(I)$. Moreover, we have $\lambda_n \to 0_+$. Now it remains to show that (neglecting sets of measure 0)

$$|x(t) + \lambda_n h_n(t)| \leq \epsilon \quad \forall\, t \in I, n \geq 1. \tag{53}$$

For $t \in I \setminus (N \cup M_n^1 \cup M_n^2)$ we have $h_n(t) = 0$; hence (53) holds since $|x| \leq \epsilon$.

For $t \in N$ we have $x(t) = 0$ and the inequality (52) implies $|x(t) + \lambda_n h_n(t)| = \lambda_n |h_n(t)| \leq \lambda_n \epsilon \leq \epsilon$.

For $t \in M_n^1$ we deduce with $|x(t)| = \epsilon$ from (51) that $0 \geq |x(t)| - \epsilon + \operatorname{sgn}(x(t))h(t) = \operatorname{sgn}(x(t))h(t)$. This implies $\operatorname{sgn}(x(t)) = -\operatorname{sgn}(h(t))$ for $t \in M_n^1 \subseteq M^1$ and hence, by $|h(t)| \leq n, \lambda_n \leq \epsilon/(2n^2)$, we deduce $|x(t) + \lambda_n h_n(t)| = |\ \epsilon \operatorname{sgn}(x) - \lambda_n |h_n| \operatorname{sgn}(x)\ | = |\epsilon - \lambda_n |h_n(t)| \leq \epsilon$.

Finally for $t \in M_n^2$ we conclude that $|x(t) + \lambda_n h_n(t)| \leq |x(t)| + \lambda_n |h_n(t)| \leq \epsilon - \epsilon/(2n) + n\epsilon/(2n^2) = \epsilon$. Hence we have shown that $x(t) + \lambda_n h_n \in S$, which implies $h \in T(S,x)$.

Acknowledgment. The author thanks Professor J. Jahn and the referees for their useful comments and suggestions.

REFERENCES

[1] J. P. Aubin and I. Ekeland, *Applied Nonlinear Analysis,* John Wiley and Sons, New York, 1984.

[2] J. P. Aubin, *Lipschitz behaviour of solutions to convex minimization problems,* Math. Oper. Res., 9 (1984), pp. 87–111.

[3] M. S. Bazaraa and C. M. Shetty, *Foundations of optimization,* Lecture Notes in Economics and Mathematical Systems 122, Springer-Verlag, Berlin, Heidelberg, New York, 1976.

[4] J. M. Borwein, *Proper efficient points for maximization with respect to cones,* SIAM J. Control Optim., 15 (1977), pp. 57–63.

[5] ———, *Weak tangent cones and optimization in a Banach space,* SIAM J. Control Optim., 16 (1978), pp. 512–522.

[6] ———, *Continuity and differentiability properties of convex operators,* Proc. London Math. Soc., 44 (1982), pp. 420–444.

[7] F. H. CLARKE, *Necessary conditions for nonsmooth problems in optimal control and the calculus of variations,* Ph.D. thesis, University of Washington, Seattle, WA, 1973.

[8] ———, *A new approach to Lagrange multipliers,* Math. Oper. Res., 1 (1986), pp. 165–175.

[9] ———, *Optimization and Nonsmooth Analysis,* Wiley Interscience, New York, 1983.

[10] V. F. DEMJANOV AND A.M. RUBINOV, *On quasidifferentiable mappings,* Math. Operationsforsch. Statist, Ser. Optim., 14 (1983), pp. 3–21.

[11] I. V. GIRSANOV, *Lectures on mathematical theory of extremum problems,* Lecture Notes in Economics and Mathematical Systems 67, Springer-Verlag, Berlin, Heidelberg, New York, 1972.

[12] M. GUIGNARD, *Generalized Kuhn–Tucker conditions for mathematical programming problems in a Banach space,* SIAM J. Control Optim., 7 (1969), pp. 232–241.

[13] J. B. HIRIART–URRUTY, *Tangent cones, generalized gradients and mathematical programming in Banach spaces,* Math. Oper. Res., 4 (1979), pp. 79–97.

[14] R. B. HOLMES, *Geometric Functional Analysis and its Applications,* Springer-Verlag, Berlin, Heidelberg, New York, 1975.

[15] L. HURWICZ, *Programming in linear spaces,* Studies in Linear and Nonlinear Programming, K. Arrow, L. Hurwicz, and H. Uzawa, eds., Stanford University Press, Stanford, CA, 1958.

[16] A. D. IOFFE, *Nonsmooth Analysis: Differential calculus of nondifferentiable mappings,* Trans. Amer. Math. Soc., 266 (1981), pp. 1–57.

[17] ———, *Necessary conditions in nonsmooth optimization,* Math. Oper. Res., 9 (1984), pp. 159–189.

[18] J. JAHN, *Mathematical Vector Optimization in Partially Ordered Linear Spaces,* Verlag Peter Lang, Frankfurt,1986.

[19] J. JAHN AND E. SACHS, *Generalized quasiconvex mappings and vector optimization,* SIAM J. Control Optim., 24 (1986), pp. 306–322.

[20] G. JAMESON, *Ordered Linear Spaces,* Lecture Notes in Mathematics 141, Springer-Verlag, Berlin, Heidelberg, New York, 1970.

[21] A. KIRSCH, W. WARTH, AND J. WERNER, *Notwendige Optimalitaetsbedingungen und ihre Anwendung,* Lecture Notes in Economics and Mathematical Systems 152, Springer-Verlag, Berlin, Heidelberg, New York, 1978.

[22] J. KLOSE, *Sensitivity analysis using the tangent derivative,* Numer. Funct. Anal. Optim., to appear.

[23] A.G. KUSRAEV, *On necessary conditions for an extremum of nonsmooth vector–valued mappings,* Dokl. Akad. Nauk SSSR Tom, 242 (1978), pp. 44–47..

[24] D.T. LUC, *Theory of vector optimization,* Lecture Notes in Economics and Mathematical Systems 319, Springer-Verlag, Berlin, Heidelberg, New York, 1990.

[25] W. MINAMI, *Weak Pareto-optimal necessary conditions in nondifferentiable multiobjective programming on a Banach space,* J. Optim. Theory Appl., 41 (1983), pp. 451–461.

[26] N. PAPAGEORGIOU, *Nonsmooth analysis on partially ordered vector spaces Part* 2 : *Nonconvex case,* Pacific J. Math., 109 (1983), pp. 463–495.

[27] A.L. PERESSINI, *Ordered Topological Vector Spaces,* Harper and Row, New York, 1967.

[28] P. PENOT, *On regularity conditions in mathematical programming,* Math. Programming Stud., 19 (1982), pp. 167–199.

[29] ———, *Open mapping theorems and linearization stability,* Numer. Funct. Anal. Optim., 8(1)(1985), pp. 21–35.

[30] T.W. REILAND, *Nonsmooth analysis of vector valued mappings with contributions to nondifferentiable programming,* Numer. Funct. Anal. Optim., 8(3,4)(1985–86), pp. 301–323.

[31] S.M. ROBINSON, *Stability theory for systems of inequalities, Part* II, SIAM J. Control Optim., 13(1976), pp. 497–513.

[32] ———, *Local structure of feasible sets in nonlinear programming, Part* I: *Regularity,* in Numerical Methods, V. Pereyra and A. Reinoza, eds., Springer-Verlag, Berlin, Heidelberg, New York, 1983.

[33] ———, *Local structure of feasible sets in nonlinear programming, Part* II: *Nondegeneracy,* Math. Programming Stud., 22 (1984), pp. 217–230.

[34] ———, *Local structure of feasible sets in nonlinear programming, Part* III: *Stability and Sensitivity,* Math. Programming Stud., 30 (1987), pp. 45–66.

[35] R.T. ROCKAFELLAR, *Generalized directional derivatives and subgradients of nonconvex functions,* Canad. J. Math., 39 (1980), pp. 257–280.

[36] ———, *Lagrange multipliers and subderivatives of optimal value function in nonlinear programming,* Math. Programming Stud., 17 (1982), pp. 28–66.

[37] W. RUDIN, *Functional Analysis,* Mc Graw–Hill, New York, 1973.

[38] E. SACHS, *Differentiability in optimization theory,* Math. Operationsforsch. Statist, Ser. Optim., 9(1978), pp. 497–513.

[39] T. STAIB, *Notwendige Optimalitätsbedingungen in der mehrkriteriellen Optimierung mit Anwendung auf Steuerungsprobleme,* Ph.D. Thesis, University of Erlangen–Nürnberg, Erlangen, FRG, 1989.

[40] L. THIBAULT, *Subdifferentials of nonconvex vector valued functions,* J. Math. Anal. Appl., 86 (1982), pp. 319–344.

[41] F. TRÖLTZSCH, *Optimality Conditions for Parabolic Control Problems and Applications,* Teubner Verlag, Leipzig, 1984.

[42] M. VALADIER, *Sous-différentiabilité de fonctions convexes à valeurs dans un espace vectoriel ordonné,* Math. Scand., 30-5 (1972), pp. 65–74.

[43] J. WARGA, *Controllability and a multiplier rule for nondifferentiable optimization problems,* SIAM J. Control Optim., 16 (1978), pp. 803–812.

[44] C. ZALINESCU, *A generalization of the Farkas lemma and applications to convex programming,* J. Math. Anal. Appl., 66 (1978), pp. 651–678.

[45] J. ZOWE AND S. KURCYUSZ, *Regularity and stability for the mathematical programming problem in Banach spaces,* Appl. Math. Optim., 5 (1979), pp. 49–62.

SIAM J. OPTIMIZATION
Vol. 2, No. 2, pp. 173-197, May 1992

001

DUAL METHODS IN ENTROPY MAXIMIZATION. APPLICATION TO SOME PROBLEMS IN CRYSTALLOGRAPHY*

ANDRÉE DECARREAU†, DANIELLE HILHORST‡, CLAUDE LEMARÉCHAL§, AND JORGE NAVAZA¶

Abstract. This paper is devoted to some infinite-dimensional optimization problems with finitely many constraints. These deal with entropy maximization, and this paper is particularly concerned with those originating from Fourier analysis.

These problems have a structure that makes them amenable to dual methods, for theoretical as well as numerical solutions. Existence results are recalled, and the use of duality to construct suitable and efficient optimization algorithms is demonstrated. Finally, the so-called phase-problem of crystallographers, which is of crucial importance in biology and pharmacology, is studied. Although it is a nonconvex optimization problem, a solution algorithm also based on duality is proposed and some numerical illustrations are given.

Key words. entropy maximization, image reconstruction, applications, large-scale problems, duality, decomposition

AMS(MOS) subject classifications. 42B05, 49A55, 49B40, 65K10, 92-08

1. Introduction. This paper deals with problems of the following general type:

$$\begin{cases} \inf H(p) & p \in B, \\ c_n(p) = 0 & n \in N, \end{cases} \tag{1.1}$$

where B is a Banach space, H is a convex function on B, each constraint c_n is a smooth function from B to $\mathbb{R}$, and N is a finite set of $|N|$ indices; we will denote by $\mathbb{R}^N$ the set of real $|N|$-tuples whose coordinates are indexed in N.

Most of our analysis will concern the linearly constrained case. With a particular application in mind, however (the so-called phase problem in X-ray crystallography [18], [20]), we will also consider a class of problems where the constraints c_n are quadratic functions. In the specific case that we consider, B is $L_1(\Omega)$, where Ω is a bounded open set in $\mathbb{R}^d$ (in most applications, $d = 3$), and the objective function has the form

$$H(p) := \int_\Omega h[p(r)]\,dr \tag{1.2}$$

where h is a strictly convex function of a single real variable.

The aim of this paper is mainly: (i) to study briefly the existence and characterization of a solution (using earlier works of Rockafellar [30], [31]; Ekeland and Temam [11]; Ben Tal, Borwein, and Teboulle [1], [2]; and Borwein and Lewis [4], [5]); (ii) to introduce a very interesting class of problems in crystallography dealing with entropy optimization and Fourier analysis; and above all (iii) to demonstrate some numerical solution schemes of (1.1), (1.2), based on duality (a technique that we think is generally overlooked in the mathematical programming community, although it goes back at

* Received by the editors February 10, 1990; accepted for publication (in revised form) March 16, 1991.

† Département de Mathématiques, Université de Poitiers, 40 avenue du Recteur Pineau, 86022 Poitiers, France.

‡ Laboratoire d'Analyse Numérique, Centre Nationale de la Recherche Scientifique and Université de Paris-Sud, 91405 Orsay, France.

§ Institut National de Recherche en Informatique et en Automatique, BP 105, 78153 Le Chesnay, France.

¶ Centre Pharmaceutique, Université de Paris-Sud, 92290 Châtenay-Malabry, France.

least to [33]; see also [13]). The new material in this paper is limited to §§ 4 and 5, devoted to (ii); the other sections should be considered as advertisement.

We find it convenient to view h in (1.2) as an *extended-valued* function (see [32] for a general introduction to extended calculus). This means that, the convex function h being defined a priori on a convex set $\operatorname{dom} h \subset \mathbb{R}$, we extend h outside its domain by setting

$$h(t) := +\infty \quad \text{for } t \notin \operatorname{dom} h.$$

Thus, we obtain a function defined on the whole $\mathbb{R}$, with values in $]-\infty, +\infty]$; $\operatorname{dom} h$ is the set of t such that $h(t) < +\infty$. Since $\operatorname{dom} h$ is convex in $\mathbb{R}$, there are two numbers m and M, with $-\infty \leqq m < M \leqq +\infty$, such that

$$h(t) < +\infty \quad \text{for } t \in]m, M[,$$

$$h(t) = +\infty \quad \text{for } t \notin [m, M].$$

As for the boundary values, we will assume that h is continuous on its domain: there holds

$$h(m) = \lim_{t \searrow m} h(t), \qquad h(M) = \lim_{t \nearrow M} h(t),$$

knowing that each of the above limits can be infinite. This assumption simply means that h is lower semicontinuous (on $\mathbb{R}$), or "closed," in the terminology of [32]. Naturally, lower semicontinuity is essential for the existence of a minimum.

Here are various examples of possible objective functions:

(1.3) $$h_1(t) := \tfrac{1}{2}t^2 \qquad (m = -\infty, M = +\infty; h_1(m) = h_1(M) = +\infty),$$

(1.4) $$h_2(t) := \exp t \qquad (m = -\infty, M = +\infty; h_2(m) = 0, h_2(M) = +\infty),$$

(1.5) $$h_3(t) := \begin{cases} -\operatorname{Log} t & \text{for } t > 0 \quad (m = 0, M = +\infty; h_3(m) = +\infty, h_3(M) = -\infty), \\ +\infty & \text{for } t \leqq 0, \end{cases}$$

(1.6) $$h_4(t) := \begin{cases} t \operatorname{Log} t & \text{for } t > 0 \quad (m = 0, M = +\infty; h_4(m) = 0; h_4(M) = +\infty), \\ +\infty & \text{for } t < 0, \end{cases}$$

(1.7) $$h_5(t) := \begin{cases} -(2\pi)^{-1/2} \exp(-s^2/2) & \text{for } t \in]0, 1[, \\ +\infty & \text{for } t \notin]0, 1[, \end{cases}$$

with s defined by $t = (2\pi)^{-1/2} \int_{-\infty}^{s} \exp(-u^2/2)\, du$. Here, $m = 0$, $M = 1$, $h_5(m) = h_5(M) = 0$.

For pedagogical purposes, we also mention (although it is not of great practical interest)

(1.8) $$h_6(t) := \begin{cases} \tfrac{1}{2}t^2 & \text{for } |t| \leqq 1, \\ +\infty & \text{for } |t| > 1. \end{cases}$$

The difference between (1.7) and (1.8) is essentially that h_5 (respectively, h_6) has infinite (respectively, finite) slopes at m and M.

Actually, h_2 and h_3 will be ruled out from our development: they do not increase at infinity and, in this case, little can be said about the existence of an optimum in (1.1), (1.2). We do need h to be *coercive*, i.e.,

(1.9) $$\frac{h(t)}{|t|} \to +\infty \quad \text{when } |t| \to \infty,$$

an assumption that will be in force throughout our paper. Beware that (1.9) is not the definition of coercivity used in, for example, [11]: it is essential that h tend to infinity faster than a linear function.

Remark 1.1. The rationale for (1.1), with H given by (1.2), can be explained as follows: the constraints represent measurements, which provide partial information concerning the unknown function p from Ω to $\mathbb{R}$. On the other hand, one knows for sure that $p(r)$ (for almost all $r \in \Omega$) must lie between m and M. Then, one seeks a p which is "best" among those functions satisfying the given properties. The wording "best" is made precise by the exact form of h between m and M; $-H$ is usually called an entropy. Observe in passing that strict convexity is a natural property of h: its role is to select a *unique* convenient p satisfying the constraints.

We refer to [23] for the definition of various entropies, in particular (1.7), based on statistical considerations. The idea is to introduce an optimal probability measure—in the sense of [19]; see also [7]—on the set of all admissible functions; then, our unknown p is defined as the averaged admissible function, in the sense of this optimal measure. Changing the set of admissible functions changes h.

Linear constraints, now, are most generally defined by

$$c_n(p) := \int_\Omega f_n(r)p(r)\,dr - z_n, \tag{1.10}$$

where each f_n is in $L_\infty(\Omega)$ and $z_n \in \mathbb{R}$.

A first example of such possible constraints is $f_n := \chi(\Omega_n)$, the characteristic function of some open set Ω_n, i.e.,

$$f_n(r) = 1 \quad \text{if } r \in \Omega_n; \quad f_n(r) = 0 \quad \text{if } r \notin \Omega_n. \tag{1.11}$$

Here, $\bigcup\{\Omega_n \colon n \in N\} \subset \Omega$. The nth constraint is therefore

$$\int_{\Omega_n} p(r)\,dr = z_n,$$

which is encountered in various problems of image reconstruction, tomography, etc. [15], [16].

Another type of constraint consists of fixing some Fourier coefficients of p, considered as the restriction to Ω (the mesh) of a periodic function on $\mathbb{R}^d$. Then, with $\Omega :=]0, 2\pi[^d$, c_n is actually complex-valued:

$$f_n(r) := \exp(in \cdot r), \tag{1.12}$$

where n is a d-tuple of d integers in $\mathbb{R}^d$, N is a finite subset of $\mathbb{N}^d$, and

$$n \cdot r := n_1 r_1 + n_2 r_2 + \cdots + n_d r_d$$

is the usual dot-product on $\mathbb{R}^d$ (note the change of notation: z_n of (1.10) is now a complex number, or a couple of real numbers). It is generally the case that $0 \in N$, which fixes the mean value of p on Ω. This kind of constraint is encountered in crystallography, where p represents the electronic density of a given crystal.

To sum up, the problem of interest to us in this paper is essentially

$$\begin{cases} \inf \displaystyle\int_\Omega h[p(r)]\,dr =: H(p), & p \in L_1(\Omega), \qquad (1.13) \\ \displaystyle\int_\Omega f_n(r)p(r)\,dr = z_n, & n \in N \qquad (1.14) \end{cases}$$

(except in §§ 4 and 5, dealing with special nonlinear constraints). The following assumptions will hold throughout.

ASSUMPTIONS 1.2 (standing assumptions).

* Ω is a bounded set in $\mathbb{R}^d$;

* h is a closed (i.e., lower semicontinuous or l.s.c.) strictly convex and coercive function of the real variable;

* N is a finite set, and the given functions f_n are in $L_\infty(\Omega)$.

2. Theoretical background—existence and characterization of a solution. Duality will play an essential role in our development, for theoretical as well as practical purposes. A useful tool for this is the concept of conjugacy: we denote by

$$k(s) := \sup\{st - h(t)\colon t \in \mathbb{R}\} \tag{2.1}$$

the so-called conjugate function of h—usually denoted h^*—and we recall that k is a l.s.c. convex function (because it is a sup of linear continuous functions). Since h is evidently minorized by some affine function:

$$h(t) \geqq a + bt \quad \text{for some } a,\ b, \text{ and all } t \text{ in } \mathbb{R}, \tag{2.2}$$

k is not identically $+\infty$ ($k(b) \leqq -a$!). In fact, coercivity of h implies that the domain dom k of k is the whole $\mathbb{R}$. For an illustration, apply (2.1) to examples (1.3)-(1.8): direct calculations give the maximal t (when it exists) for given s, and we obtain

$$k_1(s) = \tfrac{1}{2}s^2, \tag{2.3}$$

$$k_2(s) = \begin{cases} s \operatorname{Log} s - s & \text{for } s > 0, \\ 0 & \text{for } s = 0, \\ +\infty & \text{for } s < 0 \end{cases} \tag{2.4}$$

(the domain of k_2 is now $[0, +\infty[$: st; e^t has no supremum if $s < 0$)

$$k_3(s) = \begin{cases} -1 - \operatorname{Log}(-s) & \text{for } s < 0, \\ +\infty & \text{for } s \geqq 0, \end{cases} \tag{2.5}$$

$$k_4(s) = \exp(s - 1), \tag{2.6}$$

$$k_5(s) = (2\pi)^{-1/2}\left[s \int_{-\infty}^{s} \exp(-u^2/2)\, du + \exp(-s^2/2) \right], \tag{2.7}$$

$$k_6(s) = \begin{cases} \tfrac{1}{2}s^2 & \text{for } |s| \leqq 1, \\ |s| - \tfrac{1}{2} & \text{for } |s| \geqq 1. \end{cases} \tag{2.8}$$

It is a consequence of coercivity that (2.1) has a solution for all s, but strict convexity of h implies also that this solution $t(s)$ is unique, and is indeed the derivative of k:

$$k'(s) = t(s). \tag{2.9}$$

In a word, our standing assumptions, Assumptions 1.2, imply that k is a continuously differentiable convex function on $\mathbb{R}$. It results directly from (2.1) and (2.9) that

$$sk'(s) - h[k'(s)] = k(s) \geqq st - h(t) \quad \text{for all } s \text{ and } t. \tag{2.10}$$

Remark 2.1. All these classical properties are fairly well illustrated by (2.3)-(2.8). The function k is not so important per se; in fact, the important thing is the mapping $s \mapsto t(s)$ in (2.1), sometimes called the Legendre transform, which is well defined thanks to strict convexity of h. Its crucial property is that it is continuous, and can be expressed as the derivative of a smooth convex function, namely k.

The existence and value (2.9) of $k'(s)$ comes from [8]. Observe the "miracle": k' is just the partial derivative of the maximand in (2.1) (and yet, $t(s)$ does depend on s!). The following formal calculation will help us understand it: plug the value $t(s)$ in (2.1) and differentiate with respect to s, thus obtaining

$$k'(s)=t(s)+st'(s)-h'[t(s)]t'(s);$$

then realize that $s-h'[t(s)]=0$ because $t(s)$ solves (2.1). The aim of our notation s is to suggest that the argument of k is nothing but a *slope* of h. Likewise, (2.9) indicates that t is a slope of k. Indeed, the conjugacy operation applied to the convex function k gives back h.

We leave it as an exercise to compute the conjugate function of, say, $h(t):=\frac{1}{2}t^2+|t|$, to realize that differentiability of h has nothing to do with smoothness of k or $t(\cdot)$.

Naturally, just as H is obtained by integrating h, we can define the convex function

$$K(q):=\int_\Omega k[q(r)]\,dr, \tag{2.11}$$

which, because dom $k=\mathbb{R}$, is finite for all $q\in L_\infty(\Omega)$.

These preliminaries place us in a position to prove existence for (1.13), (1.14). We note first that the condition $p\in\operatorname{dom} H$ is implicit in our formulation, and can be interpreted as additional (inequality) constraints. Then, a necessary condition for existence is of course nonemptiness of the feasible domain, i.e.,

(2.12) there exists $\tilde{p}\in L_1(\Omega)$ satisfying (1.14) and such that $H(\tilde{p})<+\infty$.

THEOREM 2.2. *Under the standing assumptions, Assumptions* 1.2:

(i) *In* (1.13), *the function* $H: L_1(\Omega)\mapsto\,]-\infty,+\infty]$ *is convex and l.s.c. for the weak topology.*

(ii) *For any function* $f\in L_\infty(\Omega)$, *the sublevel-sets*

$$\left\{p\in L_1(\Omega)\colon H(p)+\int_\Omega f(r)p(r)\,dr\leqq\alpha\right\} \tag{2.13}$$

are weakly compact.

(iii) *If* (2.12) *holds, then* (1.13), (1.14) *has a unique optimal solution.*

Proof. In the terminology of [30], the h of (1.13), which does not depend explicitly on r, is a (particularly simple) normal convex integrand. In particular, if we take $p_0(r)\equiv t_0$ and $q_0(r)\equiv s_0$ with $t_0\in\operatorname{dom} h$ and s_0 arbitrary, then $H(p_0)<+\infty$ and $K(q_0)<+\infty$. With these observations, [29, Thm. 1] ensures that H is a well-defined convex function; from [29, Thm. 2], it is even a conjugate function (of K), hence l.s.c.; (i) is classical (see, for example, [11, Chap. I, Cor. 2.2]).

Now, coercivity of h is just finiteness of $k(s)$ for all s, so (ii) is [30, Cor. 2B]. Then, the proof of (iii) is classical: if $\{p_j\}$ is a feasible minimizing sequence, (ii) ensures that it has a weak cluster-point p^*. Clearly, p^* is feasible (and would be so even if N were infinite) and from (i),

$$H(p^*)\leqq\liminf H(p_j),$$

in which the right-hand side is just $\lim H(p_j)$, the optimal value of (1.13), (1.14). As for uniqueness, it trivially results from strict convexity of h, hence of H. □

Remark 2.3. Examples are known of optimization problems in L_1 whose objective function is merely increasing at infinity, and whose optimal "solutions" contain Dirac measures; see [3]. In these problems, a minimizing sequence is bounded (in L_1) but this does not suffice to imply the existence of a cluster-point (the bounded sets of $L_1(\Omega)$ are not compact, even for the weak topology: L_1 is not reflexive).

Indeed, coercivity of h is the key to (ii), which in turn is essential for (iii). The convex analysis setting of [31], used to prove (ii), somehow hides this point. We will see in § 4 a proof based on functional analysis, applied to a problem with nonlinear (hence nonconvex) constraints.

We now turn to duality: taking the Lagrange function

$$L(p,\lambda):=\int_\Omega \left\{h[p(r)]-\sum_{n\in N}\lambda_n f_n(r)p(r)\right\}dr+\sum_{n\in N}\lambda_n z_n, \tag{2.14}$$

the fundamental question is whether there is $\lambda^*\in\mathbb{R}^N$ such that $L(\cdot,\lambda^*)$ is stationary at the solution p^*. Being convex, the function $p\mapsto L$ is stationary when it is minimal, so we consider the *dual function* defined by

$$D(\lambda):=\inf\{L(p,\lambda): p\in L_1(\Omega)\}. \tag{2.15}$$

Observe the usefulness of Theorem 2.2(ii) (which results from coercivity of h): it implies that the above infimum in (2.15) is attained, just as in Theorem 2.2(iii). The next result, also classical, will show that the minimizing p is simply obtained by minimizing pointwise the integrand in (2.14). Then the role of (2.1) pops up, exhibiting a key function of r, which we denote $\lambda^TF\in L_\infty(\Omega)$, and whose value at $r\in\Omega$ is

$$\lambda^TF(r):=\sum_{n\in N}\lambda_n f_n(r) \tag{2.16}$$

(F denoting the vector $(f_n)_{n\in N}$, the above function is really the dot-product of $\lambda\in\mathbb{R}^N$ with $F(r)\in\mathbb{R}^N$).

PROPOSITION 2.4. *For each $\lambda\in\mathbb{R}^N$, L has a unique minimizer p_λ, which is actually in $L_\infty(\Omega)$. The value of p_λ at $r\in\Omega$ is*

$$p_\lambda(r)=k'[\lambda^TF(r)]. \tag{2.17}$$

The dual function (2.15) *is given by*

$$D(\lambda)=\lambda^Tz-K(\lambda^TF)=\lambda^Tz-\int_\Omega k[\lambda^TF(r)]\,dr. \tag{2.18}$$

It is concave, continuously differentiable, and its partial derivatives are

$$\frac{\partial D(\lambda)}{\partial\lambda_n}=z_n-\int_\Omega p_\lambda(r)f_n(r)\,dr, \qquad n\in N. \tag{2.19}$$

Proof. The profound reason is that K and H are conjugate to each other; see [30] again. However, the stated result is essential for numerics; we therefore give a self-contained proof, by showing a posteriori that (2.17) minimizes L.

First, observe that

$$p_\lambda\in L_\infty(\Omega)\subset L_1(\Omega)$$

because $\lambda^TF\in L_\infty(\Omega)$ and k' is continuous. Now take an arbitrary function $p\in L_1(\Omega)$ and apply (2.10) with $s=\lambda^TF(r)$ and $t=p(r)$:

$$\lambda^TF(r)p_\lambda(r)-h[p_\lambda(r)]=k[\lambda^TF(r)]\geqq\lambda^TF(r)p(r)-h[p(r)]$$

almost everywhere on Ω. Changing signs, integrating, and adding λ^Tz gives

$$L(p_\lambda,\lambda)=-\int_\Omega k[\lambda^TF(r)]\,dr+\lambda^Tz\leqq L(p,\lambda).$$

This shows that $D(\lambda)=L(p_\lambda, \lambda)$, which is exactly (2.18). Finally, the integrand in (2.18) is differentiable with respect to λ for almost every r; its derivatives are bounded for bounded λ. We can therefore differentiate under the integral (see, for example, Theorem 13.8.6 of [10]) to obtain the derivatives (2.19). □

Remark 2.5. Thus, unlike (1.13), (1.14), the optimization problem in (2.15) is easy: it has the explicit solution (2.17). This comes from decomposability of H (and of the constraints); the whole interest of the dual approach lies precisely here.

The differentiability of D is just what can be expected when one accepts (2.9): $\partial D/\partial\lambda_n$ is nothing but $\partial L/\partial\lambda_n$, the partial derivative of L with respect to λ. Once again, this is the "miracle" of [8]; see also [17] for a simple proof of it.

As already said, the constraint $H(p)<+\infty$ is implicit in the formulation (1.13), (1.14); it implies the constraints

$$m \leqq p(r) \leqq M \quad \text{for almost every } r\in\Omega,$$

which, in our infinite-dimensional situation, do not define a polyhedron. Hence, some qualification condition is needed and we make the Slater-type assumption:

(2.20) there is $\tilde{p}$ in (2.12) with $m<\tilde{p}(r)<M$ for almost every $r\in\Omega$.

Then, the existence of Lagrange multipliers relies on the following technical result, essentially due to [1], [2], [4], and [5], with a slight extension of Borwein.

LEMMA 2.6. *If* (2.20) *holds, the dual function D of* (2.18) *has a maximum.*

Proof. We use [5], in particular its Theorem 2.4 (where X is $L_1(\Omega)$, f is H, A and b symbolize the constraints (1.10), and P is $\{0\}\in\mathbb{R}^N$). For this, we prove that the $\tilde{p}$ of (2.20) is in the quasi-relative interior (qri) of dom $H\cap C$, with

$$C := \{p\in L_1(\Omega)\colon m\leqq p(r)\leqq M \text{ a.e.}\}.$$

First, observe that

$$\{p\in L_\infty(\Omega)\colon m<\operatorname{ess\,inf} p\leqq \operatorname{ess\,sup} p<M\}\subset \operatorname{dom} H,$$

so that dom H is dense in C. Now, we consider three cases.

(i) $M=-m=+\infty$. Then observe that $L_\infty(\Omega)\subset$ dom H, hence the closure of dom H is the whole $L_1(\Omega)$. Apply [5, Lemma 2.3] with dom $H=C_1\subset C_2=L_1(\Omega)$: qri dom $H=$ dom $H\ni\tilde{p}$.

(ii) In the case where either m or M but not both, is infinite, the result just follows from [5, Cor. 2.6] (by translation and/or symmetry, the case $m=0$ or $M=0$ generalizes to arbitrary m or M in a straightforward way).

(iii) Now let m and M be finite. We just observed that we may assume $m=0$, and of course $M>0$. Then define

$$\Gamma := \{(u,v)\in[L_1(\Omega)]^2\colon u\geqq 0,\, v\geqq 0,\, u+v=M \text{ a.e.}\}.$$

According to [4, Cor. 3.16], together with [4, Ex. 3.11(i)],

$$\operatorname{qri}\Gamma=\{(u,v)\in\Gamma\colon u>0,\, v>0 \text{ a.e.}\}.$$

Now define the linear continuous operator A that, to $(u,v)\in[L_1(\Omega)]^2$, associates $A(u,v)=u\in L_1(\Omega)$. Observe that

$$A\Gamma=\{p\in L_1(\Omega)\colon 0\leqq p\leqq M \text{ a.e.}\}=C.$$

Clearly, $\tilde{p}$ of (2.20) is in $A(\operatorname{qri}\Gamma)$, hence in qri $(A\Gamma)$ by virtue of [4, Prop. 2.22].

Finally, since dom H is dense in $C=A\Gamma$, the rest follows as in (i): with the help of [5, Lemma 2.3] (used with $C_2=A\Gamma$), we prove that $\tilde{p}\in\operatorname{qri}(\operatorname{dom} H)=\operatorname{qri}(\operatorname{dom} H\cap C)$. □

THEOREM 2.7. *With the standing assumptions, Assumptions* 1.2, *let* λ *be any maximizer of* D. *The corresponding* p_λ *of* (2.17) *is the solution of* (1.13), (1.14), *whose optimal value is just* $D(\lambda)$.

Proof. Expressing in (2.19) that $\nabla D(\lambda)=0$ gives

$$\int_\Omega p_\lambda(r) f_n(r)\,dr = z_n \quad \text{for all } n \in N,$$

i.e., p_λ is feasible. Also, $D(\lambda)=L(p_\lambda,\lambda)=H(p_\lambda)$. For any other feasible p, there holds

$$H(p)=L(p,\lambda)\geqq D(\lambda)=L(p_\lambda,\lambda)=H(p_\lambda). \qquad \square$$

We purposely give the optimality condition in a rather unusual form (usually, one establishes that the solution p^* of (1.1) is the p-part of a saddle point of L). In fact, our statement of Theorem 2.7 is inspired by the suitable methods to maximize (2.18): they will of course be based on some gradient-type algorithm.

A key to this approach is the strict convexity of h, which implies uniqueness of p_λ. In the language of Wolfe's duality [34], the condition $\nabla_p L(p,\lambda)=0$ can be used to eliminate the primal variable from Wolfe's dual problem. If h were merely convex (as for the example in linear programming), a dual solution λ^* would not readily furnish a primal solution: a selection among the minimizers of $L(\cdot,\lambda^*)$ should be needed. This selection would be automatically performed by special maximization methods, say, bundle methods (see [21], [29]), which incidentally would be made necessary because D is differentiable only thanks to *strict convexity* of $L(\cdot,\lambda)$.

Remark 2.8. Note a curious consequence of Theorem 2.7: up to the nonlinear but continuous mapping k', the primal solution is a linear combination of the finitely many constraint functions f_n. As such, it has the same smoothness: for example, $p^* \in C_\infty(\Omega)$ in the case of the Fourier constraints (1.12); or p^* is a step-function for constraints such as (1.11).

Primal-dual existence is studied much more deeply in [1], [2], [4], [5]. It is shown, in particular, that coercivity of h is not needed to prove Lemma 2.6. Along this line, we mention the following instructive counterexample, which can be found in [24].

With $\Omega :=]-\pi,+\pi[^3$ (the counterexample would not work with $d\leqq 2$), consider the following problem (1.12)-(1.14): take the (noncoercive) entropy of (1.5) and the four Fourier constraints

$$\int_\Omega p(r_1,r_2,r_3)\,dr_1\,dr_2\,dr_3 = 1, \tag{2.21}$$

$$\int_\Omega p(r_1,r_2,r_3)\cos r_n\,dr_1\,dr_2\,dr_3 = z, \qquad n=1,2,3, \tag{2.22}$$

with given $z\in[0,1[$. It can be proved that the feasibility assumption (2.20) holds. For this, start from the function

$$p_\alpha(r) := \begin{cases} 1/\alpha^3 & \text{if } r \in]0,\alpha[^3, \\ 0 & \text{otherwise.} \end{cases}$$

It satisfies (2.21), and (2.22) is satisfied if α solves the equation

$$\frac{\sin\alpha}{\alpha}=z,$$

which has one solution in $]0,\pi]$ if $0\leqq z<1$. Then apply [5, Thm. 2.9]: there is $\varepsilon>0$ such that p_α can be modified to take its values in $[\varepsilon,+\infty[$.

Now, the dual function D of (2.18) is (up to a constant)

$$D(\lambda_0, \lambda_1, \lambda_2, \lambda_3) = \begin{cases} \lambda_0 + z \sum_{n=1}^{3} \lambda_n + \int_\Omega \operatorname{Log}\left[-\lambda_0 - \sum_{n=1}^{3} \lambda_n \cos r_n\right] dr \\ \qquad \text{if } \lambda_0 + \sum_{n=1}^{3} \lambda_n \cos r_n < 0 \quad \text{a.e. in } \Omega, \\ -\infty \quad \text{otherwise.} \end{cases}$$

By virtue of [2, Thm. 2.1] or [5, Thm. 2.4], it has a maximum and, by symmetry, it has a maximum for $\lambda_1 = \lambda_2 = \lambda_3$. Restricting our study to this case, we set $\mu := \lambda_1/\lambda_0$ and write the above dual constraint

$$|\mu| \leqq \tfrac{1}{3} \quad \text{and} \quad \lambda_0 < 0,$$

while the function p_λ of (2.17) becomes $-P(\mu, r)/\lambda_0$, with the positive function

$$P(\mu, r) := \left[1 + \mu \sum_{n=1}^{3} \cos r_n\right]^{-1}.$$

Feasibility in (2.21), (2.22) is:

$$I_0(\mu) := \int_\Omega P(\mu, r)\, dr = -\lambda_0, \tag{2.23}$$

$$I_1(\mu) := \int_\Omega \cos r_1 \, P(\mu, r)\, dr = -z\lambda_0 = zI_0(\mu) \tag{2.24}$$

(knowing that (2.24) stands for three equations, which are identical). That is, the equation

$$R(\mu) := \frac{I_1(\mu)}{I_0(\mu)} = z$$

must have some solution in $[-\frac{1}{3}, +\frac{1}{3}]$, and then λ_0 is given by (2.23). We claim that this is not possible, at least not for all values of $z \in]0, 1[$.

The crucial point is that I_0 and I_1 are convergent integrals (for $\mu = \pm\frac{1}{3}$). At $\mu = -\frac{1}{3}$, for example, develop each $\cos r_n$ near the singularity $r_n = 0$ and use polar coordinates: P diverges at the speed ρ^{-2}, which is balanced by the element of volume $\rho^2 \cos\theta\, d\rho\, d\theta\, d\varphi$ (note: $d > 2$ is essential). We can even say more: for r in some neighbourhood B of $(0, 0, 0)$, the coefficient of μ in $P(\mu, r)$ is positive; hence $P(\mu, \cdot) \leqq P(-\frac{1}{3}, \cdot) \in L_1(B)$. We deduce by Lebesgue's Theorem [10] that $I_0(\mu)$ is continuous at $\pm\frac{1}{3}$. The same argument applies to I_1 (cut B in two parts).

Then R is a continuous function on $[-\frac{1}{3}, +\frac{1}{3}]$, and it is clear enough that $|R(\mu)| < 1$ for all $\mu \in [-\frac{1}{3}, +\frac{1}{3}]$; it has a maximum, say, $\bar{z} < 1$. Conclusion: if z is close enough to 1, a dual solution is of no help to recover a primal solution: the corresponding p_λ of (2.17) is a nice function of $L_1(\Omega)$, but not even feasible in the primal problem. In summary, Theorem 2.7 can fail in the absence of coercivity.

3. Linear constraints: Computing a solution. The way to solve (1.13), (1.14) is clearly indicated by the results of § 2. One can maximize D to obtain a solution λ^*; then, it suffices to compute the primal solution as in Theorem 2.7. This technique is by no means new, and goes back at least to [33]; see also [8].

Thus, we are in a fairly favourable situation: the dual problem is set in $\mathbb{R}^N$, the space of constraint values, much simpler than the primal space; furthermore, it has no constraints, the objective function (to be maximized) is concave, and its gradient is made available by (2.19). On the other hand, a restrictive aspect is that (2.20) must hold, in addition to the necessary assumption (2.12).

Generally speaking, the dual algorithm will work as follows.

ALGORITHM 3.1. Choose an initial $\lambda^0 \in \mathbb{R}^N$; set $j=0$.
Step 1. Compute $p^j := p_{\lambda_j}$ of (2.17) and the corresponding constraint and L-values.
Step 2. Stop if λ^j is approximately dual-optimal, i.e., if p^j is approximately feasible.
Step 3. With the help of some gradient-type optimization algorithm, select a new λ^{j+1} aimed at increasing D; increase j by 1 and loop to Step 1.

A discretization of $L_\infty(\Omega)$ is necessary in Step 1 to compute the various integrals defining the problem. These relatively easy computations are under the responsibility of the user who needs to solve (1.13), (1.14). As for Step 3, the algorithm designer will often be faced with rather big sets N, say, $|N|=10^4$; in such cases, it is appropriate to use conjugate gradient methods, or limited-memory variable metric methods [6], [25].

Remark 3.2. With D just continuously differentiable, convergence can be proved for a method of the steepest descent-type only. For more sophisticated methods, one needs additional smoothness of D, i.e., of k, which in turn requires more than strict convexity of h. To prove convergence of, say, BFGS when ∇D is Lipschitz continuous, use the lemma in § 3 of [26] in conjunction with [27].

We note the expression of the second derivatives of D, i.e., the Jacobian of the vector (2.19):

$$\frac{\partial^2 D(\lambda)}{\partial\lambda_n \partial\lambda_m} = -\int_\Omega k''[\lambda^T F(r)] f_n(r) f_m(r)\, dr,$$

which makes it possible to use second-order methods, at least formally. Needless to say, second differentiability of D requires additional properties of h.

We have to check that this dual technique does solve our primal problem. This property does not follow from Theorem 2.7: the fact that $\nabla D(\lambda^j) \to 0$, i.e., that p^j is asymptotically feasible (assuming that λ^j tends to a dual solution), does not imply that p^j tends to a primal solution.

PROPOSITION 3.3. *The function* $\lambda \mapsto p_\lambda$ *is continuous from* $\mathbb{R}^N$ *to* $L_\infty(\Omega)$.

Proof. Let $\|\lambda\|_1$ be the 1-norm of $\lambda \in \mathbb{R}^N$ and set $\|F\|_\infty := \max\{\|f_n\|_\infty : n \in N\}$. The majoration

$$\forall \lambda \in \mathbb{R}^N, \qquad \|\lambda^T F\|_\infty \leqq \|\lambda\|_1 \|F\|_\infty \tag{3.1}$$

is easy to obtain. Now, suppose $\lambda \to \lambda^*$, which, by (3.1), implies

$$\|\lambda^T F - \lambda^{*T} F\|_\infty \to 0.$$

Also, the convergent λ is bounded and (3.1) again implies that the two functions $\lambda^T F$ and $\lambda^{*T} F$ vary in some fixed interval $[-A, +A]$, on which the continuous function k' is uniformly continuous. This allows us to deduce

$$\|k'(\lambda^T F) - k'(\lambda^{*T} F)\|_\infty \to 0. \qquad \square$$

Thus, in Algorithm 3.1, the convergence of p^j to the primal solution just depends on the convergence of λ^j to a dual solution. As far as numerical algorithms are concerned, such a convergence can be forced rather easily, with the help of line-searches or similar *stabilizing* techniques. By contrast, if the λ-problem is viewed as a mere system of equations

$$(3.2) \qquad \int_\Omega k'[\lambda^T F(r)]f_n(r)\,dr = z_n, \qquad n \in N$$

expressing the fact that p_λ must be feasible (see [12]), then stability may be harder to obtain. It is very helpful to interpret (3.2) as the dual optimality condition $\nabla D = 0$, in which the concave function D must be maximized.

With any reasonable algorithm, $\{\lambda^j\}$ will be automatically bounded if D is sup-compact. This latter property, however, does not trivially hold, in particular because $\lambda \mapsto \lambda^T F$ may be rank-deficient; so the sequence $\{\lambda^j\}$ might diverge if one is not careful enough in Algorithm 3.1, especially with a Newton method. Indeed, consider the subspace

$$(3.3) \qquad Z := \left\{\zeta \in \mathbb{R}^N : \exists p \in L_1(\Omega) \text{ with } \int_\Omega f_n(r)p(r)\,dr = \zeta_n,\, n \in N\right\},$$

which is nothing but the range of the linear operator $L_1(\Omega) \mapsto \mathbb{R}^N$ mapping p onto the constraint values in (1.14). A necessary condition for existence is, of course, $z \in Z$ (not even mentioning that the p in (3.3) must also have $p(r) \in [m, M]$ for almost every $r \in \Omega$).

PROPOSITION 3.4. *Denote by*

$$Z^\perp := \{\mu \in \mathbb{R}^N : \mu^T\zeta = 0 \textit{ for all } \zeta \in Z\}$$

the orthogonal complement of Z. For all $\lambda \in \mathbb{R}^N$ *there holds*

$$(3.4) \qquad \forall \mu \in Z^\perp, \qquad D(\lambda + \mu) = D(\lambda) + \mu^T z.$$

It follows that, if $z \notin Z$, *then* $\sup D = +\infty$. *If* $z \in Z$, *then the set (possibly empty) of maximizers of D over* $\mathbb{R}^N$ *is just* $\Lambda^* + Z^\perp$, *where* Λ^* *is the set (possibly empty) of solutions of*

$$(3.5) \qquad \sup\{D(\lambda) : \lambda \in Z\}.$$

Proof. Observe that L is affine in λ: for all $p \in L_1(\Omega)$, λ and μ in $\mathbb{R}^N$,

$$L(p, \lambda + \mu) = L(p, \lambda) + \sum_n \mu_n \left[z_n - \int_\Omega f_n(r)p(r)\,dr\right].$$

By definition, the integral in the bracket is in Z, so (3.4) is established ($p_{\lambda+\mu} = p_\lambda$, indeed). If $z \notin Z$, i.e., z has a nonzero projection z' onto $Z^\perp$, take $\mu = \alpha z'$ with $\alpha \to +\infty$ to exhibit unboundedness of D. If $z \in Z$, then $\mu^T z = 0$ and the rest follows without difficulty. □

Accordingly, (3.5) is the relevant dual problem to solve: each λ in Algorithm 3.1 should actually be projected onto Z. Our next result shows how to do it.

PROPOSITION 3.5. *We have*

$$Z^\perp = \left\{\lambda \in \mathbb{R}^N : \sum_{n \in N} \lambda_n f_n(r) = 0 \text{ a.e.}\right\} = \{\lambda \in \mathbb{R}^N : \Phi\lambda = 0\}$$

where $\Phi \in \mathbb{R}^{N \times N}$ *is defined by*

$$\Phi_{nm} := \int_\Omega f_n(r)f_m(r)\,dr.$$

Proof. Apply the definitions:

$$\lambda \in Z^{\perp} \Leftrightarrow \lambda^T z = 0 \quad \text{for all } z \in Z,$$

$$\Leftrightarrow \sum_n \lambda_n \int_\Omega f_n(r)p(r)\,dr = 0 \quad \text{for all } p \in L_1(\Omega),$$

$$\Leftrightarrow \int_\Omega p(r) \sum_n \lambda_n f_n(r)\,dr = 0 \quad \text{for all } p \in L_1(\Omega),$$

which means that $\lambda^T F = 0 \in L_\infty(\Omega)$, so $\|\lambda^T F\|_2 = 0$:

$$\int_\Omega \left[\sum_n \lambda_n f_n(r)\right]^2 dr = \lambda^T \Phi \lambda = 0.$$

Finally, observe that $\lambda^T\Phi\lambda = 0$ (if and) only if $\Phi\lambda = 0$, because Φ is symmetric semipositive definite. □

In many applications, $Z = \mathbb{R}^N$, which means that $Z^\perp = \{0\}$, i.e., the functions f_n are linearly independent. This is the case, for example, with the Fourier problem (1.12) ($\Phi = (2\pi)^d I$). On the other hand, this property is unlikely with image reconstruction of the type (1.11): in these applications, the subsets Ω_n usually have a high overlap.

The above difficulty concerning boundedness of λ is purely numerical. A more fundamental one is that (2.20), or at least (2.12), may not hold, even in real-life situations: it may even be that $z \notin Z$ because of measurement errors. In this case, one is bound to replace (1.13), (1.14) by a *penalized problem*

$$\text{(3.6)} \qquad \inf \int_\Omega h[p(r)]\,dr + \frac{1}{2T}\sum_n \left[\int_\Omega p(r) f_n(r)\,dr - z_n\right]^2$$

(T plays the role of a temperature; it is supposedly small). Although this is an unconstrained problem, we are somehow back in the situation (1.13), (1.14), with an optimization problem in $L_1(\Omega)$ instead of $\mathbb{R}^N$. Quite nicely, however, all the simplicity of the dual approach is still fully alive. To see it, observe that (3.6) is clearly equivalent to the apparently more complicated ($|\cdot|$ is the Euclidean norm in $\mathbb{R}^N$)

$$\begin{cases} \inf H(p) + \dfrac{1}{2T}|\zeta|^2, & p \in L_1(\Omega), \quad \zeta \in \mathbb{R}^N \\ \displaystyle\int_\Omega f_n(r)p(r)\,dr - z_n = \zeta_n, & n \in N, \end{cases}$$

whose Lagrangian is

$$\text{(3.7)} \qquad L^T(p, \zeta, \lambda) = L(p, \lambda) + \frac{1}{2T}|\zeta|^2 + \lambda^T\zeta.$$

Needless to say, (3.7) can be minimized separately with respect to p and ζ; the p-part is not affected by T: we again obtain p_λ of (2.17). Finally, the new dual function

$$\text{(3.8)} \qquad D^T(\lambda) := \min\{L^T(p, \zeta, \lambda) : p \in L_1(\Omega), \zeta \in \mathbb{R}^N\} = D(\lambda) - \tfrac{1}{2}T|\lambda|^2$$

has a unique maximum λ^T if T is positive.

In other words, the only (beneficial) affect of passing from (1.13), (1.14) to (3.6) is to simplify the dual problem: it has a (unique) solution without additional hypothesis, in particular, without (2.12), and it is better conditioned: penalization in the primal corresponds to regularization in the dual. As a result, (3.6) is totally "safe" and all our development can be reproduced with $T>0$. We summarize the important results.

THEOREM 3.6. *Under the standing assumptions, Assumptions* 1.2:

(i) *The first derivatives of* D^T *are*

$$\frac{\partial D^T(\lambda)}{\partial\lambda_n}=z_n-\int_\Omega p_\lambda(r)f_n(r)\,dr-T\lambda_n.$$

(ii) *If* $T>0$, *the penalized function of* (3.6) *has a unique minimum* p^T *and its minimal value* $H(p^T)$ *is just* $D^T(\lambda^T)$.

(iii) *When* $\lambda\to\lambda^T$, $p_\lambda\to p^T$ *in* $L_\infty(\Omega)$.

(iv) *Thus,* p^T *can be computed by some gradient-type maximization method applied to* D^T.

Just as in Remark 3.2, the second derivatives of D^T, when they exist, are

$$\frac{\partial^2 D^T(\lambda)}{\partial\lambda_n\partial\lambda_m}=-\int_\Omega k''[\lambda^TF(r)]f_n(r)f_m(r)\,dr-T\delta_{mn}.$$

Remark 3.7. An interesting by-product of the above development is the following: suppose that one insists on solving (3.6) explicitly (instead of simply maximizing D^T). Then the optimal p^T has a priori the form (2.17): it is not necessary to search the solution in the whole $L_1(\Omega)$, but only among those functions of the form (2.17). This allows a *parameterized* penalty technique: (3.6) is equivalent to minimizing with respect to λ

$$P^T:=H[k'(\lambda^TF)]+\frac{1}{2T}\sum_n\left\{\int_\Omega f_n(r)k'[\lambda^TF(r)]\,dr-z_n\right\}^2, \tag{3.9}$$

which depends only on $|N|$ parameters. This remark, which is apparently useless (after all, D^T is much nicer than P^T), will have its importance in the next section. Finally, observe that the optimality conditions for D^T are

$$z_n-\int_\Omega p_\lambda(r)f_n(r)\,dr=T\lambda_n,\qquad n\in N.$$

They are also the optimality conditions for P^T, since minimizing P^T is maximizing D^T.

We conclude this section with an example illustrating the role of the entropy: with $d=1$ and $\Omega=]0,1[$ discretized in 256 points, Fig. 1(a) represents a function p to be reconstructed (note the 11 peaks, and note that $0\leqq p\leqq 1$). Suppose that we just measure the mean value of p, together with its Fourier coefficients $1,2,\cdots,11$. Then we impose the constraints (1.12), (1.14) with $n=0,1,\cdots,11$. Naturally, computing the true p is a highly underdetermined problem; if, for example, we complete the Fourier series by 0 (setting $z_n=0$, $n>11$) and compute the resulting inverse Fourier transform, we obtain a very bad reconstruction, given in Fig. 1(b). In order to account for additional information, the idea is therefore to introduce some entropy and to solve (1.12)-(1.14) (compare Remark 1.1). This results in maximizing the dual function (3.8), which poses no particular difficulty; it is a concave smooth function of $1+2\times 11=23$ real variables. Figure 1(c) gives the reconstruction obtained with (1.6): only 10 peaks are obtained, and the upper bound $p\leqq 1$ is not respected. An excellent reconstruction (Fig. 1(d)) is obtained with the help of an entropy resembling (1.7), except that $h(0)=h(1)=+\infty$; it is defined by

$$2k'(s)=1-\frac{1}{\frac{1}{2}s}+\frac{1}{\operatorname{th}\frac{1}{2}s},\qquad k'(0)=\tfrac{1}{2}. \tag{3.10}$$

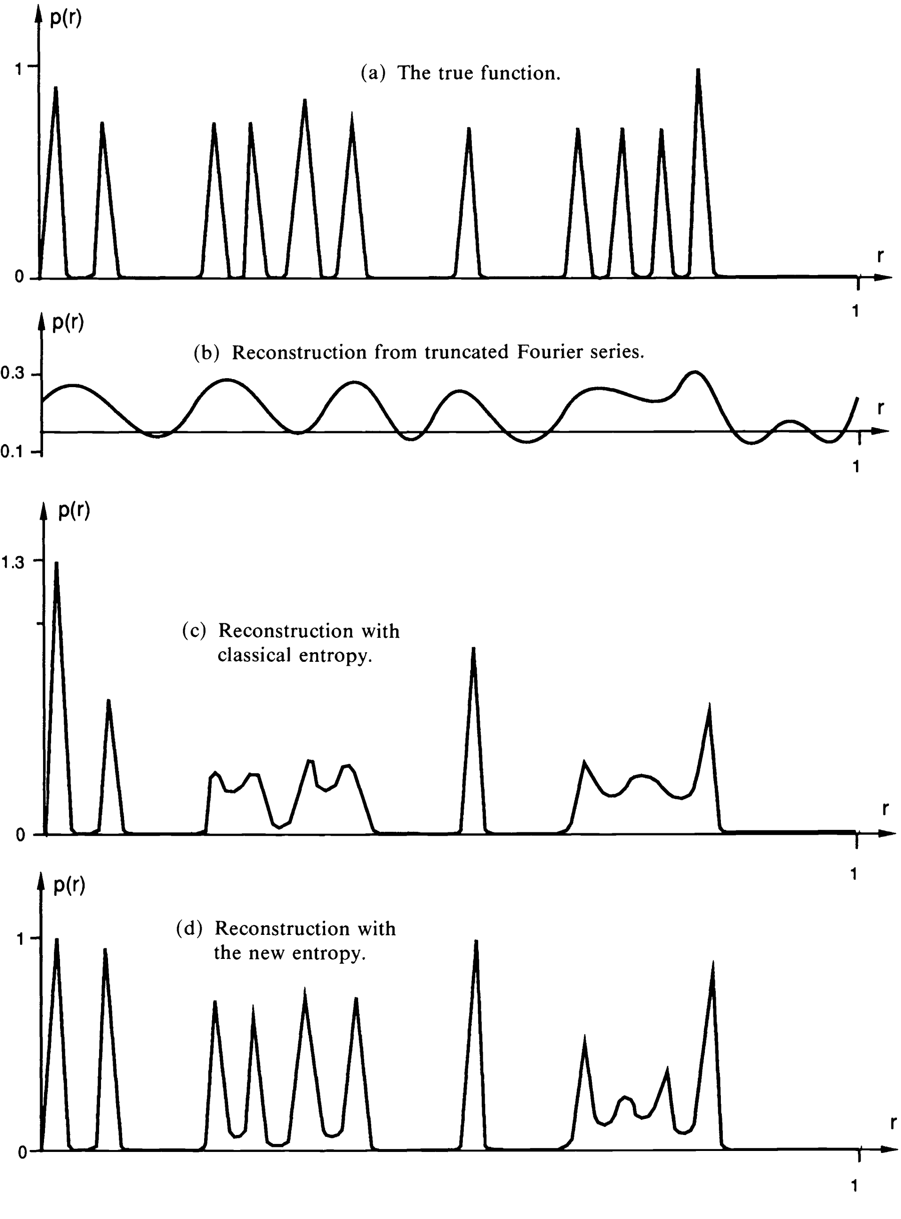


FIG. 1 *An illustrative example.*

Remark 3.8. Just as with (1.7), (2.7), the entropy associated with (3.10) is obtained from statistical considerations (see [23] again). The difference is essentially the behaviour of k' at infinity: here, $k'(s)$ converges to its limit with the speed $1/s$ (instead of $\exp(-s^2/2)$ for k_5'). It is interesting to observe that, when using such an approach, it is not h that comes out directly, but rather k'. Here, a primitive of (3.10) and the fundamental formulae (2.9), (2.10) give the corresponding entropy

$$h(t) := \frac{\frac{1}{2}s}{\operatorname{th}\frac{1}{2}s} + \operatorname{Log}\frac{\frac{1}{2}s}{\operatorname{sh}\frac{1}{2}s}, \tag{3.11}$$

where s is defined by $t = k'(s)$ of (3.10).

4. A nonlinearly constrained problem. The example above has illustrated the reconstruction of a function from its truncated Fourier series $\{z_n\}_{n\in N}$. An important case is when only some of these z_n are fully known, while the phases (in the sense of complex numbers) of the others are unknown. The physical origin of these problems is to locate the atoms of a given crystal: $p(r)$ is then the electron density at $r\in\Omega\subset\mathbb{R}^3$, the moduli of its Fourier coefficients being measured (up to some finite resolution) via X-ray diffraction. Interesting crystals are of organic nature, in which case the number of atoms per mesh is huge: 10^3 for a protein, 10^6 for a virus.

More precisely, our optimization problem (1.1) becomes

$$(4.1)\qquad \begin{cases} \inf \displaystyle\int_\Omega h[p(r)]\,dr, \\ \displaystyle\int_\Omega \cos(n\cdot r)p(r)\,dr = x_n; \quad \int_\Omega \sin(n\cdot r)p(r)\,dr = y_n, \qquad n\in N_0 \\ \left[\displaystyle\int_\Omega \cos(n\cdot r)p(r)\,dr\right]^2 + \left[\displaystyle\int_\Omega \sin(n\cdot r)p(r)\,dr\right]^2 = m_n^2, \qquad n\in M. \end{cases}$$

Here, N_0 and M form a partition of N; N_0 is the set of fully known Fourier coefficients $z=(x,y)$, M the set with unknown phases (note: $0\in N_0$, which fixes the mean value x_0 of the density p; consistent data have $y_0=0$; in extreme cases, $N_0=\{0\}$, i.e., no phase is known at all). The given moduli m_n, $n\in M$, are of course positive (otherwise, the corresponding index would be in N_0!). As for h, it may, for example, have the form (1.6) (which automatically ensures $p\geqq 0$), or (1.7) or (3.11) (from which also follows $p\leqq 1$).

We introduce a simplifying notation, based on complex calculus: we condense each pair of linear constraints into one single, complex-valued constraint, and we denote by $F_n(p)$ the nth Fourier coefficient of p:

$$F_n(p) := \int_\Omega e^{inr}\cdot p(r)\cdot dr \in \mathbb{C}$$

(to reduce the risk of confusion, a dot "$\cdot$" denotes the usual product of complex numbers, even if one of the numbers, like $p(r)$, is real). Furthermore, for $z_n = x_n + i\cdot y_n \in \mathbb{C}$, $n=1,2$, we denote

$$z_1 * z_2 := \operatorname{Re}(z_1\cdot \bar{z}_2) = \tfrac{1}{2}(z_1\cdot \bar{z}_2 + \bar{z}_1\cdot z_2) = x_1x_2 + y_1y_2,$$

the usual Euclidean product of z_1 and z_2, considered as vectors in $\mathbb{R}^2$. Note that its associated norm is the usual modulus of complex numbers. With these notations, (4.1) can be written

$$(4.2)\qquad \begin{cases} \inf H(p), \\ F_n(p) = z_n, \qquad n\in N_0, \\ |F_n(p)| = m_n, \qquad n\in M. \end{cases}$$

THEOREM 4.1. *Under the standing assumptions, Assumptions* 1.2, *there exists an optimal solution to* (4.1) *whenever its constraints are consistent.*

Proof. Consider a sublevel-set

$$(4.3)\qquad \{p\in L_1(\Omega):\ H(p)\leqq \alpha\}.$$

Convexity and coercivity imply that h is bounded below, say, by C. Then, for p in the set (4.3) and $\Omega' \subset \Omega$, we have

$$\begin{aligned}\int_{\Omega'} h[p(r)]\,dr &= \int_{\Omega} h[p(r)]\,dr - \int_{\Omega\backslash\Omega'} h[p(r)]\,dr\\ &\leqq \alpha - C \text{ meas } \Omega\backslash\Omega'\\ &\leqq \alpha + |C| \text{ meas } \Omega =: C'.\end{aligned}$$

On the other hand, let $\varepsilon > 0$ be arbitrary. We know from coercivity that there exists $\bar{t} > 0$ such that

$$|t| \leqq \frac{\varepsilon}{C'} h(t) \quad \text{for } |t| \geqq \bar{t}.$$

Thus, setting $\Omega' = \{r \in \Omega \colon |p(r)| \geqq \bar{t}\}$, we put $t = p(r)$ and deduce by integration over Ω' that

$$\int_{\Omega'} |p(r)|\,dr \leqq \frac{\varepsilon}{C'} \int_{\Omega'} h[p(r)]\,dr \leqq \varepsilon.$$

This is the so-called equi-integrability, which, together with the weak lower semicontinuity of the convex H, is equivalent to the weak compactness of the set (4.3); see [11, Chap. VIII, Thm. 1.3]. Then the end of the proof is standard: exploiting the weak continuity of the constraint functions of (4.1), we see that any weak cluster-point of a minimizing sequence is an optimal solution. □

To compute a solution, we want to use §§ 2 and 3, which allowed such an efficient numerical approach. If we dualize the nonlinear constraints of (4.1), however, there is a double difficulty: (i) the resulting Lagrange function cannot be easily minimized with respect to p and may have several minima, which implies that (ii) usually, no value of the dual variables gives back an optimal p (or even feasible). For a convenient use of §§ 2 and 3, we consider in (4.2) that the unknown Fourier coefficients $\{F_n(p)\}_{n\in M}$ are additional optimization variables. Clearly, (4.1) = (4.2) is equivalent to

$$\text{(4.4)} \qquad \begin{cases} \inf H(p), \\ F_n(p) = z_n, & n \in N = N_0 \cup M, \\ |z_n| = m_n, & n \in M, \end{cases}$$

in which the optimization variables are now: $p \in L_1(\Omega)$ and $z \in \mathbb{C}^M$ (remember that z_n is known and fixed for $n \in N_0$).

A possible Lagrange function associated with (4.4) is

$$H(p) - \sum_{n\in N} \lambda_n * (F_n(p) - z_n) + \tfrac{1}{2} \sum_{n \in M} \mu_n(|z_n|^2 - m_n^2),$$

which exhibits the following formal optimality conditions: there exist $\lambda \in \mathbb{C}^N$ and $\mu \in \mathbb{R}^M$ such that

$$\text{(4.5)} \qquad p(r) = p_\lambda(r) := k'\left[\sum_{n\in N} \lambda_n * e^{inr}\right] \quad \text{(stationarity in } p),$$

$$\text{(4.6)} \qquad \lambda_n = \mu_n \cdot F_n(p), \quad n \in M \quad \text{(stationarity in } z).$$

Needless to say, (4.5), (4.6) is neither necessary nor sufficient for p to solve (4.1) = (4.2): not even mentioning the nonlinearity of the constraints, the need for (2.20) is still present. On the other hand, (4.5) deserves attention, as it opens the way toward searching an optimum in the parameterized form (2.17) (compare Remark 3.7).

PROPOSITION 4.2. *Suppose that* (4.1) = (4.2) *has an optimal solution* $\bar{p}$. *Suppose, in addition, that* $m < \bar{p}(r) < M$ *for almost every* $r \in \Omega$. *Then there exists* $\lambda \in \mathbb{C}^N$ *such that* $\bar{p}$ *has the form* (4.5).

Proof. For $n \in N$, call $\bar{z}_n := F_n(\bar{p})$; then $\bar{p}$ is clearly a solution of (1.13), (1.14) with the right-hand side $\bar{z}$. Because the Slater assumption (2.20) is obviously satisfied (by $\tilde{p} = \bar{p}$!), the claim is just Lemma 2.6. □

Our trick (4.2) → (4.4) somehow splits the original problem into two parts: one, which is nice and convex, concerns p alone expressed by (4.5); all the nasty nonlinearities are packed into the second part, concerning z, in which p appears rather as a parameter. The situation is analogous to that of § 3: with an additional assumption of Slater type, our problem is posed in the λ-space $\mathbb{C}^N$ instead of the p-space $L_1(\Omega)$. Naturally, the λ-problem is now much more involved: we must not only have feasibility of the function (4.5), but also its Fourier coefficients must be aligned (mod π) with λ if (4.6) is to hold.

Remark 4.3. The additional Slater-type assumption needed for Proposition 4.2 can hardly be tolerated: it is an assumption that must be satisfied a posteriori by the optimal solution, and there is no way of checking it in advance. Note, incidentally, that it can be slightly relaxed to the following—hardly more tolerable:

At the optimal $\bar{p}$, $\bar{z} := F(\bar{p})$ satisfies (2.20) (possibly with $\tilde{p} \neq \bar{p}$).

On the other hand, the problem is limited to trigonometric constraint functions f_n; for this special class, (2.20) turns out to follow automatically from (2.12) in some cases; let us cite:

* when m and M are infinite (obvious);

* when $h(m) = h(M) = +\infty$ (obvious as well: indeed, any p with $H(p) < +\infty$ satisfies the Slater assumption);

* when m or M is finite but not both: this is [5, Thm. 2.9];

* when m and M are both finite but, a posteriori, there is some other function having the same Fourier coefficients as the optimal p; this can be proved by extending the proof of [5, Thm. 2.9].

The key issue underlying this question is the following: "Is it true that the set of functions lying between two finite bounds and having given Fourier coefficients is never a singleton?" If the answer is "yes," then a strictly feasible solution can be proved to exist whenever a merely feasible solution exists.

Partly because of the above technical difficulty, but mainly for practical reasons (the data z_n and m_n are highly noisy), we now turn to a penalized form of the problem, just as was done at the end of § 3. Choosing the formulation (4.4), we penalize only the linear constraints, explicitly keeping the nonlinear ones: we now solve for $p \in L_1(\Omega)$ and $z \in \mathbb{C}^M$ (z_n being known for $n \in N_0$):

$$(4.7) \qquad \begin{cases} \inf H(p) + \dfrac{1}{2T} \sum_{n \in N} |F_n(p) - z_n|^2, \\ |z_n| = m_n, \qquad n \in M. \end{cases}$$

THEOREM 4.4. *Under the standing assumptions, Assumptions* 1.2, (4.7) *has an optimal solution for any* $T > 0$. *The p-part of the solution set of* (4.7) *is also the set of minima of the function*

$$(4.8) \qquad H(p) + \frac{1}{2T} \sum_{n \in N_0} |F_n(p) - z_n|^2 + \frac{1}{2T} \sum_{n \in M} \{|F_n(p)| - m_n\}^2.$$

Proof. The existence of an optimal solution to (4.7) is exactly as in Theorem 4.1. Now, (4.7) can equivalently be solved in a hierarchical way with respect to z first, and then with respect to p. For fixed $F_n(p)$, the objective of (4.7) is minimized on $\mathbb{C}^M$ by

$$z_n = \begin{cases} m_n \cdot F_n(p)/|F_n(p)| & \text{if } F_n(p) \neq 0, \\ \text{arbitrary of modulus } m_n \text{ otherwise.} \end{cases}$$

Plugging these values back into (4.7) for $n \in M = N \backslash N_0$, we just obtain the function (4.8), whose minimization is therefore equivalent to the resolution of (4.7). □

Thus, (4.7) is nothing but the standard penalized formulation of (4.1) = (4.2); let us now address its numerical resolution. Among the possibilities, there is a hierarchical approach opposite to the one suggested by the above proof, namely, to minimize with respect to p first, a convex decomposable problem fully resorting to the dual machinery of §§ 2 and 3.

THEOREM 4.5. *Under the standing assumptions, Assumptions* 1.2, *let* $T > 0$ *and use the notation* (4.5). *The optimization problem in* λ

$$(4.9) \qquad P^T(z) := \max_{\lambda \in \mathbb{C}^N} \sum_{n \in N} \left(\lambda_n * z_n - \frac{T}{2} |\lambda_n|^2 \right) - \int_\Omega k \left[\sum_{n \in N} \lambda_n * e^{inr} \right] dr$$

has a unique solution $\lambda^T(z)$. *The maximal value* $P^T(z)$ *is convex and differentiable in* z; *its gradient is given by the formula*

$$(4.10) \qquad dP^T(z) = \sum_{n \in M} \lambda_n^T(z) * dz_n.$$

The optimization problem (4.7) *is equivalent to the problem in* $z \in \mathbb{C}^M$

$$(4.11) \qquad \min \{P^T(z) \colon |z_n| = m_n,\ n \in M\}.$$

Proof. Realize from (2.18) that the maximand in (4.9) is just the dual function D^T of (3.8). Then apply Theorem 3.6: the optimization problem in (4.9) has a unique solution $\lambda^T(z)$ and the function $P^T(z)$ is nothing but the minimal value of (4.7) with respect to p. It is convex because it is the conjugate of a convex function. Its differentiability and (4.10) follow again from [8] (differentiate the maximand in (4.9); remember Remarks 2.1 and 2.5).

Since $P^T(z)$ is just the result of a partial optimization with respect to p in (4.7), the equivalence (4.7)$\Leftrightarrow$(4.11) becomes clear. □

Let us summarize our results: the solutions of the penalized form (4.8) of (4.1) are obtained by (4.5), where λ solves (4.9) at those z solving (4.11); to compute such a z, a gradient-type minimization method can be used based on (4.10). Beware that the λ-problem (4.9) is always posed in the whole $\mathbb{C}^N$, while the z-problem (4.11) is posed in the smaller $\mathbb{C}^M$, the set of "incomplete" Fourier data. Needless to say, (4.11) has an optimal solution (the feasible domain is compact!) that is hard to find, in view of the nonconvex constraints; but at least the difficulty is now concentrated in these constraints—no trouble comes from the objective P^T.

The optimality conditions for (4.9) have been seen in Remark 3.7:

$$(4.12) \qquad T\lambda_n^T(z) = z_n - F_n[p_{\lambda^T(z)}] \quad \text{for } n \in N.$$

On the other hand, the optimality conditions obviously hold at an optimum of (4.11). Use (4.10) and draw a picture in the complex plane to realize that they are:

$$\lambda_n^T(z) \text{ and } z_n \text{ are aligned } (\text{mod} \cdot \pi) \quad \text{for } n \in M;$$

in view of (4.12), this means also that $\lambda_n^T(z)$ and $F_n[p_{\lambda^T(z)}]$ are aligned (mod · π); we recognize the original condition (4.6).

Remark 4.6. As part of the elegance of this approach, observe that Theorem 4.5 somehow unifies (4.4) and (4.7) = (4.11); it suffices to set $T=0$ in (4.9) or (4.12). This case, however, introduces the technical difficulties alluded to in Remark 4.3: P^0 may be $+\infty$, or there may be no optimal λ, or there may be several; then, regularity of P^0 may disappear.

5. Conclusion. We have studied in this paper two classes of problems dealing with entropy. The "easy" class has linear constraints only, and is most efficiently solved by a dual algorithm. The "difficult" class of § 4 contains nonconvex constraints, for which we have proposed an approach keeping as close as possible to the convex situation. This approach amounts to solving a minimax problem in $\mathbb{C}^M \times \mathbb{C}^N$:

$$\min_{|z_n|=m_n} \max_{\lambda} D^T(z, \lambda) \tag{5.1}$$

where D^T, the maximand in (4.9), is nothing but the regularized dual function of (2.18), (3.8). The original unknown variable p is searched in a parameterized form, as a function of λ coming out from (5.1).

Remark 5.1. The inherent difficulty of the problem, namely, the nonconvexity of its constraints, is conserved as such in (5.1): we have not made the least step toward global optimization. If the z-domain of (5.1) were convex, things would become "easy" (D^T is convex-concave).

We mention a problem studied in [28] which, in our notations, is

$$\min \{H(p): F_n(p) = z_n, |z_n - \zeta_n|_\infty \leqq m_n, n \in N\}$$

(the $\zeta_n \in \mathbb{C}$ and m_n are all given). Our approach is particularly well suited here as well: we end up with a formulation (5.1) having the simple constraints $|z_n - \zeta_n|_\infty \leqq m_n$. Furthermore, $L_1(\Omega)$ is discretized a priori in [28], in such a way that the internal λ-problem has an explicit solution; see Remark 5.4 below.

The z-constraints of (5.1) are most conveniently handled via polar coordinates: set

$$z_n = z_n(\varphi) = m_n \cdot e^{i \cdot \varphi_n} \quad \text{for } n \in M$$

and minimize P^T of (4.11) with respect to the unconstrained real variables φ_n. Elementary calculus gives the gradient of this new function $Q^T(\varphi) := P^T[z(\varphi)]$: from the relations

$$dz_n = i \cdot z_n \cdot d\varphi_n \quad \text{and} \quad dQ^T = dP^T = \sum_{n \in M} \lambda_n^T * dz_n,$$

we obtain (setting $\lambda_n = \alpha_n + i \cdot \beta_n$ and $z_n = x_n + i \cdot y_n$)

$$\frac{\partial Q^T}{\partial \varphi_n} = \beta_n x_n - \alpha_n y_n.$$

Observe that $\nabla Q^T = 0$ when, once again, λ_n and z_n are aligned (mod $\cdot\ \pi$). With these notations, the following algorithmic scheme suggests itself.

ALGORITHM 5.2. Choose an initial $\varphi \in \mathbb{R}^M$.

Step 1. Apply, for example, Algorithm 3.1 to compute λ maximizing D^T of (2.18), (3.8), given that $z_n = m_n \cdot e^{i \cdot \varphi_n}$.

Step 2. Stop if φ is approximately optimal, i.e., if for each $n \in M$

$$\text{phase}\,(\lambda_n) \simeq \varphi_n \quad (\text{mod } \pi).$$

Step 3. With the help of some gradient-type optimization algorithm, select a new $\varphi \in \mathbb{R}^M$ aimed at decreasing the resulting D^T and loop to Step 1.

By construction, each iteration constructs $p \in L_\infty(\Omega)$ of the form (4.5); Step 1 solves (4.12) and Step 3 aims at solving (4.6) asymptotically. The algorithm can be used with $T=0$ (see Remark 4.6), at the expense of some complications. First, the various iterates of Step 3 might well generate unbounded λ and/or $D=D^0$ in Step 1. Actually, a penalty approach is then recommended for the constraints of (5.1). A major difficulty is that (4.5), (4.6) may not hold at a solution of (4.2). In these conditions, little can be said concerning the overall convergence, as long as the argument raised in Remark 4.3 is not fixed.

Remark 5.3. An interesting question is the inversion of the min- and max-operation in (5.1): consider

$$\max_{\lambda} \min_{|z_n|=m_n} D^T(z, \lambda). \tag{5.2}$$

The internal z-minimization can be worked out explicitly, and (5.2) is just the minimization of the convex function of λ

$$\sum_{n\in M} m_n|\lambda_n| - \sum_{n\in N_0} z_n * \lambda_n + \frac{T}{2}\sum_{n\in N}|\lambda_n|^2 + \int_\Omega k\left[\sum_{n\in N}\lambda_n * e^{inr}\right] dr. \tag{5.3}$$

Now, because D^T is linear in z, (5.2) = (5.3) is also equivalent to

$$\max_{\lambda} \min_{|z_n|\leqq m_n} D^T(z, \lambda), \tag{5.4}$$

and the theory of saddle functions (see, for example, [32]) tells us that (5.4) is in turn equivalent to

$$\min_{|z_n|\leqq m_n} \max_{\lambda} D^T(z, \lambda),$$

the convexification of (5.1). Admitting that this latter problem is close to (5.1), it, or its equivalent form (5.3), can be used to initialize Algorithm 5.2.

This technique has a limited value, however: take the most unfavourable situation, with no phase known a priori, $N_0=\{0\}$. Then, it is not difficult to see that the optimal λ in (5.3) is $\lambda_n=0$ for $n\neq 0$ (observe that the corresponding p_λ is a constant function, whose Fourier coefficients are all zero; from (2.19), we know that these Fourier coefficients are precisely the derivatives $\partial K/\partial\lambda$). Unfortunately, $\lambda=0$ is just the point for which (5.4) gives exactly *no information* on z.

An important drawback of Algorithm 5.2 is the need to maximize D^T for each iterate φ. This means that Algorithm 3.1, which acts as an internal subroutine, has to be extremely fast and reliable. When $|N|$ is small enough, Newton's method is suggested and allows an exact computation of $Q^T=P^T$; but this method may become impossible to implement for larger $|N|$: to begin with, storage problems may appear with the matrix $\partial^2 D^T/\partial\lambda^2$. So far, we have not really experimented with any alternative.

Remark 5.4. For really large $|N|$, Algorithm 3.1 must be abandoned, but one can exploit the fact that trigonometric functions are orthonormal. Indeed, consider (4.2): it is reasonable to assume that N (very large) contains all the observable $|F_n(p)|$; in other words, we can reasonably fix $F_n(p)=0$ for $n\notin N$. Then, there is only one p satisfying the constraints (1.14), explicitly given by inverse Fourier transform. Based on this idea, we explain schematically how to compute $\lambda^T(z)$, even for $T>0$.

The Fourier series defines a linear operator A mapping $L_2(\Omega)$ onto a subset of $\mathbb{C}^{\mathbb{Z}^d}$. This A is invertible and A^{-1}, the inverse Fourier transform, is its adjoint A^*. With this notation, p_λ of (4.5) is nothing but the function $k'(A^*\lambda)\in L_2(\Omega)$; the λ-problem (4.12) = (4.9) is

$$Ak'(A^*\lambda)+T\lambda=z$$

(z is given: the z_n are data if $n \in N_0$, the current outer iterate if $n \in M$, and 0 otherwise). Taking the inverse Fourier transform:

$$k'(A^*\lambda) + TA^*\lambda = A^*z.$$

To compute $A^*\lambda$, we have to solve pointwise the equation in t

$$k'(t) + Tt = (A^*z)(r);$$

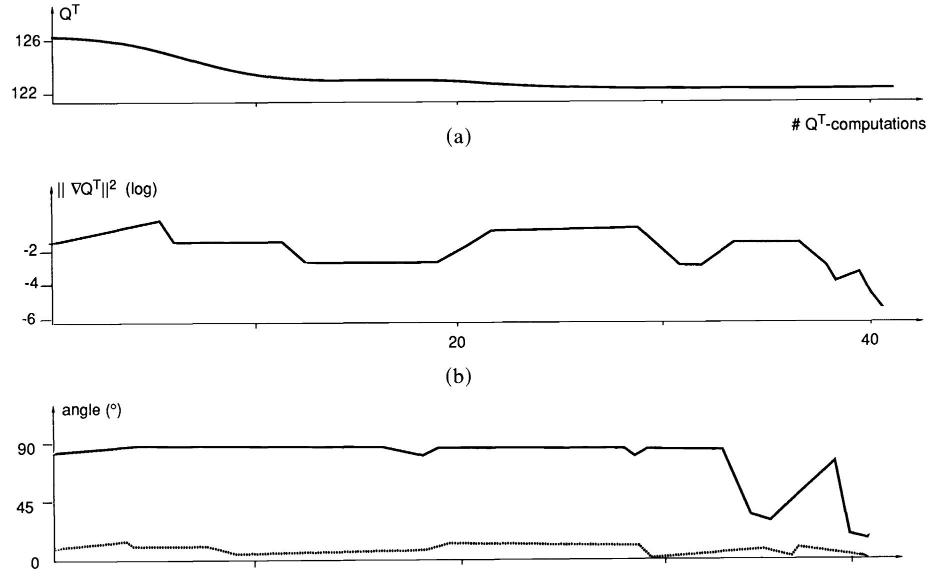
calling $q(r)$ the result, $\lambda^T(z)$ is the Fourier series of q and we are done (in view of (4.10), λ_n needs to be computed for $n \in M$ only). Note that this trick leaves the z-problem unchanged, in fact, M is still the same and the whole business just amounts to replacing N_0 by $\mathbb{N}^d \backslash M$.

For a numerical illustration, we have considered a molecule called prostaglandin, having 25 atoms. In a first experiment, we took a data-set with $|N_0| = 4$, $|M| = 185$; taking symmetry into account, the resulting optimization variables were $\lambda \in \mathbb{R}^{377}$ and $\varphi \in \mathbb{R}^{185}$. The internal λ-problem was solved by Newton's method, which gave (4.12) to within machine accuracy; then, computing the value and gradient of Q^T essentially amounted to two Fourier transforms (with $\Omega \subset \mathbb{R}^3$ discretized in $34 \times 38 \times 20 = 28424$ points); the Hessian $\nabla^2 Q^T(\varphi)$ could be computed with one more Fourier transform and the inversion of a 377×377 matrix.

Three methods were tested in Step 3 of Algorithm 5.2: the variable-memory quasi-Newton method of [25], [14], the truncated Newton method of [9], and the method of Newton with trust region [22]. Not unexpectedly, the latter two methods were substantially better, both in terms of computing time and accuracy. Roughly speaking, it can be said that quasi-Newton needed some 200 computations of $(D^T, \nabla D^T)$, and a computing time perceivably longer than the other two methods. These latter methods obtained a much better accuracy in some 20 computations of $\nabla^2 Q^T$ (by far more expensive than the corresponding 40 computations of function-gradient). A detail is worth mentioning concerning truncated Newton: initializing the stepsize for the line-search is not a straightforward task; we chose a weighted combination of 1 (suitable when Newton's equation is solved accurately) and "Fletcher's value" $-2\Delta Q/\nabla Q^T d$ (always not bad, here ΔQ is the progress in the previous iteration, d the direction).

The behaviour of truncated Newton is illustrated in Fig. 2, as a function of the number of D^T-maximizations. Figure 2(a) displays Q^T. Observe how little it varies: D^T is very flat indeed (the starting phases were randomly selected, hence supposedly far from the optimum fund). Figure 2(b) gives the corresponding evolution of $\|\nabla Q^T\|$; it is interesting to mention that the iteration where it starts really decreasing is also the iteration where $\nabla^2 Q^T$ starts having all its eigenvalues positive. Another measure of $\|\nabla Q^T\|$ is given by (4.6): call $\alpha_n \in [0°, 90°]$ the angle between the two complex numbers λ_n and $F_n(p)$; φ is stationary when all α_n are 0. Figure 2(c) displays the max and average of these angles along the iterations. With trust region, the results were better but qualitatively the same. All these results were obtained on a SUN/4 computer using single precision (about six digits).

One more detail: as predicted by Theorem 2.4 in [9] and Theorem 4.17 in [22], truncated Newton and trust region consistently converged to true local minima. In our example, Fig. 2, the eigenvalues of $\nabla^2 Q^T$ after convergence of truncated Newton were regularly spread in $[1, 10^4]$; quasi-Newton produced three eigenvalues smaller than 1, including one at -3. Thus, we have here an example in which a descent method using first derivatives only does not converge to a local minimum; such examples are not easy to construct.



(a)

(b)

(c)

FIG. 2. *Solving the phase problem by truncated Newton method.*

To give an idea of the primal aspect, Fig. 3 displays contours of the function $p(\cdot, \cdot, r_3)$, plotted for a number of equidistant values of r_3. The picture seems to go out from its frame because prostaglandin crystallizes in a nonorthogonal unit cell. This reconstruction needed a data-set with $|N_0| = 27$ and $|M| = 303$.

Finally we mention that, depending on the data-set, such reconstructions were often mediocre, having little to do with an actual electron density, and this was invariably the case when N_0 was a small set. For example, consider a data-set with $N_0 = \{0\}$, $|M| = 78$ and suppose that the 78 unknown phases are those of the exact solution. Then $\nabla Q^T \neq 0$ and, more importantly, the Hessian $\nabla^2 Q^T$ is indefinite: its spectrum is regularly spread in $[-10^2, 10^4]$, with 13 negative eigenvalues: these exact phases are far from a minimum. Worse, when initialized on these exact phases,

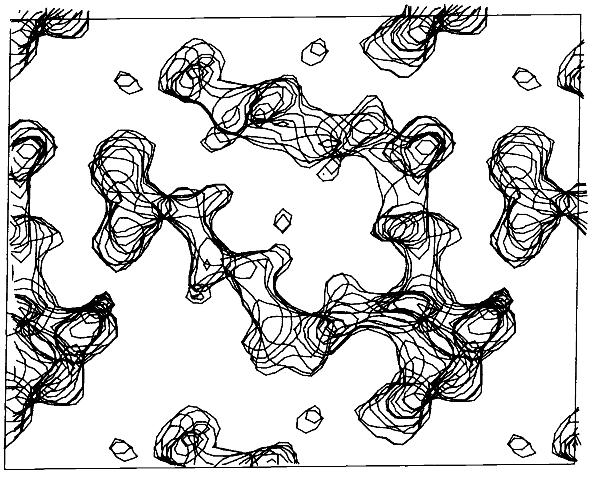


FIG. 3. *A primal reconstruction.*

Algorithm 5.2 converges to the reconstruction of Fig. 4, from which no chemist could extract any useful information. Nevertheless, take a small data-set with $|N| = 20$, but $N_0 = N$ ($M = \emptyset$, we are actually in the linear framework of § 3); then Algorithm 3.1 produces the reconstruction of Fig. 5, which points out the overall shape of the molecule, thanks to the connectivity of the electron density. Conclusion: it is fair to say that the entropy approach has limited efficiency, in the sense that it requires a good amount of information, especially concerning the phases. Other models are wanted when this information is not available.

Remark 5.5. Let us mention that more information can orient the search for a solution without making the problem easier. For example, it is known that the unknown p is an atomic distribution (a sum of Dirac measures); it is, however, difficult to take this into account. In particular, the entropy-based model is somehow self-contradictory: if the observations were neglected, it would give an absolutely flat p.

Another such item of information is symmetry, which helps to reduce the number of variables. For example, if p is even, i.e., $p(-r) = p(r)$, then its Fourier coefficients

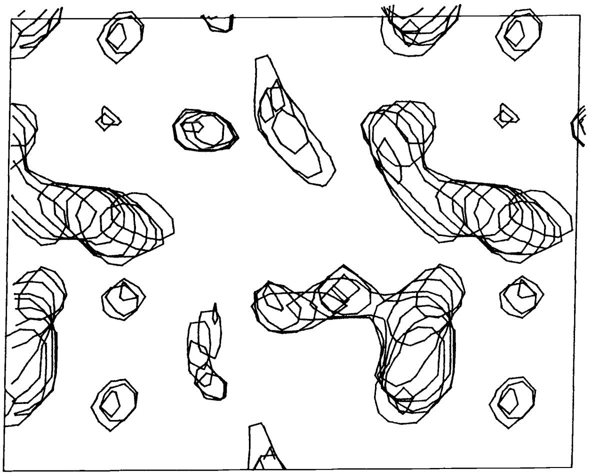


FIG. 4. *Entropy maximization can ruin a good starting point.*

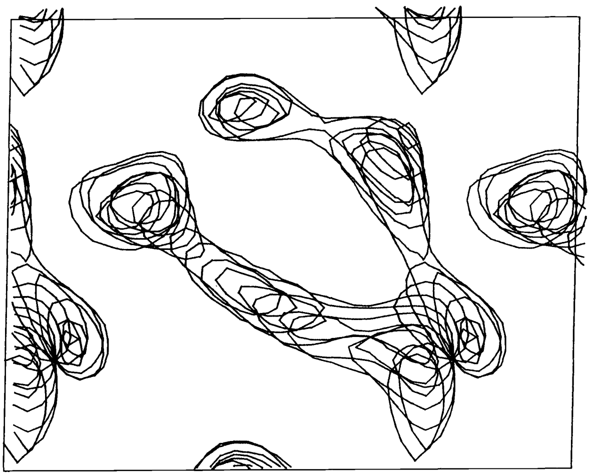


FIG. 5. *The importance of phase information.*

must be real, i.e., their phases are 0 or π. The phase-problem becomes a 0-1 programming problem (highly nonlinear); needless to say, it is strictly intractable for standard methods of combinatorial optimization.

Acknowledgment. We are indebted to anonymous referees for some very valuable suggestions, criticisms, and bibliographical references.

REFERENCES

[1] A. Ben Tal, J. M. Borwein, and M. Teboulle, *Spectral estimation via convex programming*, in Proc. Internat. Conference on Extremal Methods, University of Texas, Austin, TX, 1987.

[2] ———, *A dual approach to multidimensional L_p spectral estimation problems*, SIAM J. Control Optim., 26 (1988), pp. 985-996.

[3] M. F. Bidaut, *Un problème de contrôle optimal à fonction-coût en norme L_1*, Comptes Rendus Acad. Sci. Paris, A-28 (1975), pp. 273-276.

[4] J. M. Borwein and A. S. Lewis, *Partially finite convex programming*, Math. Programming, to appear.

[5] ———, *Duality relationships for entropy-like minimization problems*, SIAM J. Control Optim., 29 (1991), pp. 325-338.

[6] A. A. Buckley and A. Lenir, *QN-like variable storage conjugate gradients*, Math. Programming, 15 (1978), pp. 200-210.

[7] D. Dacunha-Castelle and M. Duflo, *Probabilités et Statistiques*, Masson, Paris, 1982.

[8] J. M. Danskin, *The theory of maxmin with applications*, SIAM J. Appl. Math., 4 (1966), pp. 641-655.

[9] R. S. Dembo and T. Steihaug, *Truncated-Newton algorithms for large-scale unconstrained optimization*, Math. Programming, 26 (1983), pp. 190-212.

[10] J. Dieudonné, *Eléments d'Analyse, tome* 2, Gauthier-Villars, Paris, 1968.

[11] I. Ekeland and R. Temam,*Convex Analysis and Variational Problems*, North-Holland, Amsterdam, 1976.

[12] J. Eriksson, *A note on solution of large sparse maximum entropy problems with linear equality constraints*, Math. Programming, 18 (1980), pp. 146-154.

[13] A. M. Geoffrion, *Duality in nonlinear programming: A simplified application-oriented development*, SIAM Rev., 13 (1971), pp. 1-37.

[14] J. C. Gilbert and C. Lemaréchal, *Some numerical experiments with variable-storage quasi-Newton algorithms*, Math. Programming B, 45 (1989), pp. 407-435.

[15] S. F. Gull and G. J. Daniell, *Image reconstruction from incomplete and noisy data*, Nature, 272 (1978), pp. 686-690.

[16] R. Gordan and G. T. Herman, *Three-dimensional reconstruction from projections: A review of algorithms*, Internat. Rev. Cytology, 38 (1974), pp. 111-151.

[17] R. C. Grinold, *Lagrangian subgradients*, Management Sci., 17 (1970), pp. 185-188.

[18] H. Hauptman and J. Carle, *Solution of the phase-problem.* I: *The centrosymmetric crystal*, American Crystallographic Association, Monograph 3, Polycrystal Book Service, Pittsburgh, Pennsylvania, 1953.

[19] E. T. Jaynes, *Information theory and statistical mechanics*, Phys. Rev., 106 (1957), pp. 620-630.

[20] A. Klug, *Joint probability distributions of structure factors and the phase-problem*, Acta Crystallographica, 11 (1958), pp. 515-543.

[21] D. Medhi, *Decomposition of structured large-scale optimization problems and parallel optimization*, Ph.D. thesis, University of Wisconsin, Madison, WI, 1987.

[22] J. Moré, *Recent developments in algorithms and software for trust region methods*, in Mathematical Programming, the State of the Art, A Bachem, M. Grötschel, B. Korte, eds., Springer-Verlag, Heidelberg, 1983, pp. 258-287.

[23] J. Navaza, *The use of nonlocal constraints in maximum-entropy electron density reconstruction*, Acta Crystallographica, A42 (1986), pp. 212-223.

[24] R. Nityananda and R. Narayan, *Maximum entropy image reconstruction—a practical noninformation-theoretic approach*, J. Astrophys. Astron., 3 (1982), pp. 419-450.

[25] J. Nocedal, *Updating quasi-Newton matrices with limited storage*, Math. Comp., 35 (1980), pp. 773-782.

[26] B. T. Poljak, *Iterative algorithms for singular minimization problems*, in Nonlinear Programming 4, O. L. Mangasarian, R. R. Meyer, S. M. Robinson, eds., Academic Press, New York, 1981, pp. 147-166.

[27] M. J. D. Powell, *Some global convergence properties of a variable metric algorithm for minimization without exact line-searches*, in Nonlinear Programming, R. W. Cottle, C. E. Lemke, eds., SIAM-AMS Proceedings IX, Society for Industrial and Applied Mathematics, Philadelphia, PA, 1976.

[28] M. J. D. POWELL, *An algorithm for maximizing entropy subject to simple bounds*, Math. Programming, 42 (1988), pp. 171-180.

[29] S. M. ROBINSON, *Bundle-based decomposition: Conditions for convergence*, Analyse non linéaire, 6 (1989), pp. 435-448.

[30] R. T. Rockafellar, *Integrals which are convex functionals*, Pacific J. Math., 24 (1968), pp. 525-539.

[31] ———, *Integrals which are convex functionals*, II, Pacific J. Math., 39 (1971), pp. 439-469.

[32] ———, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.

[33] H. UZAWA, *Iterative methods for concave programming*, in Studies in Linear and Nonlinear Programming, K. J. Arrow, L. Hurwicz, H. Uzawa, eds., Stanford University Press, Stanford, CA, 1972.

[34] P. WOLFE, *A duality theorem for nonlinear programming*, Quart. Appl. Math., 19 (1961), pp. 239-244.

SIAM J. OPTIMIZATION
Vol. 2, No. 2, pp. 198-209, May 1992

A LOW COMPLEXITY INTERIOR-POINT ALGORITHM FOR LINEAR PROGRAMMING*

MICHAEL J. TODD†

Abstract. This paper describes an interior-point algorithm for linear programming that is almost as simple as the affine-scaling method and yet achieves the currently best complexity of $O(\sqrt{n}\, t)$ iterations to attain precision t. The basic algorithm needs neither dual estimates nor lower bounds, although its analysis is based on Ye's results for the primal-dual potential function. Some computationally preferable variants are also presented.

Key words. linear programming, interior-point methods

AMS(MOS) subject classification. 90C05

1. Introduction. The polynomial-time interior-point algorithms that have been developed in the past few years can be roughly classified as follows:

(i) projective-scaling algorithms, stemming from Karmarkar's original method [16], such as those of Anstreicher [1], de Ghellinck and Vial [9], and Todd and Burrell [28];

(ii) path-following methods, which attempt to follow closely the central path studied by Megiddo [20] and Bayer and Lagarias [6], such as the dual algorithm of Renegar [26], the primal algorithm of Gonzaga [11], and the primal-dual algorithms of Kojima, Mizuno, and Yoshise [17], [18] and Monteiro and Adler [24], [25]; and

(iii) potential-reduction algorithms, such as those of Gonzaga [12], Ye [31], Freund [8], Kojima, Mizuno, and Yoshise [19], Gonzaga [13], and Anstreicher [2], [3].

While the derivations of these methods follow very different lines, the search directions employed are invariably linear combinations of two directions: the affine-scaling direction and the centering direction, which try, respectively, to improve the objective function and to drive the current iterate towards the analytic center (Sonnevend [27]) of the feasible region. This property of the search direction was noted in several papers: Yamashita [30], Gonzaga [10], Mitchell and Todd [22], Zimmerman [32], and the recent survey of den Hertog and Roos [15]. (In case an algorithm does not require a feasible starting point and generates infeasible iterates, a third "feasibility" direction is also included in the search direction; see, e.g., de Ghellinck and Vial [9] and Anstreicher [2].)

The affine-scaling direction mentioned above is the basis of the affine-scaling algorithm first proposed by Dikin [7] and rediscovered by Barnes [4] and Vanderbei, Meketon, and Freedman [29]. This method is believed not to be polynomial on the basis of results of Megiddo and Shub [21], although a variant that includes centering steps does possess a polynomial time bound (Barnes, Chopra, and Jensen [5]). The convergence results assume a step a fixed proportion of the way either to the boundary of the feasible region [29] or to the boundary of the inscribed ellipsoid [7], [4], which corresponds to a step of fixed Euclidean length in the transformed space.

* Received by the editors May 23, 1990; accepted for publication (in revised form) April 2, 1991. This research was supported in part by National Science Foundation grant DMS-8904406 and by the National Science Foundation, the Air Force Office of Scientific Research, and the Office of Naval Research through National Science Foundation grant DMS-8920550.

† School of Operations Research and Industrial Engineering, College of Engineering, Cornell University, Ithaca, New York 14853.

In this paper we propose a new algorithm whose search direction is a very simple combination of the affine-scaling and constant-cost centering directions. The step length is a constant in the transformed space. This simple algorithm attains the best-known complexity for the number of iterations without requiring the generation of lower bounds on the objective value or of dual feasible iterates. The proof, however, does make use of results of Ye [31] concerning the primal–dual potential function used in his potential-reduction algorithm.

Complexity results given in the literature typically address the case of a linear programming problem with integer data, and bound the computational work in terms of the number of inequalities n (the number of variables in a standard form problem) and the length L of the input (the total number of bits necessary to describe the problem). Hence, after suitable initialization, the projective-scaling algorithms require $O(nL)$ iterations (and $O(n^{3.5}L)$ or $O(n^4L)$ arithmetic operations in total) and the path-following and potential-reduction methods, $O(\sqrt{n}\, L)$ iterations (and $O(n^3L)$ or $O(n^{3.5}L)$ arithmetic operations). This number of iterations guarantees a feasible solution that is close enough to optimal that an exact solution can be obtained with modest additional computational effort. However, we feel it is more appropriate for linear programming (where the data are usually regarded as real) to state the complexity results in terms of n and a parameter t, which represents the precision required as well as the quality (initial objective value and "closeness" to the central path) of the initial solution. Our algorithms require $O(\sqrt{n}\, t)$ or $O(nt)$ iterations in this sense, and easily translate to $O(\sqrt{n}\, L)$ or $O(nL)$ iteration methods in the integer data case.

Section 2 describes the basic algorithm, and the $O(\sqrt{n}\, t)$ complexity result is derived in § 3. Section 4 describes two variants, one of which maintains this complexity, while the other requires $O(nt)$ iterations. These variants sacrifice some of the simplicity of the basic algorithm for improved practical behavior, and, in particular, recur lower-bound estimates of the optimal value; the algorithms then bear a strong resemblance to those of Gonzaga [12], Ye [31], and Freund [8]. Section 5 contains the results of preliminary computational experience showing the superiority of the second variant. This illustrates a phenomenon which has also been observed elsewhere: the better practical versions of interior-point algorithms frequently do not have the best theoretical complexity bounds. Our results also suggest the most important modifications in making the basic algorithm efficient; line searches are essential, and then the tightest lower bounds significantly improve performance.

2. The basic algorithm. We consider the linear programming problem in standard form:

$$(P) \qquad \begin{aligned} &\min c^T x, \\ &Ax = b, \\ &x \geqq 0, \end{aligned}$$

where A is $m \times n$. Let $F(P) := \{x \in \mathbb{R}^n\colon Ax = b,\ x \geqq 0\}$ denote its feasible region, and $F_+(P) := \{x \in F(P)\colon x > 0\}$ the relative interior of the feasible region. We assume that $F_+(P)$ is nonempty, and that (P) has a nonempty bounded set of optimal solutions. Let $v(P)$ denote the optimal value of (P).

We suppose that an initial point $x^0 \in F_+(P)$ is available. At iteration k we will have the current iterate $x^k \in F_+(P)$, and we define a scaled problem $(\bar{P})$ as follows. Let $X_k :=$ diag (x^k) be the diagonal matrix with the components of x^k down its diagonal, and consider the affine transformation $x \rightarrow \bar{x} := X_k^{-1}x$. The image of x^k under this transformation is e, the vector of ones in $\mathbb{R}^n$. In terms of $\bar{x}$, (P) becomes

$$(\bar{P}) \qquad \begin{aligned} &\min \bar{c}^T\bar{x}, \\ &\bar{A}\bar{x} = b, \\ &\bar{x} \geqq 0, \end{aligned}$$

where $\bar{A} := AX_k$ and $\bar{c} := X_k c$.

In the scaled problem $(\bar{P})$, there are two very important directions. Let $P_{\bar{A}}$ denote projection into the null space of $\bar{A}$. The first direction is the affine-scaling direction

$$(1) \qquad -\bar{c}_p := -P_{\bar{A}}\bar{c};$$

the second is the projection of the negative gradient of the barrier function

$$(2) \qquad p(\bar{x}) := -\sum_j \ln \bar{x}_j$$

evaluated at the point $\bar{x} = e$, which is

$$(3) \qquad e_p := P_{\bar{A}} e.$$

If $\bar{c}_p = 0$, then it is easy to see that all feasible points of $(\bar{P})$ have the same objective function value and so are optimal, and hence x^k is optimal in (P). Henceforth, we assume that $\bar{c}_p \neq 0$.

Most of the directions we are concerned with are combinations of the form

$$(4) \qquad \bar{d}_\beta := -\beta\bar{c}_p + e_p$$

of our two basic directions, for some scalar β. In particular, the direction $\bar{d}_\alpha$, where

$$(5) \qquad \alpha := \bar{c}_p^T e / \bar{c}_p^T \bar{c}_p,$$

will be very important to us. We note that it has three properties:

$$(6) \qquad \bar{d}_\alpha = \operatorname{argmin}\{\|\bar{d}\| : \bar{d} = \bar{d}_\beta \text{ for some } \beta\};$$

$$(7) \qquad \bar{d}_\alpha^T \bar{c}_p = 0; \quad \text{and}$$

$$(8) \qquad \bar{d}_\beta^T \bar{d}_\alpha = \bar{d}_\alpha^T \bar{d}_\alpha \quad \text{for all } \beta.$$

It is easy to see that $\bar{d}_\alpha$ is the steepest descent direction for the barrier function in the set $\{\bar{x}: \bar{A}\bar{x} = b,\ \bar{c}^T\bar{x} = \bar{c}^T e\}$, so we call $\bar{d}_\alpha$ the *constant-cost centering direction.* This direction appears in the centered version of the affine-scaling algorithm due to Barnes, Chopra, and Jensen [5] and in the monotonic versions of the standard-form projective variant (Anstreicher [1]) and of the scaled potential algorithm (Anstreicher [3]).

The direction of our algorithm is then chosen as follows:

Case 1. $\|\bar{d}_\alpha\| \geqq .3$. Then set

$$(9) \qquad \bar{d} = \bar{d}_\alpha / \|\bar{d}_\alpha\|.$$

Case 2. $\|\bar{d}_\alpha\| < .3$. Then set

$$(10) \qquad \bar{d} = -\bar{c}_p / \|\bar{c}_p\|.$$

Thus our direction is proportional either to the constant-cost centering direction or the affine-scaling direction, and in either case it is normalized to have length 1.

Having defined the direction $\bar{d}$, we take a step of length .2, so that

$$(11) \qquad \bar{x}_+ = e + .2\bar{d}$$

in the transformed space, and then

$$(12) \qquad x^{k+1} = X_k(e + .2\bar{d}) = x^k + .2X_k\bar{d}.$$

Since $\|\bar{d}\| = 1$, $\bar{x}_+ > 0$ and it is easy to check that $\bar{A}\bar{x}_+ = b$; hence $x^{k+1} \in F_+(P)$. Moreover, $\bar{c}^T\bar{x}_+ \leqq \bar{c}^Te$ (strictly if Case 2 obtains), so that $c^Tx^{k+1} \leqq c^Tx^k$.

3. Analysis. In this section we show that, if x^0 is suitably chosen, in $O(\sqrt{n}\ t)$ iterations we will have an iterate $x^k \in F_+(P)$ with

$$c^Tx^k - v(P) \leqq 2^{-t}. \tag{13}$$

The argument uses the results proved by Ye [31] in his scaled potential-reduction algorithm.

The dual of (P) is

$$(D) \qquad \begin{aligned} &\max b^Ty, \\ &A^Ty + s = c, \\ &s \geqq 0, \end{aligned}$$

and, for any $x \in F(P)$ and (y, s) feasible in (D), the duality gap is $b^Ty - c^Tx = x^Ts \geqq 0$. Let $F(D) = \{s \in \mathbb{R}^n \colon A^Ty + s = c$ for some y and $s \geqq 0\}$ and $F_+(D) = \{s \in F(D) \colon s > 0\}$. The fact that (P) has a nonempty bounded optimal solution set implies that $F_+(D) \neq \emptyset$. For an $x \in F_+(P)$ and $s \in F_+(D)$, and for any $q \geqq 0$, we define the primal-dual potential function (with parameter q) to be

$$\begin{aligned} \phi_q(x, s) &:= q \ln (x^Ts) - \sum_j \ln x_j - \sum_j \ln s_j - n \ln n \\ &= (q - n) \ln (x^Ts) - \sum_j \ln \frac{x_js_j}{x^Ts/n} \\ &\geqq (q - n) \ln (x^Ts) \end{aligned} \tag{14}$$

since the $x_js_j/(x^Ts/n)$ terms are positive with arithmetic mean one. We also use $\phi(x, s)$ to denote $\phi_{\bar{q}}(x, s)$ where $\bar{q} := n + \sqrt{n}$.

The condition we require on $x^0 \in F_+(P)$ is that, for some $s^0 \in F_+(D)$ (which need not be known), $\phi(x^0, s^0) = O(\sqrt{n}\ t)$. (When the data of (P) are integer, Monteiro and Adler [24], [25] show how to construct a related linear programming problem for which such an initial (x^0, s^0) can easily be obtained, with $t = L$, the size of the input.)

Suppose that at each iteration we can reduce ϕ by a constant. Then in $O(\sqrt{n}\ t)$ iterations we will have (x^k, s^k), with $\phi(x^k, s^k) \leqq -\sqrt{n}\ t$, so that by (14),

$$c^Tx^k - v(P) \leqq (x^k)^Ts^k \leqq 2^{-t},$$

as required.

We aim to show that this constant reduction is achieved even though the iterates s^k are not explicitly computed. For each $x^k \in F_+(P)$, we use an associated $s^k = s(x^k) \in F_+(D)$ that minimizes $\phi(x^k, \cdot)$. First we show the existence and uniqueness of such an s^k. (In fact, it is not hard to see that s^k is on the central path [6], [20] for the dual.)

PROPOSITION 1. *If $\hat{x} \in F_+(P)$, then* $\inf\{\phi(\hat{x}, s) \colon s \in F_+(D)\}$ *is attained by a unique $\hat{s} \in F_+(D)$. Write $\hat{s} = s(\hat{x})$. If $\tilde{x} \in F_+(P)$ with $c^T\tilde{x} \leqq c^T\hat{x}$, then $c^T\hat{x} - \hat{x}^T\hat{s} \leqq c^T\tilde{x} - \tilde{x}^T\tilde{s}$, where $\hat{s} = s(\hat{x})$, $\tilde{s} = s(\tilde{x})$.*

Proof. Choose any $\tilde{s} \in F_+(D)$; then we can confine the minimization to those $s \in F_+(D)$ with $\phi(\hat{x}, s) \leqq \phi(\hat{x}, \tilde{s})$. By (14) this shows that we can add the constraint $\hat{x}^Ts \leqq \mu$ for some μ, and since $\hat{x} > 0$, that s can be confined to a bounded set. Next, since $\phi(\hat{x}, s) \geqq \sqrt{n} \ln (\hat{x}^Ts) - \ln (\hat{x}_js_j) + \ln (\hat{x}^Ts) - \ln n$, and $\hat{x}^Ts \geqq c^T\hat{x} - v(\text{P}) > 0$, we can further restrict s_j to be at least some positive $\underline{s}_j$, and this argument applies to each j. But $\phi(\hat{x}, \cdot)$ is continuous on the compact set $\{s \in F(D) \colon \hat{x}^Ts \leqq \mu,\ s_j \geqq \underline{s}_j$ all $j\}$, so it attains its minimum there, and existence follows. Uniqueness is then implied by the argument of [28, Lemma 2.3]. So we can write $\hat{s} = s(\hat{x})$.

Now suppose $\hat{s} = s(\hat{x})$, $\tilde{s} = s(\tilde{x})$, where $c^T\tilde{x} \leqq c^T\hat{x}$. If $c^T\tilde{x} = c^T\hat{x}$, then $\phi(\hat{x}, \cdot)$ and $\phi(\tilde{x}, \cdot)$ differ by a constant, so $\hat{s} = \tilde{s}$ and the second part follows easily; both sides of the inequality equal $b^T\hat{y}$, where $(\hat{y}, \hat{s})$ is feasible in (D). So assume that $c^T\tilde{x} < c^T\hat{x}$. Then from $\phi(\hat{x}, \hat{s}) \leqq \phi(\hat{x}, \tilde{s})$ and $\phi(\tilde{x}, \tilde{s}) \leqq \phi(\tilde{x}, \hat{s})$, we deduce that $\ln \hat{x}^T\hat{s} + \ln \tilde{x}^T\tilde{s} \leqq \ln \hat{x}^T\tilde{s} + \ln \tilde{x}^T\hat{s}$. Suppose $(\hat{y}, \hat{s})$ and $(\tilde{y}, \tilde{s})$ are feasible in (D). By exponentiating the last inequality and simplifying, we find

$$(c^T\hat{x} - c^T\tilde{x})(b^T\hat{y} - b^T\tilde{y}) \leqq 0.$$

Then $b^T\hat{y} \leqq b^T\tilde{y}$, which yields the second part.

We will show that $\phi(x^{k+1}, s(x^k)) \leqq \phi(x^k, s(x^k)) - .02$ for each k, so that

$$(15) \qquad \phi(x^{k+1}, s(x^{k+1})) \leqq \phi(x^k, s(x^k)) - .02$$

a fortiori. This will prove the desired complexity result. (In effect, we are working with the primal-only potential function

$$(16) \qquad \psi(x) := \min\{\phi(x, s) \colon s \in F_+(D)\} = \phi(x, s(x)),$$

which is the only such function we know that can ensure an $O(\sqrt{n}\, t)$ iteration bound.)

Note that $\phi(\Lambda^{-1}x, \Lambda s) = \phi(x, s)$ for any positive definite diagonal matrix Λ. We can therefore always scale so that our current iterate x^k is e, and it is straightforward to check that the algorithm of § 2 is invariant under such scaling. We will therefore assume until the statement of Theorem 1 that such a scaling has already been performed, so that $x^k = e$, and we omit the overbars in our notation, so that $c_p = P_A c$, $e_p = P_A e$, $d_\beta = -\beta c_p + e_p$, $\alpha = c_p^T e / c_p^T c_p$, and d is our search direction. We wish to show that

$$(17) \qquad \phi(e + .2d, s(e)) \leqq \phi(e, s(e)) - .02;$$

this will then imply (15), as desired.

A key point is that, for any $q \geqq 0$,

$$-P_A(\nabla_x \phi_q(e, s(e))) = -P_A\left(\frac{q}{e^T s(e)}\, s(e) - e\right) = -\frac{q}{e^T s(e)}\, c_p + e_p$$

(since $A^T y + s(e) = c$ for some y, $P_A s(e) = P_A c = c_p$), and this is of the form d_β for some $\beta \geqq 0$. We now have the following result.

LEMMA 1. *For any* $q \geqq n + \sqrt{n}$,

$$(18) \qquad \left\| -\frac{q}{e^T s(e)}\, c_p + e_p \right\| \geqq .4.$$

Proof. Assume the contrary, so that $\|h\| < .4$ where $h = (q/e^T s(e))c_p - e_p$ for some $q \geqq n + \sqrt{n}$. Then, for some y,

$$A^T y + \frac{e^T s(e)}{q}(e + h) = c,$$

so that

$$(19) \qquad \bar{s} := \frac{e^T s(e)}{q}(e + h) \in F_+(D).$$

Moreover, the associated duality gap is

$$(20) \qquad e^T\bar{s} = \frac{e^T s(e)}{q}(e^T(e + h)) \leqq \frac{e^T s(e)}{n + \sqrt{n}}(n + .4\sqrt{n}) \leqq \left(1 - \frac{.6\sqrt{n}}{n + \sqrt{n}}\right) e^T s(e).$$

Finally, Lemma 2 of Ye [31] shows that

$$\left\| \bar{s} - \frac{e^T\bar{s}}{n} e \right\| \leqq .5 \frac{e^T\bar{s}}{n}; \tag{21}$$

note that Ye's argument does not require that q (his ρ) equals $n+\sqrt{n}$. We can then argue as in Theorem 1 of Ye [31] that

$$\phi(e, \bar{s}) \leqq \phi(e, s(e)) - \frac{.6}{2} + \frac{(.5)^2}{2(1-.5)} < \phi(e, s(e)),$$

which contradicts our choice of $s(e)$. Hence (18) must hold.

We also require the following standard result; see, for instance, Ye [31].

LEMMA 2. *If* $Ad = 0$, $\|d\| = 1$, *and* $0 < \gamma < 1$, *then*

$$\phi_q(e+\gamma d, s) \leqq \phi_q(e, s) + \gamma \nabla_x \phi_q(e, s)^T d + \frac{\gamma^2}{2(1-\gamma)}. \tag{22}$$

Now we consider the two cases of our algorithm, and prove that in each case (17) holds. Define δ by $-P_A \nabla_x \phi(e, s(e)) = d_\delta$.

First assume that $\|d_\alpha\| \geqq .3$ so that $d = d_\alpha / \|d_\alpha\|$. Then

$$\begin{aligned}\nabla_x \phi(e, s(e))^T d &= (P_A \nabla_x \phi(e, s(e)))^T d_\alpha / \|d_\alpha\| \\ &= -d_\delta^T d_\alpha / \|d_\alpha\| \\ &= -d_\alpha^T d_\alpha / \|d_\alpha\| = -\|d_\alpha\| \leqq -.3,\end{aligned}$$

where the first equality holds since d is in the null space of A, the second holds by definition of δ, and the third by (8). Hence with $\gamma = .2$, Lemma 2 yields

$$\begin{aligned}\phi(e+.2d, s(e)) &\leqq \phi(e, s(e)) - .2 \times .3 + \frac{(.2)^2}{2(1-.2)} \\ &\leqq \phi(e, s(e)) - .02.\end{aligned} \tag{23}$$

Now suppose that $\|d_\alpha\| < .3$. Then $\delta > \alpha$ (otherwise $\alpha = q/e^T s(e)$ for some $q \geqq n + \sqrt{n}$, contradicting Lemma 1) and, in fact, since $\|d_\delta\| \geqq .4$, d_δ makes an angle of at least $\arccos(\frac{3}{4})$ with d_α, and hence an angle of at most $\arccos(\frac{5}{8})$ with $-c_p$ (see Fig. 1). Hence, with $d = -c_p/\|c_p\|$, we have $d_\delta^T d \geqq \frac{5}{8} \times .4 = .25$. Using Lemma 2 we can conclude that

$$\begin{aligned}\phi(e+.2d, s(e)) &\leqq \phi(e, s(e)) - .2 \times .25 + \frac{(.2)^2}{2(1-.2)} \\ &\leqq \phi(e, s(e)) - .02.\end{aligned}$$

To summarize, we have shown our convergence result.

THEOREM 1. *At each iteration of the algorithm of* § 2,

$$\phi(x^{k+1}, s(x^{k+1})) \leqq \phi(x^k, s(x^k)) - .02. \tag{24}$$

If $\phi(x^0, s^0) = O(\sqrt{n}\ t)$ *for some* $s^0 \in F_+(D)$, *then after* $O(\sqrt{n}\ t)$ *iterations we have* x^k *with*

$$c^T x^k - v(P) \leqq 2^{-t}.$$

4. Refinements. The simple algorithm of § 2 does not perform well in practice. Indeed, (12) shows that each component of x^k can decrease by at most 20 percent in each iteration, so that the best we can hope for is linear convergence with ratio .8. In

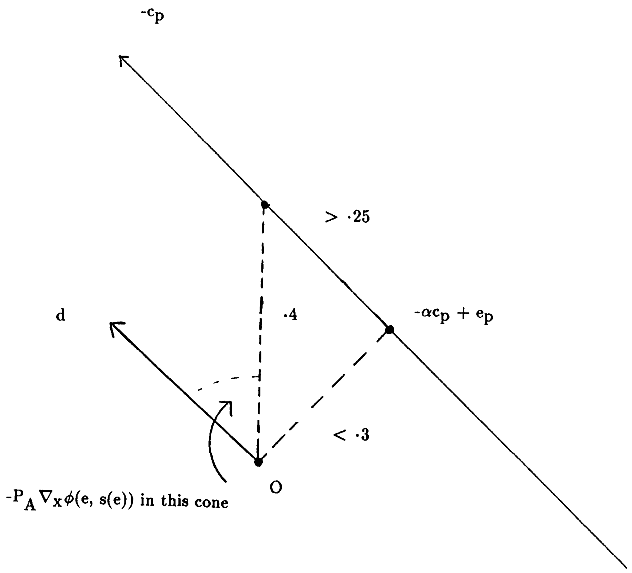


FIG. 1

this section we describe refinements that attempt to improve the convergence while maintaining the simple structure of the algorithm. Usually, an interior-point method can be improved significantly by incorporating a line search. Here, however, the potential function on which we would like to perform a line search cannot easily be computed, and hence complications arise. We use the insights obtained from the analysis of § 3 and update lower bounds on the optimal value. The resulting methods bear a strong resemblance to those of Gonzaga [12], Ye [31], and Freund [8].

Our first variant maintains (24), and hence provides an $O(\sqrt{n}\ t)$ iteration method ($O(\sqrt{n}\ L)$ for a problem with integer data), while our second variant is preferable in practice but has only an $O(nt)$ iteration bound.

At each iteration, we try to update the current lower bound z_k, then choose a search direction and make a line search in this direction to minimize approximately some potential function. In the second variant we only need $z_k \leqq v(P)$. In the first, we require further that $c^Tx^k - z_k \geqq (x^k)^Ts(x^k)$, and when we update z_k to z_{k+1}, we insist that z_{k+1} also satisfies this inequality. Since our algorithm is monotonic, Proposition 1 then implies that $c^Tx^{k+1} - z_{k+1} \geqq (x^{k+1})^Ts(x^{k+1})$. In both variants we can initialize with $z_0 = -\infty$.

To describe the iteration, it is again convenient to assume that the problem has been scaled, if necessary, so that the current iterate is e. Suppose c_p, e_p, and α have been calculated. In the first variant, z_k is updated as follows.

(25)
If $\|d_\alpha\| \geqq .4$, $z_{k+1} := z_k$.
Otherwise, let $\varepsilon > \alpha$ be such that $\|d_\varepsilon\| = .4$.
If $\varepsilon \leqq 0$, $z_{k+1} := z_k$.
If $\varepsilon > 0$, let $z_+ := c^Te - (n + \sqrt{n})/\varepsilon$ and $z_{k+1} := \max\{z_k, z_+\}$.

This method of updating z_k is very similar to that of Gonzaga [12, § 5] and closely related to those of Freund [8] and Ye [31]. We then have Lemma 3.

LEMMA 3. *Assume that $c^Te - z_k \geqq e^Ts(e)$, so that z_k is a valid lower bound on $v(P)$. Then if z_{k+1} is updated by (25), $c^Te - z_{k+1} \geqq e^Ts(e)$ also, so that z_{k+1} is a valid lower bound, too.*

Proof. There is nothing to show if $z_{k+1}=z_k$. Otherwise, $z_{k+1}=z_+$. Because of Lemma 1, $(n+\sqrt{n})/e^Ts(e)\geqq \varepsilon=(n+\sqrt{n})/(c^Te-z_+)$, which gives the desired inequality. Then z_{k+1} is a valid lower bound because it is at most $c^Te-e^Ts(e)$, the value of the dual solution corresponding to $s(e)$.

Next we compute the search direction d as follows. First, let

$$\zeta:=\begin{cases} q/(c^Te-z_{k+1}) & \text{if } z_{k+1}>-\infty, \\ 0 & \text{if } z_{k+1}=-\infty, \end{cases} \tag{26}$$

for $q=n+\sqrt{n}$. Note that, if $\|d_\alpha\|<.4$, then $\zeta\geqq\varepsilon>\alpha$, so that $\|d_\zeta\|\geqq .4$ by definition of ε, while if $\|d_\alpha\|\geqq .4$, then $\|d_\zeta\|\geqq .4$ by (6).

Case A. $\zeta<\alpha$. Then $\|d_\alpha\|\geqq .4$. In this case, set

$$d=\frac{d_\alpha}{\|d_\alpha\|}. \tag{27}$$

Case B. $\zeta\geqq\alpha$. Let

$$\tilde{d}=\frac{d_\zeta}{\|d_\zeta\|}+\frac{-c_p}{\|c_p\|},\qquad d=\frac{\tilde{d}}{\|\tilde{d}\|}. \tag{28}$$

(Note that Gonzaga [12], Ye [31], and Freund [8], given a lower bound z_{k+1} and resulting ζ with $\|d_\zeta\|$ sufficiently large, would choose $d=d_\zeta$.)

LEMMA 4. *If d is defined as above, then*

$$\nabla_x\phi_q(e, s(e))^Td\leqq -.28$$

for any $q\geqq 0$ if $z_{k+1}=-\infty$ and for any $q\geqq n+\sqrt{n}$ if $z_{k+1}>-\infty$.

Proof. Define δ by $-P_A\nabla_x\phi_q(e, s(e))=d_\delta$. Then $\nabla_x\phi_q(e, s(e))^Td=-d_\delta^Td$. In Case A, $d=d_\alpha/\|d_\alpha\|$ and $-d_\delta^Td=-d_\delta^Td_\alpha/\|d_\alpha\|=-d_\alpha^Td_\alpha/\|d_\alpha\|=-\|d_\alpha\|\leqq -.4$ using (8).

In Case B, Lemma 3 and (26) show that $\delta\geqq\zeta\geqq\alpha$ as long as $z_{k+1}=-\infty$ ($\zeta=0$, $\delta\geqq 0$ for any $q\geqq 0$) or $q\geqq n+\sqrt{n}$ (using Lemma 1). Hence $\|d_\delta\|\geqq\|d_\zeta\|\geqq .4$. Now d is a unit vector that bisects the angle between the two extreme directions for d_δ, namely, d_ζ and $-c_p$. Since $\zeta\geqq\alpha$, these two directions form a nonobtuse angle, so the angle between d_δ and d is at most $\pi/4$. Then

$$-d_\delta^Td\leqq -\|d_\delta\|\cos(\pi/4)\leqq -.4\cos(\pi/4)\leqq -.28.$$

Now we search on the half-line $\{e+\lambda d\colon \lambda\geqq 0\}$. If $z_{k+1}=-\infty$, we seek to minimize

$$\phi_0^P(x, -\infty):=-\sum_j \ln x_j, \tag{29}$$

the barrier function. It is easy to see that, in either Case A or Case B, ϕ_0^P can be decreased by at least .03. Moreover, since $c^Td\leqq 0$, we obtain at least as great a decrease in the primal potential function

$$\phi_q^P(x, z):=q\ln(c^Tx-z)-\sum_j \ln x_j \tag{30}$$

for any $q\geqq n$ and $z\leqq v(P)$, and hence in $\phi(x, s(e))$, since this differs by a constant from $\phi_q^P(x, z)$ with $q=n+\sqrt{n}$ and $z=c^Te-e^Ts(e)$.

If $z_{k+1}>-\infty$, we seek to minimize

$$\phi_q^P(x, z_{k+1}) \tag{31}$$

for $q=n+\sqrt{n}$. Again, our updating scheme for z_k ensures that this can be decreased by at least .03, and that $\phi(x, s(e))$ can be reduced by at least as much.

Hence our first variant of the basic algorithm preserves (24), while employing directions probably closer to the negative of the projected scaled gradient of the appropriate potential function and allowing line searches to achieve greater reductions in such potential functions.

The following argument, which improves the author's original and is due to Ye (private communication), shows that a finite lower bound must be generated in $O(\sqrt{n}\ t)$ iterations. Indeed, $\phi(x^0, s^0) = O(\sqrt{n}\ t)$ implies $\phi_n(x^0, s^0) = O(\sqrt{n}\ t)$. Until a lower bound is generated, ϕ_n decreases by a constant at each iteration, while it is bounded below by zero. This yields the desired complexity.

Once again, we have a monotonic algorithm, so all iterates lie in the compact set $\{x \in F(P): c^Tx \leqq c^Tx^0\}$. Let $\bar{\xi} := \max\{e^Tx/n: x \in F(P),\ c^Tx \leqq c^Tx^0\}$ and let $\underline{\xi}^0 := (\prod_{j=1}^n x_j^0)^{1/n}$, the geometric mean of the components of x^0. Our remarks above on decreasing ϕ_q^P ensure that

$$\phi_q^P(x^k, z_k) \leqq \phi_q^P(x^0, z_k) - .03k \tag{32}$$

for $q = n + \sqrt{n}$, and using the estimates above and the fact that $n \leqq n + \sqrt{n} \leqq 2n$, we find

$$(c^Tx^k - z_k) \leqq (\bar{\xi}/\underline{\xi}^0) \cdot \exp(-.01k/n) \cdot (c^Tx^0 - z_k). \tag{33}$$

Assuming that $c^Tx^0 - z_0$ and $\bar{\xi}/\underline{\xi}^0$ are bounded by 2^t, we can obtain from (33) an $O(nt)$ bound on the number of iterations to obtain $c^Tx^k - z_k \leqq 2^{-t}$. This contrasts with the $O(\sqrt{n}\ t)$ bound in Theorem 1, which is still valid; one reason for the difference is that $v(P)$ is not known in Theorem 1, whereas here z_k is a known (and possibly not very tight) lower bound on $v(P)$. By contrast, the algorithms of Ye [31] and Freund [8, § 4] require only $O(\sqrt{n}\ t)$ iterations to obtain this inequality, where now z_k is the lower bound associated with an explicitly computed dual solution.

This first variant of the basic algorithm generates better directions than the original, but it still converges rather slowly in practice. The reason appears to be that the lower bounds generated are rather poor, so that in the line search to minimize $\phi_{n+\sqrt{n}}^P(x, z_{k+1})$, the barrier term $-\sum_j \ln x_j$ forces a small step size, typically of the order of a tenth of the maximum feasible step size.

In order to obtain better lower bounds, we reconsider the argument used in Lemma 1. For any $\beta > 0$, if $\|-\beta c_p + e_p\| \leqq 1$, we find that $s := (1/\beta)(e + \beta c_p - e_p) = c_p + (e - e_p)/\beta$ belongs to $F(D)$, and the associated duality gap is $e^Tc_p + e^T(e - e_p)/\beta = e^Tc_p + \|e - e_p\|^2/\beta$, which is decreasing as a function of β. In fact, it is not even necessary that $\|-\beta c_p + e_p\| \leqq 1$; as long as s is nonnegative, it provides such a bound. Hence we have Lemma 5.

LEMMA 5. *If there is some* $\beta > 0$ *with* $c_p + (e - e_p)/\beta \geqq 0$, *let* $\bar{\beta}$ *be the maximum such. Then* $z_+ := c^Te - c_p^Te - \|e - e_p\|^2/\bar{\beta}$ *is a lower bound on* $v(P)$.

This lower bound was also obtained independently by Gonzaga [14]. In fact, it is also implicit in the statement on performing a line search for Δ in Freund [8, § 3].

Thus our second variant sets z_+ as above if the criterion of the lemma is met, and then sets

$$z_{k+1} := \max\{z_k, z_+\}.$$

In this second variant, z_{k+1} is always a lower bound, but we may not have $c^Te - z_{k+1} \geqq e^Ts(e)$, as we did before. Given z_{k+1}, we define the search direction d as above (see (27), (28)) and perform a line search on (29) or (31) as before. However, since we no longer have $z_{k+1} \leqq c^Tx^k - (x^k)^Ts(x^k)$, we may not obtain a corresponding decrease in $\phi(x, s(e))$. Nevertheless, our line search can guarantee a decrease of at least .03 in $\phi_0^P(x, -\infty)$ (and in $\phi_q^P(x, z)$ for any z and any q) if $z_{k+1} = -\infty$ and the same decrease in $\phi_q^P(x, z_{k+1})$ for $q=n+\sqrt{n}$ if $z_{k+1}>-\infty$. Thus (32) and (33) remain valid, and we deduce that the second variant provides an $O(nt)$ iteration algorithm. This remains true if q in (26) and (31) is $O(n)$, rather than $n+\sqrt{n}$.

5. Computational results. We conclude the paper by giving the results of some very preliminary computational testing. These results suggest that the key refinement is the incorporation of a line search, followed by the use of improved lower bounds.

The test problems were obtained as follows. For a given m and n, we generated each entry of A, y, and s as an independent standard normal random variable; then set $b=Ae$ and $c=A^Ty+|s|$, where $|s|=(|s_j|)$; the initial solution was $x^0=e$. We describe the results of several variants. In all of them, we generated a sequence of lower bounds z_k, even if they were not used in the algorithm, in order to terminate when $(c^Tx-z_k)/\max\{1, |c^Tx|\}$ was less than 10^{-4}. Whenever a search direction d^k was obtained, we checked this termination criterion at the point $x=x^k+\lambda_{\max}d_k$, where $\lambda_{\max}=\max\{\lambda\colon x^k+\lambda d^k\geqq 0\}$, so that the algorithm could terminate in a reasonable number of iterations even if it chose rather small step sizes.

The first variant differs from the basic algorithm of § 2 in three respects. It uses a sequence of lower bounds, generates improved directions, and performs a line search on a suitable potential function. The second variant adds to this an improved lower bound update. For one 50 × 100 problem, we tested all combinations of these features, as well as trying $q=n+\sqrt{n}$ and $q=2n$ for the second variant. Thus we tried nine algorithms: the basic algorithm of § 2 (with termination based on lower bounds, as in the first variant), then this basic method with a line search, the basic method with improved search directions, and the basic method with improved lower bounds (which only affect the termination criterion in this case); next, the basic algorithm with two of these features, namely, with improved directions and improved lower bounds, with a line search and improved lower bounds and with a line search and improved directions (the first variant); and finally, with all three features and $q=n+\sqrt{n}$ or $q=2n$ (the second variant). The results are given in Table 1, which also presents a typical value of $\lambda_k/\lambda_{\max}$, where λ_k is the step size chosen and $\lambda_{\max}$ the maximum feasible step size. All runs used PRO-MATLAB [23] Version 3.5e on a Sun SPARCstation 1.

It is clear that the most significant enhancement is the incorporation of a line search; when this is present, improved lower bounds are more important than better directions because they allow longer step sizes to be used. Finally, all three features together allow a considerable decrease in the number of iterations required, especially for $q=2n$. (If we also recur the improved lower bounds in the first variant, but only

TABLE 1
Computational results on a 50 × 100 *problem.*

Method	Number of iterations	Typical $\lambda_k/\lambda_{\max}$
Basic	347	.04
Basic with line search	98	.13
Basic with improved directions	347	.04
Basic with improved bounds	282	.04
Basic with improved directions and improved bounds	262	.04
Basic with line search and improved bounds	27	.46
Basic with line search and improved directions (first variant)	95	.14
Second variant, $q=n+\sqrt{n}$	12	.87
Second variant, $q=2n$	11	.98

use them in the termination criterion (thus maintaining the theoretical $O(\sqrt{n}\ t)$ bound), then the number of iterations required decreases only slightly, to 77. Thus the difficulty is slow primal convergence, not a poor termination criterion.) Similar results hold for other test problems solved.

We also solved ten random 50×100 problems using the second variant. With $q = n + \sqrt{n}$, the average number of iterations was 12.2, with $\lambda_k / \lambda_{\max}$ typically .87; with $q = 2n$, the figures become 11.0 and .97. Five random 100×200 problems needed an average of 14.0 iterations with $q = n + \sqrt{n}$ and 12.2 with $q = 2n$. For 150×300 problems the figures were 14.4 and 13.0, respectively, and for 200×400 problems, 15.4 and 13.6. The typical step size ratio was similar to those reported above.

Finally, a single 100×200 problem, which required 16 or 14 iterations for the second variant, needed 161 for the first variant (144 if termination was based on the improved lower bounds) and 626 for the basic algorithm; and a 200×400 problem, also needing 16 or 14 iterations for the second variant, required 240 (or 202) for the first variant and 739 for the basic algorithm.

Acknowledgment. I am very grateful to the referees for their helpful comments, which considerably improved the presentation.

REFERENCES

[1] K. M. Anstreicher, *A monotonic projective algorithm for fractional programming*, Algorithmica, 1 (1986), pp. 483-498.

[2] ———, *A combined phase* I-*phase* II *scaled potential algorithm for linear programming*, CORE Discussion Paper 8939, Catholic University of Louvain, Louvain, Belgium, 1989.

[3] ———, *On monotonicity in the scaled potential algorithm for linear programming*, Linear Algebra Appl., 152 (1991), pp. 223-232.

[4] E. R. Barnes, *A variation of Karmarkar's algorithm for solving linear programming problems*, Math. Programming, 36 (1986), pp. 174-182.

[5] E. R. Barnes, S. Chopra, and D. L. Jensen, *A polynomial time version of the affine scaling algorithm*, Report RC 13965, IBM T. J. Watson Research Center, Yorktown Heights, NY, 1988.

[6] D. Bayer and J. C. Lagarias, *The nonlinear geometry of linear programming*: I. *Affine and projective scaling trajectories*, and II. *Legendre transform coordinates and central trajectories*, Trans. Amer. Math. Soc., 314 (1989), pp. 499-526 and 527-581.

[7] I. I. Dikin, *Iterative solution of problems of linear and quadratic programming*, Dokl. Akad. Nauk SSSR, 174 (1967), pp. 747-748. (English translation: Soviet Math. Dokl., 8 (1967), pp. 674-675.)

[8] R. Freund, *Polynomial-time algorithms for linear programming based only on primal scaling and projected gradients of a potential function*, Math. Programming, 51 (1991), pp. 203-222

[9] G. de Ghellinck and J.-Ph. Vial, *A polynomial Newton method for linear programming*, Algorithmica, 1 (1986), pp. 425-453.

[10] C. Gonzaga, *Search directions for interior linear programming methods*, Algorithmica, 6 (1991), pp. 153-181.

[11] ———, *An algorithm for solving linear programming in $O(n^3L)$ operations*, in Progress in Mathematical Programming, N. Megiddo, ed., Springer-Verlag, Berlin, 1988, pp. 1-28.

[12] ———, *Polynomial affine algorithms for linear programming*, Math. Programming, 49 (1990), pp. 7-21.

[13] ———, *Large step path-following methods for linear programming, Part* II: *Potential reduction method*, SIAM J. Optimization, 1 (1991), pp. 280-292.

[14] ———, *On lower bound updates in primal potential reduction methods for linear programming*, Report ES-227/90, COPPE, Federal University of Rio de Janeiro, Brazil, 1989.

[15] D. den Hertog and C. Roos, *A survey of search directions in interior-point methods for linear programming*, Report 89-65, Faculty of Technical Mathematics and Informatics, Delft University of Technology, Delft, The Netherlands, 1989.

[16] N. Karmarkar, *A new polynomial time algorithm for linear programming*, Combinatorica, 4 (1984), pp. 373-395.

[17] M. KOJIMA, S. MIZUNO, AND A. YOSHISE, *A primal-dual interior point method for linear programming*, in Progress in Mathematical Programming, N. Megiddo, ed., Springer-Verlag, Berlin, 1988, pp. 29-47.

[18] ———, *A polynomial algorithm for a class of linear complementarity problems*, Math. Programming, 44 (1989), pp. 1-26.

[19] ———, *An* $O(\sqrt{n}\,L)$ *iteration potential reduction algorithm for linear complementarity problems*, Math. Programming, 50 (1991), pp. 331-342.

[20] N. MEGIDDO, *Pathways to the optimal set in linear programming*, in Progress in Mathematical Programming, N. Megiddo, ed., Springer-Verlag, Berlin, 1988, pp. 131-158.

[21] N. MEGIDDO AND M. SHUB, *Boundary behavior of interior point algorithms in linear programming*, Math. Oper. Res., 14 (1989), pp. 97-146.

[22] J. E. MITCHELL AND M. J. TODD, *On the relationship between the search directions in the affine and projective variants of Karmarkar's linear programming algorithm*, in Contributions to Operations Research and Economics, B. Cornet and H. Tulkens, eds., M.I.T. Press, Cambridge, MA, 1989, pp. 237-250.

[23] C. B. MOLER, J. LITTLE, S. BANGERT, AND S. KLEIMAN, *Pro-Matlab User's Guide*, MathWorks, Sherborn, MA, 1987.

[24] R. C. MONTEIRO AND I. ADLER, *Interior path following primal-dual algorithms, Part* I: *Linear programming*, Math. Programming, 44 (1989), pp. 27-41.

[25] ———, *Interior path following primal-dual algorithms, Part* II: *Convex quadratic programming*, Math. Programming, 44 (1989), pp. 43-66.

[26] J. RENEGAR, *A polynomial-time algorithm based on Newton's method for linear programming*, Math. Programming, 40 (1989), pp. 59-93.

[27] G. SONNEVEND, *An analytical centre for polyhedrons and new classes of global algorithms for linear* (*smooth, convex*) *programming*, Lecture Notes in Control and Information Sciences 84, Springer-Verlag, New York, 1986, pp. 866-876.

[28] M. J. TODD AND B. P. BURRELL, *An extension of Karmarkar's algorithm for linear programming using dual variables*, Algorithmica, 1 (1986), pp. 409-424.

[29] R. J. VANDERBEI, M. S. MEKETON, AND B. A. FREEDMAN, *A modification of Karmarkar's linear programming algorithm*, Algorithmica, 1 (1986), pp. 395-407.

[30] H. YAMASHITA, *A polynomially and quadratically convergent method for linear programming*, Report, Mathematical Systems Institute, Tokyo, Japan, 1986.

[31] Y. YE, *An* $O(n^3 L)$ *potential reduction algorithm for linear programming*, Math. Programming, 50 (1991), pp. 239-258.

[32] U. ZIMMERMAN, *Search directions for a class of projective methods*, Z. Oper. Res., 34 (1990), pp. 353-379.

SIAM J. OPTIMIZATION
Vol. 2, No. 2, pp. 210-241, May 1992

003

AN SQP AUGMENTED LAGRANGIAN BFGS ALGORITHM FOR CONSTRAINED OPTIMIZATION*

R. H. BYRD†, R. A. TAPIA‡, AND YIN ZHANG§

Abstract. In this research an effective algorithm for nonlinearly constrained optimization using the structured augmented Lagrangian secant update recently proposed by Tapia is presented. The algorithm is globally defined, and uses a new and reliable method for choosing the Lagrangian augmentation parameter that does not require prior knowledge of the true Hessian. Considerable numerical experimentation with this algorithm, both embedded in a merit-function line search SQP framework and without line search, is presented. The algorithm is compared to the widely used damped BFGS secant update of Powell, which, like the one in this paper, was designed to circumvent the lack of positive definiteness in the Hessian of the Lagrangian. It is also established that when the algorithm converges it converges R-superlinearly, which is a strong result in that it makes no assumptions on the approximate Hessian or the augmentation parameter. An immediate corollary is a new result in unconstrained optimization: whenever the unconstrained BFGS secant method converges, it does so Q-superlinearly. This study has led to the conclusion that, when properly implemented, Tapia's structured augmented Lagrangian BFGS secant update has strong theoretical properties, and in experiments, is very competitive with Powell's damped BFGS update.

Key words. BFGS secant method, augmented Lagrangian, SQP methods, superlinear convergence, constrained optimization

AMS(MOS) subject classifications. 49D37, 65K05, 90C30

1. Introduction. In this work, we will be concerned with the equality-constrained optimization problem

$$\text{(1.1)}\qquad \begin{array}{ll}\text{minimize} & f(x),\\ \text{subject to} & h(x)=0,\end{array}$$

where $f:\mathbf{R}^n\to\mathbf{R}$, $h:\mathbf{R}^n\to\mathbf{R}^m (m<n)$, and f and h are generally nonlinear. The Lagrangian function associated with problem (1.1) is the function

$$\text{(1.2)}\qquad \ell(x,\lambda)=f(x)+\lambda^T h(x),$$

where $\lambda\in\mathbf{R}^m$ is called the vector of Lagrange multipliers or simply the Lagrange multiplier. We will be examining algorithms for solving this problem based on successive quadratic programming that make use of a modification of the Lagrangian in (1.2), the *augmented Lagrangian.*

As usual, ∇ will denote the gradient operator, ∇^2 the Hessian operator, and subscripts on these quantities signify partial differentiation. We will denote $\nabla f(x)$ by $g(x)$ and the matrix whose columns are $\nabla h_1(x), \nabla h_2(x), \cdots, \nabla h_m(x)$ by $A(x)$. On occasion, we employ the convention of writing g_k for $g(x_k)$ and g_* for $g(x_*)$, and similarly for other functions and other arguments. This usage should be clear from

* Received by the editors February 10, 1990; accepted for publication (in revised form) April 25, 1991. This research was sponsored by Strategic Defense Initiative Organization/Army Research Office, Air Force Office of Scientific Research grant 85-0243, Department of Energy grant DEFG05-86ER 25017, National Science Foundation grant CCR87024303, Air Force Office of Scientific Research grant 85-0251, and Army Research Office contract DAAL03-88-K-0086.

† Department of Computer Science, University of Colorado, Boulder, Colorado 80309.

‡ Department of Mathematical Sciences, Rice University, Houston, Texas 77251-1892.

§ Department of Mathematics and Statistics, University of Maryland, Catonsville, Maryland 21228.

the context. We will use x_* to denote a local solution of problem (1.1) and λ_* to denote a Lagrange multiplier vector, satisfying $\nabla_x \ell(x_*, \lambda_*) = 0$.

In unconstrained optimization the BFGS secant update has emerged as the secant update of choice. The convergence analysis of BFGS secant methods requires that the Hessian matrix that is being approximated be positive definite at the solution. Furthermore, this requirement is satisfied at any nonsingular local minimizer.

It is well known that a formal extension of the BFGS secant method can be made from unconstrained optimization to constrained optimization (problem (1.1)) by employing the so-called successive quadratic programming (SQP) framework. In anticipation of our later needs, we now state this formal extension in a line search globalization environment.

ALGORITHM 1.1 (Line search SQP Lagrangian BFGS method). Given $x_0 \in \mathbf{R}^n$ and a symmetric $B_0 \in \mathbf{R}^{n\times n}$, for $k = 1, 2, \cdots$, until convergence do

$$\begin{aligned} x_{k+1} &= x_k + \tau_k d_k, \\ \lambda_{k+1} &= \Lambda(x_k, x_{k+1}, B_k), \end{aligned} \tag{1.3}$$

$$\begin{aligned} s_k &= x_{k+1} - x_k, \\ y_k^\ell &= \nabla_x \ell(x_{k+1}, \lambda_{k+1}) - \nabla_x \ell(x_k, \lambda_{k+1}), \end{aligned} \tag{1.4}$$

$$B_{k+1} = B_k - \frac{B_k s_k s_k^T B_k}{s_k^T B_k s_k} + \frac{y_k^\ell y_k^{\ell T}}{y_k^{\ell T} s_k}, \tag{1.5}$$

where the line search direction d_k is the solution of the quadratic programming subproblem

$$\begin{aligned} &\text{minimize} \quad g_k^T d + \tfrac{1}{2} d^T B_k d, \\ &\text{subject to} \quad h_k + A_k^T d = 0, \end{aligned} \tag{1.6}$$

and the step-length τ_k is chosen to decrease a given merit (line search) function. The matrix B_k is interpreted as an approximation to $\nabla_x^2 \ell(x_k, \lambda_k)$. The function Λ in (1.3) is an updating formula for λ. A common choice for λ_{k+1} in (1.3) is the multiplier associated with the solution d_k of the subproblem (1.6). Observe that B_{k+1} satisfies the secant equation $B_{k+1} s_k = y_k^\ell$.

There is a major flaw in Algorithm 1.1. This flaw will be obvious once we invoke the following assumptions, which are standard in the theory of quasi-Newton methods for problem (1.1). They will be assumed throughout this paper.

ASSUMPTIONS.

A1. f and h_i have second derivatives that are Lipschitz continuous in an open, convex neighborhood $D \subset \mathbf{R}^n$ of the local solution x_*.

A2. $A(x_*)$ has full rank.

A3. $p^T \nabla_x^2 \ell(x_*, \lambda_*) p > 0$ for all $p \neq 0$ satisfying $A(x_*)^T p = 0$.

Note that A2 implies that λ_* is unique.

The deficiency of Algorithm 1.1 is that the local convergence theory for BFGS secant methods requires $\nabla_x^2 \ell(x_*, \lambda_*)$ to be positive definite and yet satisfaction of this condition is not guaranteed by the standard assumptions A1–A3.

When a line search globalization strategy is added to a BFGS secant method, it is essential that the approximation Hessian matrices B_k be positive definite. The well-known hereditary positive definiteness property of the BFGS secant update is that positive definite B_k leads to positive definite B_{k+1} if and only if $y_k^{\ell T} s_k > 0$. If $\nabla_x^2 \ell(x_*, \lambda_*)$

is not positive definite, we cannot guarantee the condition $y_k^{\ell T} s_k > 0$ even locally, i.e., for x_k and x_{k+1} near x_*, let alone globally. The desire to enforce this condition globally will play a major role in the present research.

Alternative formulations of the SQP Lagrangian BFGS secant method that circumvent the lack of positive definiteness of $\nabla_x^2 \ell(x_*, \lambda_*)$ have been challenging researchers now for many years. Perhaps the first alternative considered was replacing the Lagrangian with the augmented Lagrangian associated with problem (1.1) (see Han [12] and Tapia [24]). This latter function is

$$L(x, \lambda, \rho) = \ell(x) + \frac{\rho}{2} h(x)^T h(x), \qquad (\rho \geqq 0). \tag{1.7}$$

Observe that the Hessian of the augmented Lagrangian at a local solution of problem (1.1) has the form

$$H_*(\rho) \equiv \nabla_x^2 L(x_*, \lambda_*, \rho) = \nabla_x^2 \ell(x_*, \lambda_*, \rho) + \rho A(x_*) A(x_*)^T. \tag{1.8}$$

It is well known that for any augmentation parameter ρ greater than a threshold value $\bar{\rho}$, $H_*(\rho)$ is positive definite; therefore, if y_k is defined as (we will use y_k as a generic term and different choices of y_k will be denoted by different superscripts)

$$y_k^L = \nabla_x L(x_{k+1}, \lambda_{k+1}, \rho) - \nabla_x L(x_k, \lambda_{k+1}, \rho), \tag{1.9}$$

we can guarantee that near the solution $y_k^{LT} s_k > 0$ for ρ sufficiently large.

We arrive at the (line search) SQP augmented Lagrangian BFGS secant method for problem (1.1) by replacing y_k^ℓ in (1.4) with y_k^L from (1.9). The Broyden-Dennis-Moré theory was used by Han [12], Tapia [24], and Glad [9] to establish local and Q-superlinear convergence in the pair (x, λ) for a version of this algorithm under the standard assumptions A1-A3. Fontecilla, Steihaug, and Tapia [8] showed that the convergence in x is actually Q-superlinear.

Though theoretically attractive, this alternative has serious practical problems. First, a priori knowledge of the threshold value $\bar{\rho}$ for a given problem is generally unavailable. Second, the attempt to use large ρ seems to present severe numerical problems; see the examples given by Tapia [24] and Nocedal and Overton [16]. See Appendix B of Tapia [25] for some interesting comments on this issue. We emphasize that y_k^L given by (1.9) has the serious disadvantage that at some iterations it may not be possible to choose ρ sufficiently large so that $y_k^{LT} s_k$ is positive (even though it must be possible near the solution).

Another direction taken to circumvent the lack of positive definiteness of $\nabla_x^2 \ell(x_*, \lambda_*)$ is to use the BFGS secant update in the context of reduced Hessian (or projected Hessian) methods. In contrast to full Hessian methods, reduced Hessian methods approximate the Hessian restricted to the null space of the Jacobian of the constraints, where it is expected to be positive definite. Since the concern of the present work is full Hessian methods, we refer interested readers to Coleman and Conn [4], Nocedal and Overton [16], and Byrd and Nocedal [2] for further references on reduced Hessian methods. Fenyes [6] and Fontecilla [7] proposed full Hessian methods that have some of the flavor of the reduced Hessian methods.

Powell [19] proposed another modification to the (line search) SQP Lagrangian BFGS secant method that compensates for the lack of positive definiteness in the Hessian at the solution. Despite the fact that the true Hessian of the Lagrangian may not be positive definite at a solution, Powell chose to maintain a positive definite matrix by modifying y_k^ℓ whenever necessary. The modified y_k^P (say) has the form

$$y_k^P = \theta_k y_k^\ell + (1 - \theta_k) B_k s_k, \tag{1.10}$$

where the parameter θ_k is contained in $(0, 1]$. Notice from (1.5), if $\theta_k = 0$, then $B_{k+1} = B_k$; while if $\theta_k = 1$, we obtain B_{k+1} as the full BFGS update of B_k. For this reason, with Griewank [10], we refer to the use of (1.10) in (1.5) as the damped update. Powell chose θ_k so that

$$y_k^{PT} s_k \geqq \eta s_k^T B_k s_k$$

is always satisfied for some $\eta \in (0, 1)$. More specifically, the number $\theta_k \in (0, 1]$ is given the value

$$\theta_k = \begin{cases} 1, & y_k^{\ell T} s_k \geqq \eta s_k^T B_k s_k, \\ (1-\eta) s_k^T B_k s_k / (s_k^T B_k s_k - y_k^{\ell T} s_k), & \text{otherwise.} \end{cases}$$

A value for η of 0.2 was proposed in [19] and 0.1 in [21]. This technique preserves positive definiteness of B_k even far from the solution, and therefore the subproblems (1.6) are always well posed. Powell's damped BFGS secant method has proved to be very successful computationally (see Hock and Schittkowski [14], for example). However, a proof of local convergence is not known for this algorithm. Given convergence, Powell [18] proves an R-superlinear rate, but only under the assumption of uniform bounds involving the approximate Hessians. Practically, although Powell's damped BFGS update works very well in general, it does sometimes encounter difficulties (see Powell [21]).

Recently, Tapia [25] suggested two new BFGS secant updates based on the structure of the augmented Lagrangian. He was able to prove that the corresponding SQP methods gave local and Q-superlinear convergence in the variable x under the standard assumptions and the assumption that the augmentation parameter ρ was greater than a threshold value $\bar{\rho}$. No guidelines or heuristics were given for choosing the augmentation parameter ρ.

It is worth mentioning that all the above techniques except for Powell's damped update have been restricted primarily to a local framework.

The objective of the current research is to first develop effective guidelines for choosing the augmentation parameter in Tapia's BFGS structured augmented Lagrangian secant algorithm (SALSA). This choice must globally produce a y_k such that $y_k^T s_k > 0$ so that the positive definiteness of approximate Hessians will be maintained. We then describe a practical implementation of SALSA, and make a theoretical and experimental investigation of its behavior.

The bulk of our numerical study of SALSA will be accomplished by using it in an SQP framework in conjunction with a line search on an ℓ_1 merit-function. Because of the demonstrated effectiveness of Powell's damped BFGS algorithm (which we will refer to as PDA) on many problems, we compare SALSA and PDA in this context. However, in order to demonstrate that differences observed are not purely consequences of the line search strategy employed, we also include comparisons of the local versions of both algorithms (i.e., without line search).

Our theoretical results are an advance over what has been shown about SALSA and other augmented Lagrangian-based SQP methods. We analyze the algorithm and its adaptive procedure for choosing the augmentation parameter ρ_k, without assuming that this parameter is chosen greater than some threshold value. We show, under only assumptions A1–A3 and no assumptions whatsoever on the approximate Hessians and the choice of the augmentation parameter, that if SALSA converges, then the convergence in x is R-superlinear. This is similar to, although somewhat stronger than, the result of Powell for PDA, which is mentioned above. Additionally, our theorem implies, as an immediate corollary, the new result that under the standard assumptions only, i.e., no assumptions on the approximate Hessians, whenever the BFGS secant method for unconstrained optimization converges it converges Q-superlinearly.

This paper is organized as follows. In § 2, we briefly present SALSA as Tapia proposed it. In § 3, we discuss some critical issues concerning the globalization and implementation of SALSA and describe a complete algorithm. In particular, we develop a choice for the augmentation parameter ρ, propose a merit function, and present the complete line search algorithm. Section 4 is devoted entirely to demonstrating the convergence rate result discussed above. Our numerical results comparing SALSA and Powell's damped BFGS algorithm are given in § 5. Section 6 contains concluding remarks.

2. The use of structure in the augmented Lagrangian. SALSA was designed to take advantage of the structure present in the Hessian of the augmented Lagrangian function for problem (1.1). By way of motivation, observe that the Hessian of the augmented Lagrangian (1.8) displays significant structure in that there is a clear separation between the first- and second-order information.

Recall that the Lagrangian $\ell(x, \lambda)$ is given by (1.2) and that the augmented Lagrangian $L(x, \lambda, \rho)$ is given by (1.7). We use the superscripts ℓ and L to denote quantities associated with the Lagrangian and the augmented Lagrangian, respectively. The superscript S is used in place of the superscript L when the quantity in question has been derived using the structure of the Hessian of the augmented Lagrangian.

From the definitions of y_k^L and y_k^ℓ (see (1.9) and (1.4)),

$$\begin{aligned} y_k^L &= y_k^\ell + \rho(A_{k+1}h_{k+1} - A_k h_k) \\ &= y_k^\ell + \rho\left(\sum_{i=1}^m h_{k+1}^{(i)} \nabla_x^2 h_{k+1}^{(i)} + A_{k+1}A_{k+1}^T\right) s_k + O(\|s_k\|^2) \\ &= y_k^\ell + \rho A_{k+1}A_{k+1}^T s_k + O(\sigma_k^2), \end{aligned}$$

where in this case we use the superscript (i) to denote the ith component of the vector h_k and

$$\sigma_k = \max\{\|x_{k+1} - x_*\|, \|x_k - x_*\|\}. \tag{2.1}$$

Eliminating the second-order term of σ_k from y_k^L, we have

$$y_k^S = y_k^\ell + \rho A_{k+1}A_{k+1}^T s_k. \tag{2.2}$$

It should be noted that the use of y_k^S in place of y_k^L does not prevent the local analysis for secant methods from being carried out, since the difference between y_k^L and y_k^S is $O(\sigma_k^2)$.

For the sake of completeness, we present the line search SQP structured augmented Lagrangian BFGS secant algorithm, SALSA, in its entirety instead of merely making appropriate changes in Algorithm 1.1.

ALGORITHM 2.1 (SALSA). Given $x_0 \in \mathbf{R}^n$ and a symmetric positive definite matrix $B_0 \in \mathbf{R}^{n \times n}$, for $k = 1, 2, \cdots$, until convergence do

$$\begin{aligned} x_{k+1} &= x_k + \tau_k d_k, \\ \lambda_{k+1} &= \Lambda(x_k, x_{k+1}, B_k), \qquad (2.3) \\ s_k &= x_{k+1} - x_k, \end{aligned}$$

$$y_k^S = \nabla_x \ell(x_{k+1}, \lambda_{k+1}) - \nabla_x \ell(x_k, \lambda_{k+1}) + \rho A_{k+1} A_{k+1}^T s_k, \tag{2.4}$$

$$B_{k+1}^L = B_k^L - \frac{B_k^L s_k s_k^T B_k^L}{s_k^T B_k^L s_k} + \frac{y_k^S y_k^{ST}}{y_k^{ST} s_k}, \tag{2.5}$$

where the line search direction d_k is the solution of the quadratic programming subproblem (1.6) with B_k^L in place of B_k, and the step-length τ_k is chosen to decrease a given merit-function. The matrix B_{k+1}^L is interpreted as an approximation to $\nabla_x^2 L(x_{k+1}, \lambda_{k+1}, \rho)$.

In SALSA, the approximate Hessian of the augmented Lagrangian B_{k+1}^L satisfies the following structured form of the augmented Lagrangian secant equation:

$$B_{k+1}^L s_k = y_k^S = y_k^\ell + \rho A_{k+1} A_{k+1}^T s_k. \tag{2.6}$$

For ρ large enough, the local positivity of $y_k^{ST} s_k$ is guaranteed and consequently the hereditary positive definiteness of B_k^L is achieved. Even globally, $y_k^{ST} s_k$ can be made positive by increasing ρ, as long as $A_{k+1}^T s_k \neq 0$. To see this, note that

$$y_k^{ST} s_k = y_k^{\ell T} s_k + \rho \| A_{k+1}^T s_k \|_2^2. \tag{2.7}$$

However, as discussed in § 3, some back-up strategy is needed to make $y_k^{ST} s_k > 0$ when $A_{k+1}^T s_k$ is numerically zero and $y_k^{\ell T} s_k \leqq 0$.

It is interesting to note that while we have been viewing SALSA as an SQP augmented Lagrangian secant method, it can be equivalently viewed as an SQP structured Lagrangian secant method. To see this recall that $\nabla_x^2 L_* = \nabla^2 \ell_* + \rho A_* A_*^T$; thus it is quite natural to consider B_{k+1}^ℓ defined by

$$B_{k+1}^\ell = B_{k+1}^L - \rho A_{k+1} A_{k+1}^T.$$

Now from (2.6), we see that B_{k+1}^ℓ satisfies the Lagrangian secant equation

$$B_{k+1}^\ell s_k = y_k^\ell. \tag{2.8}$$

Moreover, B_{k+1}^ℓ is positive definite on the null space of A_{k+1}^T, since on this space it coincides with B_{k+1}^L. It also follows that the corresponding quadratic programming subproblem (1.6) using B_{k+1}^ℓ will have the same solution. Hence SALSA can be viewed as an SQP Lagrangian secant method with the highly desirable property that B_{k+1}^ℓ is positive definite on the null space of A_{k+1}^T.

In SALSA the structure in the Hessian of the augmented Lagrangian was utilized only in the definition of y_k^S, but not in the definition of B_k^L. Tapia [25] considered utilizing the structure in both definitions and derived what he called the augmented-scale BFGS secant update. Essentially, he was able to show that this complete use of structure led to cancellations throughout the SQP method and the resulting algorithm could be viewed as an SQP Lagrangian secant method where only the part of the BFGS secant update corresponding to the scale was changed.

Initially, we experimented with the SQP augmented-scale BFGS secant method and found that it does not lend itself to a line search globalization. This is due to the fact that the Hessian approximations are not necessarily positive definite. For this reason, we decided to restrict our attention to SALSA. However, the augmented-scale BFGS secant update may find use in a trust-region globalization.

3. Development of SALSA. In the previous section we discussed why we believe that the SALSA updating procedure, that is, using (2.5) with (2.4), should be a good one. However, several important issues associated with the development of the algorithm SALSA remain to be addressed. In this section we first discuss a weighted form of the augmentation, and then we take up the essential issue of choosing the augmentation parameter ρ. We discuss the issues of subproblem solution, multiplier estimates, and line search, which must be addressed for any SQP algorithm, and finally we give a precise statement of the algorithm. We mention that the current version of the code is given primarily for the purpose of testing the viability of SALSA and performing numerical comparative studies. Further effort is needed to optimize each component of this algorithm.

3.1. Weighted augmentation. The Hessian of the (unweighted) augmentation term $\rho h(x)^T h(x)$ at any feasible point, in particular at a solution x_*, is of the form $\rho A(x)A(x)^T$. If the constraints are badly scaled, then the matrix $A(x)A(x)^T$ may be ill conditioned (here the condition number of a singular matrix is defined to be the ratio of its largest and smallest nonzero singular values) and can have negative effects on the updating process through the use of $A_{k+1}A_{k+1}^T s_k$ in y_k^S. It is natural to scale the constraints by using a weighted augmentation term $\rho h(x)^T W(x)h(x)$, which produces at x_* a well-conditioned Hessian matrix $\rho A(x_*)\,W(x_*)A(x_*)^T$. The matrix $W(x) \in \mathbf{R}^{m\times m}$ is called a weighting matrix and should be positive definite in the area of interest. Under the assumption that $A(x)$ has full rank for all x_k, a good choice for the weighting matrix seems to be

$$W(x) = [A(x)^T A(x)]^{-1}$$

because we can write

$$A_k W_k A_k^T = A_k(A_k^T A_k)^{-1}A_k^T = Y_k Y_k^T,$$

where Y_k is any orthonormal basis for the range space of A_k. Clearly, the matrix $Y_k Y_k^T$ always has unity condition number. Moreover, as long as a weighting matrix $W(x)$ and its inverse are uniformly bounded in norm, all our theoretical results remain valid. Based on the above consideration, we therefore use the matrices $Y_k Y_k^T$ instead of $A_k A_k^T$ in our algorithm. Specifically, we define

$$y_k^S = y_k^\ell + \rho_k Y_{k+1} Y_{k+1}^T s_k. \tag{3.1}$$

In our computational experiments, this weighting technique worked somewhat better on the whole, and we did find examples for which it significantly improved the robustness of the algorithm when compared to the unweighted version.

3.2. Choice of parameter ρ. A fundamental issue in using the augmented Lagrangian in a secant algorithm is the choice of the augmentation parameter ρ, and this is thus an issue for SALSA also. Although, as mentioned in the introduction, any value of ρ greater than the threshold value $\bar{\rho}$ will make the Hessian of the augmented Lagrangian $H_*(\rho) \equiv \nabla_x^2 L_* = \nabla_x \ell_*^2 + \rho A_* A_*^T$ positive definite, $\bar{\rho}$ depends on the unknown matrix $\nabla_x \ell_*^2$.

The practice of choosing a large ρ from the very beginning has proved to be computationally ineffective for the SQP augmented Lagrangian secant method. Not surprisingly, as was also observed by Martinez [15], we found that the same ineffectiveness also exists for the structured version, SALSA.

An alternative approach that we consider here is to choose ρ_k just large enough so that $y_k^{ST} s_k$ is positive. The formulation of SALSA provides a natural framework for doing this. As we can see from the definition of y_k^S in (2.4), ρ can be increased whenever needed to make $y_k^{ST} s_k$ *sufficiently* positive, as long as $A_{k+1}^T s_k \neq 0$. The difficult question here is what is meant by sufficiently positive. Suppose we choose ρ_k just large enough so that $y_k^{ST} s_k = \beta s_k^T s_k$. If β is a very small positive constant, then B_{k+1}^L will be nearly singular (having an eigenvalue less than or equal to β) whenever $y_k^{\ell T} s_k \leqq 0$. If β is reasonably large, then we get a poor approximation to H_* whenever the smallest eigenvalue of the reduced Hessian $Z_*^T H_* Z_*$, where Z_* is an orthonormal basis for the null space of A_*^T, is much less than β (provided that s_k has a significant component in the null space). However, if we impose the condition

$$\frac{y_k^{ST} s_k}{\| Y_{k+1}^T s_k \|^2} \geqq \nu, \tag{3.2}$$

we have a condition on $y_k^{ST} s_k$ that is inactive when s_k is in the null space of A_{k+1}^T and the Hessian of the Lagrangian is positive definite on that null space, and avoids near-singularity of B_{k+1}^L when s_k has a significant range space component $Y_{k+1}^T s_k$. Consequently, as will be shown in Theorem 3.1, imposing the condition (3.2) solves the problem near the solution.

However, when this positive definiteness fails, as it may far from the solution, we argue that this bound should be larger. This is because the term $y_k^S y_k^{ST} / y_k^{ST} s_k$ in the BFGS updating formula (2.5) can get excessively large when $y_k^{ST} s_k$ is small relative to $\|y_k^S\|^2$. To demonstrate this phenomenon, let us suppose that we are in a situation where $y_k^{\ell T} s_k < 0$ and $\| Y_{k+1}^T s_k \| \ll \|y_k^\ell\|$. If ρ_k is chosen such that $y_k^{ST} s_k$ is comparable to $\| Y_{k+1}^T s_k \|^2$, then $y_k^{ST} s_k \ll \|y_k^\ell\|$. On the other hand, the magnitude of $\|y_k^S\|$ (i.e., $\|y_k^\ell + \rho_k Y_{k+1} Y_{k+1}^T s_k \|$) can be as large as the dominant term $\|y_k^\ell\|$. Consequently, the rank-one matrix $y_k^S y_k^{ST} / y_k^{ST} s_k$ can be excessively large, since its unique nonzero eigenvalue is $\|y_k^S\|^2 / y_k^{ST} s_k$. As a result, the newly updated matrix B_{k+1}^L could be badly ill conditioned. To see this, observe that a lower bound for the spectrum condition number of B_{k+1}^L is

$$\frac{y_k^{ST} B_{k+1}^L y_k^S / y_k^{ST} y_k^S}{s_k^T B_{k+1}^L s_k / s_k^T s_k} \geqq \left(\frac{\|y_k^S\| \, \|s_k\|}{y_k^{ST} s_k} \right)^2.$$

In deriving the above estimate, we used the facts $B_{k+1}^L s_k = y_k^S$ and

$$y_k^{ST} B_{k+1}^L y_k^S \geqq (y_k^{ST} y_k^S)^2 / y_k^{ST} s_k.$$

Now it should be clear that the condition number of B_{k+1}^L will be large when $y_k^{ST} s_k$ is small relative to $\|y_k^S\| \, \|s_k\|$. In experiments we have observed algorithm failures due to this behavior. However, these failures were avoided by requiring in addition that $y_k^{ST} s_k \geqq |y_k^\ell s_k|$.

Therefore, combining this condition with (3.2) yields the following strategy for choosing ρ_k at each iteration. Whenever $\| Y_{k+1}^T s_k \|$ is *sufficiently positive*, we choose $\rho_k > 0$ such that

$$y_k^{ST} s_k = y_k^{\ell T} s_k + \rho_k s_k^T Y_{k+1} Y_{k+1}^T s_k \geqq \max\{ |y_k^{\ell T} s_k|, \nu \| Y_{k+1}^T s_k \|^2 \}, \tag{3.3}$$

where ν is a positive constant. The condition of $\| Y_{k+1}^T s_k \|$ being sufficiently positive will be discussed in the next subsection. In that case we need a back-up strategy, which will also be discussed in the next subsection.

It is straightforward to show that (3.3) is equivalent to requiring

$$y_k^{ST} s_k \geqq \max \left\{ \frac{\rho_k}{2}, \nu \right\} \| Y_{k+1}^T s_k \|^2. \tag{3.4}$$

In our implementation, we set $\nu = 0.01$. We choose ρ_k to be the smallest nonnegative value satisfying (3.3), which implies that we choose $\rho_k = 0$ if

$$y_k^{\ell T} s_k \geqq \nu \| Y_{k+1}^T s_k \|^2. \tag{3.5}$$

It can be easily seen that when s_k is in the null space of A_{k+1}^T and x_k is near x_*, which implies $y_k^{\ell T} s_k > 0$ under assumption A3, condition (3.5) will hold. On the other hand, when the step has a significant range space component $Y_{k+1} Y_{k+1}^T s_k$, near-singularity of B_{k+1}^L is avoided because $s_k^T B_{k+1}^L s_k \geqq \nu \| Y_{k+1}^T s_k \|^2$. Moreover, the condition $y_k^{ST} s_k \geqq |y_k^{\ell T} s_k|$ is designed to prevent the deterioration of B_{k+1}^L due to relatively small $y_k^{ST} s_k$. Our computational experiments have shown that this heuristic condition works quite well. In addition, as a result of enforcing (3.3), y_k^S has the following nice property.

THEOREM 3.1. *Under assumptions* A1-A3, *if* ρ_k *is chosen to satisfy* (3.3), *then there is a constant* M_1 *such that*

$$y_k^{ST} s_k \geqq M_1 \| s_k \|^2 \tag{3.6}$$

for all x_k *and* x_{k+1} *sufficiently close to* x_* *and* λ_{k+1} *sufficiently close to* λ_*.

Proof. Let $\hat\rho$ be some value such that $H_*(\hat\rho) = \nabla_x^2 \ell(x_*, \lambda_*) + \hat\rho A(x_*) W(x_*) A(x_*)^T$ is positive definite, and let μ_1 be the smallest eigenvalue of $H_*(\hat\rho)$.

Case 1. $\| Y_{k+1}^T s_k \|^2 \leqq \mu_1 \| s_k \|^2 / (3\hat\rho)$. Then for some constant $C > 0$,

$$\begin{aligned} y_k^{ST} s_k &\geqq y_k^{\ell T} s_k \\ &\geqq s_k^T H_*(\hat\rho) s_k - C\sigma_k \| s_k \|^2 - \hat\rho s_k^T Y_{k+1} Y_{k+1}^T s_k \\ &\geqq (\mu_1 - C\sigma_k) \| s_k \|^2 - \hat\rho \| Y_{k+1}^T s_k \|^2 \\ &\geqq \frac{\mu_1}{3} \| s_k \|^2 \end{aligned}$$

when $\sigma_k = \max\,(\|x_k - x_*\|, \|x_{k+1} - x_*\|, \|\lambda_{k+1} - \lambda_*\|) \leqq \mu_1/(3C)$.

Case 2. $\| Y_{k+1}^T s_k \|^2 > \mu_1 \| s_k \|^2 / (3\hat\rho)$. Then by (3.3)

$$y_k^{ST} s_k \geqq \nu \| Y_{k+1}^T s_k \|^2 \geqq \frac{\nu\mu_1}{3\hat\rho} \| s_k \|^2. \tag{3.7}$$

In either case our result holds with $M_1 = \min\,[\mu_1/3,\ \nu\mu_1/3\hat\rho]$. □

Note that the only property of the matrix Y_{k+1} used in the proof was the existence of $\hat\rho$ such that $H_*(\hat\rho)$ is positive definite. This means that Theorem 3.1 also holds for any choice of y_k such that $y_k - y_k^S = O(\sigma_k \| s_k \|)$, or one using A_{k+1} in place of Y_{k+1}.

A nice feature of this result is that it shows that we can pick ρ_k so that y_k^S acts as though $H_*(\rho_k)$ were positive definite (it satisfies (3.6)) even though we do not know whether we have chosen ρ_k large enough to make $H_*(\rho_k)$ positive definite.

3.3. A back-up strategy. Theorem 3.1 seems to indicate that we have a good strategy for choosing the augmentation parameter and maintaining positive definiteness of B_k in a neighborhood of the solution. In fact, it actually allows us to maintain positive definiteness whenever $Y_{k+1}^T s_k$ is nonzero.

However, in the above strategy for making $y_k^{ST} s_k$ positive, there is one case that the structured augmented Lagrangian approach is incapable of handling: that is, when

$$Y_{k+1}^T s_k = 0 \quad \text{and} \quad y_k^{\ell T} s_k \leqq 0. \tag{3.8}$$

This is analogous to the case $y_k^T s_k \leqq 0$ in unconstrained optimization. We have shown that this will not happen when the current iterate is already close to a solution, but globally this may happen. In addition, if $y_k^{\ell T} s_k \leqq 0$ and $\| Y_{k+1}^T s_k \|$ is not zero but very small, the choice of ρ given by (3.3) would be excessively large. Therefore, in these cases we need a back-up strategy for preserving positive definiteness, and we need a rule for deciding between the back-up strategy and the SALSA update.

A possible option for such a back-up strategy is to just not update, i.e., set $B^L_{k+1} = B^L_k$ whenever the case (3.8) occurs. However, in experiments with this strategy we have observed that once an update is skipped, the algorithm often continues not to update for a number of iterations without much progress, requiring a large number of iterations to solve the problem. The problem with the not-to-update strategy seems to be its sacrifice of a self-correcting mechanism. This sacrifice may cause problems in the following way. Suppose the not-to-update strategy is invoked when the step s_k is very small due to very large elements in the matrix B^L_k as well as small $\|h(x_k)\|$. Because s_k is small, $x_{k+1}(=x_k + s_k)$ will be close to x_k. Since $y^{\ell T}_k s_k < 0$, we would like $s^T_k B^L_{k+1} s_k$ to be small. Instead, the update is skipped and B^L_{k+1} continues to be large. As a result, s_{k+1} is again very small and has a direction close to that of s_k (because $B^L_{k+1} = B^L_k$ and $x_{k+1} \approx x_k$). At the $(k+1)$st iteration the update will be skipped again and this process can be repeated for many steps.

Having been convinced that skipping updates is not a good strategy, we adopt the following back-up strategy. Noting from (3.1) that y^ℓ_k is augmented by a constant times the projection of s_k on the range space of A_{k+1}, it seems natural to use s_k itself whenever its projection on the range space of A_{k+1} is too small. Therefore, whenever (3.5) is violated and

$$\|Y^T_{k+1} s_k\| < \min\{\beta_1, \|s_k\|\}\|s_k\|, \tag{3.9}$$

we replace $Y_{k+1} Y^T_{k+1} s_k$ in (3.1) by s_k. Here $\beta_1 < 1$ is a small positive number. We choose the value $\beta_1 = 0.01$, which seems to work well experimentally. When using this back-up strategy we choose ρ_k such that

$$y^{ST}_k s_k = y^{\ell T}_k s_k + \rho_k s^T_k s_k \geqq \max\left(|y^{\ell T}_k s_k|, \nu\|Y^T_{k+1} s_k\|^2\right) \tag{3.10}$$

is satisfied, which is analogous to condition (3.3).

Condition (3.9) is designed to ensure that the back-up strategy is eventually turned off as x_k approaches x_*. This is due to the fact that $\|Y^T_{k+1} s_k\|$ is of order $O(\|s_k\|^2)$. This is the subject of the following result.

THEOREM 3.2. *Assume* A1–A3. *If condition* (3.9) *holds, x_k and x_{k+1} are sufficiently close to x_* and λ_{k+1} is sufficiently close to λ_*, then condition* (3.5) *is satisfied and therefore the back-up strategy is not selected.*

Proof. Suppose condition (3.9) holds and let $Z_{k+1} \in \mathbf{R}^{n \times (n-m)}$ be such that its columns form an orthonormal basis for the null space of A^T_{k+1}. It follows from $\|Y^T_{k+1} s_k\| < \beta_1 \|s_k\|$ and $s_k = Z_{k+1} Z^T_{k+1} s_k + Y_{k+1} Y^T_{k+1} s_k$ that

$$\|Y^T_{k+1} s_k\|^2 < \frac{\beta_1^2}{1-\beta_1^2}\|Z^T_{k+1} s_k\|^2.$$

Substituting the above into $\|Y^T_{k+1} s_k\| < \|s_k\|^2$, we obtain

$$\|Y^T_{k+1} s_k\| < \|Z^T_{k+1} s_k\|^2 + \|Y^T_{k+1} s_k\|^2 < \frac{\|Z^T_{k+1} s_k\|^2}{1-\beta_1^2}. \tag{3.11}$$

Let

$$\bar{G}_k = \int_0^1 \nabla^2_x \ell(x_k + \tau s_k, \lambda_{k+1})\, d\tau.$$

Then we have

$$\begin{aligned} y_k^{\ell T} s_k = s_k^T \bar{G}_k s_k &= s_k^T Z_{k+1}(Z_{k+1}^T \bar{G}_k Z_{k+1}) Z_{k+1}^T s_k \\ &\quad + s_k^T Y_{k+1}(Y_{k+1}^T \bar{G}_k Y_{k+1}) Y_{k+1}^T s_k + 2 s_k^T Y_{k+1}(Y_{k+1}^T \bar{G}_k Z_{k+1}) Z_{k+1}^T s_k \\ &= s_k^T Z_{k+1}(Z_{k+1}^T \bar{G}_k Z_{k+1}) Z_{k+1}^T s_k + O(\|Z_{k+1}^T s_k\|^3) \end{aligned}$$

by (3.11). By assumption A3, for x_k and x_{k+1} sufficiently close to x_* and λ_{k+1} to λ_*,

$$s_k^T Z_{k+1}(Z_{k+1}^T \bar{G}_k Z_{k+1}) Z_{k+1}^T s_k \geqq \mu \|Z_{k+1}^T s_k\|^2$$

for some constant $\mu > 0$. Therefore, by (3.11),

$$y_k^{\ell T} s_k \geqq \frac{\mu}{2} \|Z_{k+1}^T s_k\|^2 \geqq \frac{\mu}{2}(1-\beta_1^2)\| Y_{k+1}^T s_k \| \geqq \nu \| Y_{k+1}^T s_k \|^2,$$

for $\| Y_{k+1}^T s_k \|$ sufficiently small, and condition (3.5) is satisfied. □

3.4. Subproblem solution and multiplier estimates. Our procedure for computing the solution of (1.6) is as follows. A QR decomposition of A_k is first performed, namely,

$$A_k = (Y_k \ \ Z_k)\begin{pmatrix} R_k \\ 0 \end{pmatrix} = Y_k R_k, \tag{3.12}$$

where $Y_k \in \mathbf{R}^{n \times m}$ is an orthonormal basis for the range space of A_k, $Z_k \in \mathbf{R}^{n \times (n-m)}$ is an orthonormal basis for the null space of A_k^T, and R_k is an m by m upper triangular matrix. The solution d_k of the subproblem (1.6) is given by

$$d_k = Y_k Y_k^T d_k + Z_k Z_k^T d_k, \tag{3.13}$$

where

$$Y_k^T d_k = -R_k^{-T} h_k \quad \text{and} \quad Z_k^T d_k = -(Z_k^T B_k^L Z_k)^{-1} Z_k^T (g_k + B_k^L Y_k Y_k^T d_k). \tag{3.14}$$

The multiplier associated with the QP subproblem (1.6) is

$$\lambda_{k+1}^{\mathrm{QP}} = -(A_k^T A_k)^{-1} A_k^T (g_k + B_k^L d_k). \tag{3.15}$$

We use this multiplier estimate in defining y_k^ℓ. Another possible choice for the multiplier estimate is the least-squares estimate

$$\lambda_{k+1}^{\mathrm{LS}} = -(A_{k+1}^T A_{k+1})^{-1} A_{k+1}^T g_{k+1}. \tag{3.16}$$

However, in our experiments we found that use of this value resulted in significantly more failures than the use of (3.15). Therefore, we will use the QP multiplier estimate to form y_k^ℓ in our numerical tests, that is,

$$y_k^\ell = \nabla_x \ell(x_{k+1}, \lambda_{k+1}^{\mathrm{QP}}) - \nabla_x \ell(x_k, \lambda_{k+1}^{\mathrm{QP}}). \tag{3.17}$$

3.5. Line search. In order to test the viability of SALSA in a line search globalization framework, we need to specify a merit-function for the algorithm. Our purpose here is not to determine the best merit-function, but to use a simple robust function to provide some context for testing our updating strategy. We choose a merit-function of the form

$$\phi(x, w) = f(x) + \sum_{i=1}^{m} w^{(i)} |h^{(i)}(x)|. \tag{3.18}$$

This type of merit-function was first used in an SQP algorithm by Han [13] and was later also used by Powell [19].

Let $\phi_k(\tau)=\phi(x_k+\tau d_k, w_k)$, $\tau \geqq 0$, and let $\phi'_k(0)$ be the directional derivative of $\phi(x, w_k)$ with respect to x in the direction d_k—the solution of subproblem (1.6). Then

$$\phi'_k(0)=g_k^T d_k-\sum_{i=1}^m w^{(i)}|h_k^{(i)}|,$$

which follows from the fact that d_k satisfies the constraints of subproblem (1.6)

$$\nabla h_k^{(i)T} d_k=-h_k^{(i)}, \qquad i=1,2,\cdots,m.$$

It has been shown by Han [13] that a sufficient condition for $\phi'_k(0)<0$ is

$$w_k^{(i)}>|(\lambda_{k+1}^{\mathrm{QP}})^{(i)}| \tag{3.19}$$

for all i, where $\lambda_{k+1}^{\mathrm{QP}}$ is the Lagrange multiplier associated with the kth QP subproblem. Han [13] proves a global convergence result, assuming that (3.19) is eventually satisfied for a constant w. This holds under his conditions if the weights are chosen to be monotone increasing. However, it has been observed that the performance of this merit function is rather sensitive to the choice of the weights w. Too large a w can also slow down convergence. Powell [19] first used in his code VF02AD a strategy that allowed w to fluctuate; more specifically,

$$w_k^{(i)}=\max\{|(\lambda_{k+1}^{\mathrm{QP}})^{(i)}|, 0.5(|(\lambda_{k+1}^{\mathrm{QP}})^{(i)}|+w_{k-1}^{(i)})\}.$$

Though this strategy has been shown [14] to be computationally successful, it does not meet Han's condition for global convergence. Moreover, Chamberlain [3] constructed an example that shows that Powell's strategy of choosing w can lead to cycling instead of convergence.

We performed numerical tests using monotonically increasing weights and found that this strategy resulted in a large number of failures with both Powell's method and SALSA. This was particularly true when we used nonstandard starting points that were far from the solutions or ill-conditioned initial Hessian approximations. It seemed to occur fairly often that an early estimate of the Lagrange multiplier would be much larger than the true multiplier. Then the corresponding large weight, kept large by the monotonicity requirement, would cause the line search to take very short steps, sometimes leading to failure.

In order to have a more meaningful comparison in a realistic environment, we used the following simple nonmonotone strategy. We define at the kth iteration

$$w_k^{(i)}=\mu_k(|\lambda_{k+1}^{\mathrm{QP}})^{(i)}|+\delta), \qquad i=1,2,\cdots,m,$$

where $\mu_k \geqq 1$ and $\delta>0$ (here we choose $\delta=0.0001$) and $\lambda_{k+1}^{\mathrm{QP}}$ is the Lagrange multiplier estimate obtained by solving the quadratic programming subproblem. Although a value of $\mu_k=1$ does give a descent direction (see (3.19)), we found that we were able to take full steps more often if μ_k was chosen large enough so that

$$\phi'_k(0)\leqq -|g_k^T d_k|.$$

Consequently, the formula we used for choosing μ_k was

$$\mu_k=\max\left\{1, 2g_k^T d_k \Big/ \left|\sum_{i=1}^m (|(\lambda_{k+1}^{\mathrm{QP}})^{(i)}|+\delta)h_k^{(i)}\right|\right\}.$$

This is somewhat similar to a condition proposed by Powell [20] in the context of a monotone strategy. Of course, we can make no global convergence guarantees for this nonmonotone strategy, and it is certainly possible that instances of cycling like those discussed by Chamberlain [3] could occur. However, based on our experiments, the likelihood of cycling seems to be extremely low (it was never observed) for equality-constrained problems. In addition, it should be noted that most proofs of convergence involving quasi-Newton methods and merit-functions (except that given in [2] for the reduced Hessian case) assume the boundedness of $\|B_k\|$ or $\|B_k^{-1}\|$, a property which, even locally, does not follow from our analysis of this method. Thus even if we used monotone increasing weights, we would have only a very weak guarantee of global convergence.

A back-tracking strategy is used in our line search to determine a step-length τ_k satisfying the sufficient decrease condition

$$\phi_k(\tau_k) \leqq \phi_k(0) + \alpha\tau_k\phi_k'(0), \tag{3.20}$$

where $0 < \alpha < \frac{1}{2}$. Here we choose $\alpha = 0.1$. We always start from $\tau_k^{(0)} = 1$. If $\tau_k^{(j)}$ satisfies (3.20), we let $\tau_k = \tau_k^{(j)}$; otherwise,

$$\tau_k^{(j+1)} = \max\left\{0.1, \min\left\{0.9, \frac{0.5\phi_k'(0)\tau_k^{(j)}}{\phi_k(0) + \phi_k'(0)\tau_k^{(j)} - \phi_k(\tau_k^{(j)})}\right\}\right\}\tau_k^{(j)}.$$

The formula on the right-hand side comes from a restricted quadratic interpolation. We limit the number of back-trackings to 10; if $j > 10$, we abort the line search and terminate the algorithm. The above back-tracking procedure is basically the one used by Powell [19].

It is well known that the nonsmoothness of the merit-function $\phi(x, w)$ may prevent a step-length of one from being taken near the solution even though it is a good choice. This phenomenon is commonly called the Maratos effect. It is certainly an issue that should be adequately addressed in a production code, but it does not happen very often and we therefore took no specific measures to combat it. The Maratos effect does not appear to have been a major factor in our numerical experiments; in only a very small number of cases was a step-length of less than one taken within the last three iterations of a run and it never happened within the last two iterations.

3.6. Algorithm description. Now we are ready to describe the complete form of SALSA. We suppose that all the quantities involved in the algorithms have already been evaluated before they are used.

ALGORITHM 3.1 (SALSA).

Step 0. Choose positive constants tol > 0, ν, β_1, a positive integer mxiter, $x_0 \in \mathbf{R}^n$, and a symmetric positive definite matrix $B_0^L \in \mathbf{R}^{n\times n}$. Set $k = 0$.

Step 1. If the stopping criterion $\|(Z_k^T g_k, h_k)\|_2 \leqq$ tol is satisfied, exit.

Step 2. If B_k^L is numerically indefinite, stop; otherwise solve the subproblem (1.6) for the search direction d_k and the QP Lagrange multiplier estimate $\lambda_{k+1}^{\text{QP}}$, using (3.13), (3.14), and (3.15).

Step 3. Perform the line search to determine the step-length τ_k. If the number of back-tracking iterations exceeds 10, stop; otherwise, set $x_{k+1} = x_k + \tau_k d_k$ and $s_k = x_{k+1} - x_k$.

Step 4. Calculate y_k^ℓ given by (3.17). If $y_k^{\ell T} s_k \geqq \nu\|Y_{k+1}^T s_k\|^2$, set $y_k^S = y_k^\ell$. Otherwise, set $y_k^S = y_k^\ell + \rho_k v_k$, where

$$\rho_k = (\max\{|y_k^{\ell T} s_k|, \nu\|Y_{k+1}^T s_k\|^2\} - y_k^{\ell T} s_k)/\|v_k\|^2,$$

$$v_k = \begin{cases} Y_{k+1}Y_{k+1}^T s_k, & \text{if } \|Y_{k+1}^T s_k\| \geqq \min\{\beta_1, \|s_k\|\}\|s_k\|, \\ s_k, & \text{otherwise.} \end{cases}$$

Use the updating formula (2.5) to obtain B^L_{k+1}.

Step 5. If $k >$ mxiter, then stop (too many iterations); otherwise, increment k by one and go to Step 1.

Since Powell's damped BFGS method is one of the most efficient methods currently available, for the purpose of comparison we also implemented Powell's damped BFGS method and ran it side by side along with SALSA. Our implementation of Powell's damped BFGS algorithm is the following, and for simplicity we will refer to it as the PD algorithm, or simply PDA.

ALGORITHM 3.2 (PDA). All steps are identical to Algorithm 3.1 (SALSA) except for Step 4, where Powell's damped BFGS update is used.

Evidently, discrepancies in the numerical performance of Algorithms SALSA and PDA should be largely due to the use of the two different updating schemes: the structured augmented Lagrangian BFGS update or Powell's damped BFGS update.

4. Convergence rate of SALSA. Now we consider the convergence rate of the algorithm developed in the previous section. In this paper we will analyze only the local behavior of SALSA. Therefore we will assume that the sequence generated by SALSA converges to a local minimizer satisfying assumptions A1–A3, and that a step-length of one is eventually taken at each iteration. A proof of convergence based on a line search on the merit-function, as in [1], would require more knowledge of merit-functions than currently exists. As already mentioned, augmented Lagrangian quasi-Newton algorithms have been analyzed by Han [12] and Tapia [24], [25] under the assumption that ρ_k is chosen larger than the threshold value and is eventually constant. Their analysis is similar to the theory of Broyden, Dennis, and Moré for unconstrained optimization and establishes that, if x_0 and B_0 are sufficiently good initial approximations, then the sequence $\{(x_k, \lambda_k)\}$ converges to (x_*, λ_*) Q-superlinearly. Actually, Tapia [25] established that $x_k \to x_*$ Q-superlinearly. Because of our weaker and more implementable assumptions on the choice of ρ, we cannot prove local convergence when B_0 is a good enough approximation, but we can prove that if the iterates converge to the solution, they converge R-superlinearly.

We would like our analysis to apply to a wider class of implementations of SALSA than the detailed Algorithm 3.1. To achieve this we will base our analysis on the following generalized version of SALSA, which differs from Algorithm 3.1 in that step-lengths of one are always taken, stopping conditions are removed, and a wider class of augmentation terms and multiplier estimates is allowed.

ALGORITHM 4.1 (Generalized local version of SALSA).

Step 0. Initialize $x_0 \in \mathbf{R}^n$ and a symmetric positive definite matrix $B^L_0 \in \mathbf{R}^{n \times n}$. Set $k = 0$.

Step 1. Solve the subproblem (1.6) for the search direction d_k using (3.13) and (3.14).

Step 2. Set $x_{k+1} = x_k + d_k$.

Step 3. Choose the matrix $\hat{A}_{k+1} = A_{k+1} W^{1/2}_{k+1} + O(\|s_k\|)$, where W_{k+1} is taken from a bounded set of positive definite matrices whose inverses are also bounded.

Step 4. Calculate $y^\ell_k = \nabla_x \ell(x_{k+1}, \lambda_{k+1}) - \nabla_x \ell(x_k, \lambda_{k+1})$. If $y^{\ell T}_k s_k \geqq \nu \|\hat{A}^T_{k+1} s_k\|^2$, set $y^S_k = y^\ell_k$. Otherwise, set $y^S_k = y^\ell_k + \rho_k v_k$, where

$$\rho_k = (\max\{|y^{\ell T}_k s_k|, \nu \|\hat{A}^T_{k+1} s_k\|^2\} - y^{\ell T}_k s_k) / \|v_k\|^2,$$

$$v_k = \begin{cases} \hat{A}_{k+1} \hat{A}^T_{k+1} s_k, & \text{if } \|\hat{A}^T_{k+1} s_k\| \geqq \min\{\beta_1, \|s_k\|\} \|s_k\|, \\ s_k, & \text{otherwise.} \end{cases}$$

Use the updating formula (2.5) to obtain B^L_{k+1}.
Step 5. Increment k by one and go to Step 1.

Note that we do not specify the Lagrange multiplier estimate λ_k in y^ℓ_{k+1} in Step 4 of the algorithm; however in theory we will require that $\lambda_k \to \lambda_*$. (For some choices of multiplier estimate such as λ_k^{LS}, convergence of the multipliers is a consequence of convergence of $\{x_k\}$, but this is not immediate for λ_k^{QP}.) The form of $\hat{A}_k$ allows many possible choices for y^S_k depending on the choice of W_{k+1} (see § 3.1). It is easy to see that the following choices of y^S_k are of the specified form, $y^\ell_k + \rho_k A_{k+1} W_{k+1} A^T_{k+1} s_k + O(\|s_k\|^2)$, and are thus covered by our analysis:

$$y^S_k = y^\ell_k + \rho_k Y_{k+1} Y^T_{k+1} s_k,$$

$$y^S_k = y^\ell_k + \rho_k A_{k+1} A^T_{k+1} s_k,$$

$$y^S_k = y^\ell_k + \rho_k Y_k Y^T_k s_k,$$

$$y^S_k = y^\ell_k + \rho_k A_k (h_{k+1} - h_k).$$

In the analysis to follow, we will use y_k in place of y^S_k and B_k in place of B^L_k for simplicity.

The main purpose of this section will be to prove the following result.

THEOREM 4.1. *Assume that the sequence* $\{x_k\}$ *is generated by Algorithm* 4.1, *and assumptions* A1-A3 *hold. If* $x_k \to x_*$ *and* $\lambda_k \to \lambda_*$, *then* $x_k \to x_*$ *R-superlinearly.*

In order to prove this convergence theorem, however, we first define some useful quantities and prove the intermediate results, Lemmas 4.1-4.4. After proving the theorem we will then point out an interesting application to unconstrained optimization. Note that Theorem 4.1 is similar to the rate of convergence result proved by Powell for his damped algorithm, except that this result makes no boundedness assumptions on the approximate Hessians. Our analysis uses some of the techniques developed by Powell in his proof.

By assumptions A1-A3, we know that there is a value $\hat{\rho} \geqq 0$ such that the matrix $\nabla^2_x L(x_*, \lambda_*, \hat{\rho})$ is positive definite. Given the uniform boundedness of $\{W^{-1}_{k+1}\}$, this value may also be chosen so that

$$s^T_k(y^\ell_k + \hat{\rho}\hat{A}_{k+1}\hat{A}^T_{k+1}s_k) > 0 \tag{4.1}$$

for x_k, x_{k+1}, and λ_k sufficiently close to their solution values. For purposes of analysis, we select one such $\hat{\rho}$ and we define the matrix $H_* = \nabla^2_x L(x_*, \lambda_*, \hat{\rho})$, which will be used as a weighting matrix. We define two quantities which measure the accuracy of B_k along the step direction s_k: the ratio of quadratic forms,

$$q_k = \frac{s^T_k B_k s_k}{s^T_k H_* s_k} \tag{4.2}$$

and

$$\cos\theta_k = \frac{s^T_k B_k s_k}{\|H^{1/2}_* s_k\| \, \|H^{-1/2}_* B_k s_k\|}, \tag{4.3}$$

the cosine of the angle between $B_k s_k$ and $H_* s_k$, measured in the $H^{-1/2}_*$ weighted norm. These two quantities, which ideally have value one, thus measure how close the magnitude and direction of $B_k s_k$ correspond to the magnitude and direction of $H_* s_k$. We now show that these two quantities provide rough bounds on the ratio of $\|s_k\|$ to the error, and on the ratio of successive errors. We will also use the notation $e_k = x_k - x_*$ in what follows.

LEMMA 4.1. *Given assumptions* A1–A3, *there exist constants* γ_1 *and* γ_2 *such that if* x_k *is sufficiently close to* x_*, *and* s_k *solves* (1.6), *then*

$$\gamma_1\left(1+\frac{q_k}{\cos\theta_k}\right)^{-1} \leqq \frac{\|s_k\|}{\|e_k\|} \leqq \gamma_2\left(\frac{1}{\cos\theta_k}+\frac{1}{q_k}\right) \tag{4.4}$$

and

$$\frac{\|e_{k+1}\|}{\|e_k\|} \leqq 1+\gamma_2\left(\frac{1}{\cos\theta_k}+\frac{1}{q_k}\right). \tag{4.5}$$

Proof. By the way the step is computed, $\|B_k s_k\| \geqq \|Z_k^T B_k s_k\| = \|Z_k^T g_k\|$. Therefore,

$$\begin{aligned}\|s_k\| &\geqq \frac{\|s_k\|}{\|B_k s_k\|}\,\|Z_k^T g_k\| \\ &= \frac{s_k^T B_k s_k}{\|B_k s_k\|\,\|s_k\|}\,\frac{s_k^T s_k}{s_k^T B_k s_k}\,\|Z_k^T g_k\| \\ &\geqq \gamma_1'\,\frac{s_k^T B_k s_k}{\|H_*^{-1/2} B_k s_k\|\,\|H_*^{1/2} s_k\|}\,\frac{s_k^T H_* s_k}{s_k^T B_k s_k}\,\|Z_k^T g_k\|\end{aligned}$$

for some constant γ_1', since H_* is positive definite. Thus

$$\|s_k\| \geqq \gamma_1'\,\frac{\cos\theta_k}{q_k}\,\|Z_k^T g_k\|.$$

Looking at the normal component of the step we see that

$$\|s_k\| \geqq \|A_k(A_k^T A_k)^{-1} A_k^T s_k\| = \|-A_k(A_k^T A_k)^{-1} h_k\| \geqq \hat{\gamma}_1 \|h_k\|$$

for some constant $\hat{\gamma}_1$. Then, in the neighborhood of a minimizer satisfying assumptions A1–A3, we have

$$\begin{aligned}\|x_k - x_*\| &\leqq \gamma(\|Z_k^T g_k\| + \|h_k\|), \\ &\leqq \frac{\gamma q_k}{\gamma_1' \cos\theta_k}\,\|s_k\| + \frac{\gamma}{\hat{\gamma}_1}\,\|s_k\|,\end{aligned}$$

and the left inequality of (4.4) follows immediately.

To establish the other side note that

$$\begin{aligned}s_k^T s_k &= \frac{s_k^T s_k}{s_k^T B_k s_k}\, s_k^T(Z_k Z_k^T B_k s_k + Y_k Y_k^T B_k s_k) \\ &\leqq \frac{s_k^T s_k}{s_k^T B_k s_k}\,(\|s_k\|\,\|Z_k^T g_k\| + \|Y_k^T s_k\|\,\|B_k s_k\|).\end{aligned}$$

Therefore

$$\|s_k\| \leqq \frac{s_k^T s_k}{s_k^T B_k s_k}\,\|Z_k^T g_k\| + \hat{\gamma}_2\,\frac{\|B_k s_k\|\,\|s_k\|}{s_k^T B_k s_k}\,\|h_k\|,$$

and since H_* is positive definite,

$$\|s_k\| \leqq \gamma_2'\left(\frac{s_k^T H_* s_k}{s_k^T B_k s_k} + \frac{\|H_*^{-1/2} B_k s_k\|\,\|H_*^{1/2} s_k\|}{s_k^T B_k s_k}\right)(\|Z_k^T g_k\| + \|h_k\|),$$

from which the right inequality follows immediately. Inequality (4.5) follows from (4.4) upon noting that, by the triangle inequality,

$$\frac{\|e_{k+1}\|}{\|e_k\|} \leqq 1 + \frac{\|s_k\|}{\|e_k\|}. \qquad \square$$

Actually, the previous lemma could have been proved with any positive definite matrix replacing H_* in the definition of q_k and $\cos\theta_k$. However, in the next lemma the use of H_* is essential to establishing the more precise result that if q_k and $\cos\theta_k$ are sufficiently close to 1, then the ratio of successive errors can be made arbitrarily small.

LEMMA 4.2. *Under the conditions of Lemma* 4.1,

$$\|e_{k+1}\| = O(\|e_k\|^2 + \|Z_k^T(B_k - H_*)s_k\|) \tag{4.6}$$

$$= O\left(\|e_k\|^2 + \left(\frac{q_k^2}{\cos^2\theta_k} - 2q_k + 1\right)^{1/2} \|s_k\|\right). \tag{4.7}$$

Proof. First we decompose the error into two parts and consider each separately. Observe that

$$\begin{aligned}
\|Z_k^T H_* e_{k+1}\| &= \|Z_k^T H_*(e_k + s_k)\| \\
&= \|Z_k^T[\nabla L(x_k, \lambda_*, \hat{\rho}) - \nabla L(x_*, \lambda_*, \hat{\rho}) + B_k s_k] \\
&\quad + Z_k^T(H_* - B_k)s_k\| + O(\|e_k\|^2) \\
&= \|Z_k^T[\nabla L(x_k, \lambda_*, \hat{\rho}) - g_k + (H_* - B_k)s_k]\| + O(\|e_k\|^2) \\
&\leqq \|Z_k^T(H_* - B_k)s_k\| + O(\|e_k\|^2).
\end{aligned}$$

The range space component of the error is given by

$$A_k^T e_{k+1} = A_k^T e_k + A_k^T s_k \tag{4.8}$$

$$= h_k + O(\|e_k\|^2) - h_k = O(\|e_k\|^2). \tag{4.9}$$

The total error is related to these two parts by

$$\|e_{k+1}\| = \left\| \begin{bmatrix} Z_k^T H_* \\ A_k^T \end{bmatrix}^{-1} \begin{bmatrix} Z_k^T H_* e_{k+1} \\ A_k^T e_{k+1} \end{bmatrix} \right\|,$$

and by assumptions A1–A3 the matrix

$$\begin{bmatrix} Z_k^T H_* \\ A_k^T \end{bmatrix}^{-1} = [Z_k(Z_k^T H_* Z_k)^{-1} \quad H_*^{-1} A_k (A_k^T H_*^{-1} A_k)^{-1}]$$

is bounded for all x_k in some neighborhood of x_*. Therefore

$$\|e_{k+1}\| = O\left(\left\| \begin{matrix} Z_k^T H_* e_{k+1} \\ A_k^T e_{k+1} \end{matrix} \right\| \right) = O(\|e_k\|^2 + \|Z_k^T(B_k - H_*)s_k\|),$$

which is just (4.6).

To establish (4.7), note that

$$\frac{\|H_*^{-1/2}(B_k - H_*)s_k\|^2}{\|H_*^{1/2} s_k\|^2} = \frac{q_k^2}{\cos^2\theta_k} - 2q_k + 1,$$

which, since H_* is nonsingular, implies that the right-hand side of (4.6) is of the same order as the right-hand side of (4.7). $\square$

Having established the effect of the quantities q_k and $\cos\theta_k$ on the length of the computed step and the error at the next point, we now consider the issue of how these two key quantities are related to the BFGS update. To that end we define, for any positive definite matrix B, the quantity

$$\psi(B) = \operatorname{trace}(H_*^{-1/2} B H_*^{-1/2}) - \log\det(H_*^{-1/2} B H_*^{-1/2}),$$

which may be considered as a measure of the deviation of B from H_*. Note that ψ is a strictly convex function over the set of positive definite matrices, and it has a unique minimizer at $B = H_*$, as is discussed by Byrd and Nocedal [1].

We now show that $\{\rho_k\}$ is bounded and that the update has an important self-correcting property with respect to ψ. Close to the solution if q_k or θ_k deviates significantly from 1, and if s_k is close to the null space, then $\psi(B)$ is decreased (i.e., B_{k+1} is closer to H_*). The self-correction relation (4.10) established below is analogous to the one of Lemma 7 in [18] except that it uses the Ψ function in a manner similar to equation (2.9) of [1] instead of a weighted Frobenius norm.

LEMMA 4.3. *If* A1–A3 *are satisfied, then for all x_k and x_{k+1} sufficiently close to x_* and all λ_{k+1} sufficiently close to λ_*, there exists an upper bound for the augmentation parameter ρ_k chosen by Algorithm* 4.1. *In addition, for any bounded choice of ρ_k, if B_k is positive definite, the updated matrix produced by the algorithm satisfies*

$$\psi(B_{k+1}) \leqq \psi(B_k) - \frac{q_k}{\cos^2\theta_k} + \log q_k + 1 + \gamma_4\sigma_k + \gamma_3\left(\frac{\|h_k\|}{\|s_k\|}\right)^2, \tag{4.10}$$

where $\sigma_k = \max\{\|e_k\|, \|e_{k+1}\|, \|\lambda_{k+1} - \lambda_*\|\}$, and γ_3 and γ_4 are constants.

Proof. By Theorem 3.2, sufficiently close to the solution, the back-up strategy is not used, and the value of ρ_k chosen in Step 4 of Algorithm 4.1 is the smallest value satisfying (3.3), or equivalently, (3.4). Since the value $\hat\rho$ is such that sufficiently close to x_* (4.1) holds, then it follows that a value of ρ_k as large as $2\hat\rho + \nu$ will satisfy (3.4).

By the definition of ψ, and since $\det(B_{k+1}) = (y_k^T s_k / s_k^T B_k s_k)\det(B_k)$,

$$\psi(B_{k+1}) = \psi(B_k) + \operatorname{trace}\left(H_*^{-1/2}\left(-\frac{B_k s_k s_k^T B_k}{s_k^T B_k s_k} + \frac{y_k y_k^T}{y_k^T s_k}\right)H_*^{-1/2}\right) - \log\left(\frac{y_k^T s_k}{s_k^T B_k s_k}\right) \tag{4.11}$$

$$= \psi(B_k) - \frac{\|H_*^{-1/2} B_k s_k\|^2}{s_k^T B_k s_k} + \frac{y_k^T H_*^{-1} y_k}{y_k^T s_k} - \log\frac{y_k^T s_k}{s_k^T H_* s_k} + \log\frac{s_k^T B_k s_k}{s_k^T H_* s_k}. \tag{4.12}$$

By Steps 2 and 3 of the algorithm,

$$\begin{aligned} y_k &= H_* s_k - \hat\rho A_{k+1} A_{k+1}^T s_k + \rho_k \hat A_{k+1} \hat A_{k+1}^T s_k + O(\sigma_k\|s_k\|) \\ &= H_* s_k + A_{k+1}(\rho_k W_{k+1} - \hat\rho I) A_{k+1}^T s_k + O(\sigma_k\|s_k\|). \end{aligned}$$

This means that

$$\begin{aligned} y_k^T H_*^{-1} y_k &= y_k^T s_k + y_k^T H_*^{-1} A_{k+1}(\rho_k W_{k+1} - \hat\rho I) A_{k+1}^T s_k + O(\sigma_k\|s_k\|^2) \\ &= y_k^T s_k + s_k^T A_{k-1}(\rho_k W_{k+1} - \hat\rho I) A_{k+1}^T s_k \\ &\quad + s_k^T A_{k+1}(\rho_k W_{k+1} - \hat\rho I) A_{k+1}^T H_*^{-1} A_{k+1}(\rho_k W_{k+1} - \hat\rho I) \\ &\quad \cdot A_{k+1}^T s_k + O(\sigma_k\|s_k\|^2), \end{aligned}$$

so that using Theorem 3.1, the fact that $A_{k+1}^T s_k = h_k + O(\sigma_k \|s_k\|)$, and the uniform bound on W_{k+1},

$$\frac{y_k^T H_*^{-1} y_k}{y_k^T s_k} = 1 + O\left(\frac{\|h_k\|}{\|s_k\|}\right)^2 + O(\sigma_k). \tag{4.13}$$

In addition,

$$y_k^T s_k = s_k^T H_* s_k + s_k^T A_{k+1}(\rho_k W_{k+1} - \hat{\rho} I) A_{k+1}^T s_k + O(\sigma_k \|s_k\|^2),$$

so that

$$-\log \frac{y_k^T s_k}{s_k^T H_* s_k} = -\log\left[1 - O\left(\frac{\|h_k\|}{\|s_k\|}\right)^2 + O(\sigma_k)\right] \tag{4.14}$$

$$= O\left(\frac{\|h_k\|}{\|s_k\|}\right)^2 + O(\sigma_k) \tag{4.15}$$

for σ_k and $\|h_k\|/\|s_k\|$ sufficiently small. Since, by Theorem 3.1, $y_k^T s_k / s_k^T H_* s_k$ is bounded away from zero, (4.15) also holds if either σ_k or $\|h_k\|/\|s_k\|$ are not small. Substituting (4.13) and (4.15) into (4.12) and using (4.2) and (4.3) we get

$$\psi_{k+1} \leqq \psi_k - \frac{q_k}{\cos^2 \theta_k} + \log q_k + 1 + O(\sigma_k) + O\left(\frac{\|h_k\|}{\|s_k\|}\right)^2. \qquad \square$$

To analyze the iterates produced by the algorithm we would like bounds on the ratios $\|s_j\|/\|e_j\|$ and $\|e_{j+1}\|/\|e_j\|$. Such bounds would hold at each iterate if we had bounds on the quantities $\|B_k\|$ and $\|Z_k^T B_k^{-1} Z_k\|$, as is shown in [18], but we have not assumed and cannot establish such bounds on B_k. However, the self-correcting property of Lemma 4.3 based on the departure of q_k and $\cos \theta_k$ from 1, can be used together with the bounds in Lemma 4.1 to bound the average behavior of any large subset of the iterates.

LEMMA 4.4. *Assume that the sequence $\{x_k\}$ is generated by Algorithm* 4.1 *and that assumptions* A1–A3 *hold. If $x_k \to x_*$ and $\lambda_k \to \lambda_*$, then there is a constant β such that for any $k > 0$ and any subset S of $\{1, \cdots, k\}$,*

$$\left[\prod_{j \in S} \frac{\|e_{j+1}\|}{\|e_j\|}\right] < \beta^k. \tag{4.16}$$

In addition, for any $p \in (0, 1)$ there are constants β_1 and β_2 such that for any $k > 0$ the set

$$\mathcal{J}_k = \left\{ j \in [1, k]: \beta_1 \leqq \frac{\|s_j\|}{\|e_j\|} \leqq \beta_2 \right\} \tag{4.17}$$

contains at least pk elements.

Proof. Summing up the recursion (4.10) established in the previous lemma, we have that

$$0 \leqq \psi_{k+1} \leqq \psi_0 + \sum_{j=0}^{k} \left[-\frac{q_j}{\cos^2 \theta_j} + \log q_j + 1 + \gamma_4 \sigma_j + \gamma_3 \left(\frac{\|h_j\|}{\|s_j\|}\right)^2 \right].$$

Since $\|A(x)\|$ is uniformly bounded near x_*, the quantity $\|h_k\|/\|s_k\|$ is bounded above for all k so that

$$0 \leqq \sum_{j=0}^{k} \left[-\frac{q_j}{\cos^2 \theta_j} + \log q_j + 1 \right] + k\gamma',$$

for some constant γ'. Alternatively,

$$\sum_{j=0}^{k}\left[\frac{q_j}{\cos^2\theta_j}-\log q_j-1\right]\leqq k\gamma'. \tag{4.18}$$

Now note that by Lemma 4.1 for any j (since we may assume without loss of generality that $\gamma_2\geqq 1$),

$$\begin{aligned}
\log\frac{\|e_{j+1}\|}{\|e_j\|}&\leqq\log\gamma_2\left(1+\frac{1}{\cos\theta_j}+\frac{1}{q_j}\right)\\
&=\log\gamma^2+\log\left(q_j+\frac{q_j}{\cos\theta_j}+1\right)-\log q_j\\
&\leqq\log\gamma_2+\frac{q_j}{\cos\theta_j}+q_j-\log q_j\\
&\leqq\log\gamma_2+\frac{q_j}{\cos\theta_j}+2q_j-3\log q_j\\
&\leqq\log\gamma_2+3\left(\frac{q_j}{\cos\theta_j}-\log q_j\right).
\end{aligned}$$

Therefore,

$$\begin{aligned}
\sum_{j\in S}\log\frac{\|e_{j+1}\|}{\|e_j\|}&\leqq 3\sum_{j\in S}\left[\frac{q_j}{\cos^2\theta_j}-\log q_j-1\right]+k(3+\log\gamma_2)\\
&\leqq 3\sum_{j=0}^{k}\left[\frac{q_j}{\cos^2\theta_j}-\log q_j-1\right]+k(3+\log\gamma_2)\\
&\leqq(3\gamma'+3+\log\gamma_2)k
\end{aligned}$$

by the fact that all terms in the sum are nonnegative and by (4.18). The nonnegativity of the terms in the sum follows from the fact that

$$\frac{q_j}{\cos^2\theta_j}-\log q_j-1=(-\log\cos^2\theta_j)+\left(\frac{q_j}{\cos^2\theta_j}-1-\log\frac{q_j}{\cos^2\theta_j}\right), \tag{4.19}$$

and, by the properties of the logarithm, both expressions in parentheses are nonnegative.

The first result follows by taking the exponential of both sides of (4.19), and letting $\beta=\gamma_2\,e^{e\gamma'+3}$. To establish the second result, we apply to (4.18) the same argument as in the proof of Theorem 2.1 of Byrd and Nocedal [1]. The relation (4.18) implies that for any k, at least pk of the (nonnegative) terms in the sum are less than or equal to $\gamma'/(1-p)$. For these terms (4.19) implies a positive lower bound on $\cos\theta_j$ and upper and lower bounds on q_j. Then the existence of the constants β_1 and β_2 in (4.17) follows from Lemma 4.1. □

Now we are ready to prove our main result, which we restate here.

THEOREM 4.1. *Assume that the sequence* $\{x_k\}$ *is generated by Algorithm* 4.1 *and assumptions* A1–A3 *hold. If* $x_k\to x_*$ *and* $\lambda_k\to\lambda_*$, *then* $x_k\to x_*$ *R-superlinearly.*

Proof. Suppose that the convergence is not R-superlinear. Then there exists a positive constant r and a subsequence $\mathcal{K}$ such that

$$\|e_k\|>r^k\quad\text{for all } k\in\mathcal{K}. \tag{4.20}$$

We will derive a contradiction from this assumption. Consider the recursion established in Lemma 4.3:

$$\psi_{k+1} \leqq \psi_k - \frac{q_k}{\cos^2 \theta_k} + \log q_k + 1 + \gamma_4 \sigma_k + \gamma_3 \left(\frac{\|h_k\|}{\|s_k\|} \right)^2 .$$

Let

$$\mathcal{P}_k = \left\{ j \in [1, k] : \frac{\|h_j\|}{\|s_j\|} \geqq \sqrt{\frac{\sigma_j}{\gamma_3}} \right\} \tag{4.21}$$

and let $\pi_k = \frac{1}{k} |\mathcal{P}_k|$, where $|\cdot|$ denotes cardinality.

Case 1. $\{\pi_k\}_{k \in \mathcal{K}}$ converges to 0.

Note that for $k \in \mathcal{K}$, Lemma 4.3 implies that

$$0 \leqq \psi_{k+1} \leqq \psi_0 + \sum_{j=0}^{k} \left[-\frac{q_j}{\cos^2 \theta_j} + \log q_j + 1 + \gamma_4 \sigma_j + \sigma_j \right] + k \pi_k \gamma_5$$

for a constant γ_5, since $\|h_j\| / \|s_j\|$ is uniformly bounded above. Therefore,

$$\frac{1}{k} \sum_{j=0}^{k} \left[\frac{q_j}{\cos^2 \theta_j} - \log q_j - 1 \right] \leqq \frac{1}{k} \psi_0 + \frac{1}{k} \sum_{j=0}^{k} (\gamma_4 + 1) \sigma_j + \pi_k \gamma_5 . \tag{4.22}$$

Since we are assuming that $\{e_k\}$ and a subsequence of $\{\pi_k\}$ converge to 0, the right-hand side and, thus, the left-hand side, of (4.22) converge to 0 for this subsequence. Therefore for any $\delta > 0$ there exists k_0 such that if $k > k_0$ and $k \in \mathcal{K}$, then

$$\frac{1}{k} \sum_{j=0}^{k} \left[\frac{q_j}{\cos^2 \theta_j} - \log q_j - 1 \right] \leqq \frac{\delta}{2} .$$

Since each summand is nonnegative, this implies that $q_j / \cos^2 \theta_j - \log q_j - 1 \leqq \delta$ for at least $k/2$ values of $j \leqq k$.

Now note that

$$\frac{q_j}{\cos^2 \theta_j} - \log q_j - 1 = \left[\frac{q_j}{\cos^2 \theta_j} - 1 - \log \frac{q_j}{\cos^2 \theta_j} \right] + \left[-\log \cos^2 \theta_j \right]$$

and both quantities in square brackets are nonnegative, so that by choosing δ sufficiently small we can make $|q_j - 1|$ and $1 - \cos \theta_j$ arbitrarily small for half the iterates. By Lemma 4.1 the quantity $\|s_j\| / \|e_j\|$ is bounded above for those iterates. Now consider (4.7), and note that the quantity $(q_k^2 / \cos^2 \theta_k - 2 q_k + 1)^{1/2}$ is zero when $q_k = \cos \theta_k = 1$ and is continuous at that point; so by (4.7), $\|e_{j+1}\| / \|e_j\|$ can be made arbitrarily small for those iterates.

Therefore, we have that for any $\varepsilon > 0$ there exists k_0 such that if $k > k_0$, $k \in \mathcal{K}$, then $\|e_{j+1}\| / \|e_j\| < \varepsilon$ for $k/2$ values of $j \leqq k$. Let $S = \{ j \leqq k : \|e_{j+1}\| / \|e_j\| < \varepsilon \}$. This implies that

$$\prod_{j=1}^{k} \frac{\|e_{j+1}\|}{\|e_j\|} = \prod_{j \in S} \frac{\|e_{j+1}\|}{\|e_j\|} \prod_{j \notin S} \frac{\|e_{j+1}\|}{\|e_j\|} \leqq \varepsilon^{k/2} \beta^k ,$$

using the bound (4.16).

By choosing ε small enough we see that $(\prod_{j=1}^{k} (\|e_{j+1}\| / \|e_j\|))^{1/k}$ is arbitrarily small for all sufficiently large $k \in \mathcal{K}$, thus contradicting (4.20) in Case 1.

Case 2. There is an infinite subset $\mathcal{K}' \subset \mathcal{K}$ and a constant $\hat{\pi} > 0$ such that $\pi_k \geqq \hat{\pi}$ for all $k \in \mathcal{K}'$.

Apply Lemma 4.4 with $p > 1 - (\hat{\pi}/2)$ (note that $\hat{\pi} \leqq 1$). Consider $k \in \mathcal{K}'$ and $\mathcal{J}_k$, the set of iterates defined by (4.17). Now define the set

$$\mathcal{T}_k = \{ j \in \mathcal{P}_k \cap \mathcal{J}_k : j - 1 \in \mathcal{J}_k \}.$$

The number of elements in $\mathcal{P}_k$ that are not in $\mathcal{T}_k$ is no more than the number of indices $j \leqq k$ such that j or $j-1$ is not in $\mathcal{J}_k$, which is at most twice the cardinality of the set $[1, k] - \mathcal{J}_k$. Therefore,

$$\begin{aligned} |\mathcal{T}_k| &\geqq |\mathcal{P}_k| - 2(k - |\mathcal{J}_k|) \\ &\geqq \hat{\pi}k - 2(k - pk) \\ &= (\hat{\pi} + 2p - 2)k = \tau k, \end{aligned} \tag{4.23}$$

where $\tau = \hat{\pi} + 2p - 2$ is positive by our choice of p.

For any $j \in \mathcal{T}_k$, by (4.17) and (4.21),

$$\|e_j\|^{3/2} \leqq \frac{1}{\beta_1} \|s_j\| \, \|e_j\|^{1/2} \leqq \frac{\sqrt{\gamma^3}}{\beta_1} \|h_j\|.$$

Expanding this h_j, we get

$$\|h(x_j)\| \leqq \|h(x_{j-1}) + A_{j-1}^T s_{j-1}\| + \gamma_6 \|s_{j-1}\|^2 = \gamma_6 \|s_{j-1}\|^2$$

for some constant γ_6. Applying (4.17) at $j-1$ gives

$$\|e_j\|^{3/2} \leqq \frac{\sqrt{\gamma_3}\,\gamma_6}{\beta_1} \|s_{j-1}\|^2 \leqq \frac{\sqrt{\gamma_3}\,\gamma_6 \beta_2^2}{\beta_1} \|e_{j-1}\|^2.$$

Thus for the at least τk indices in $\mathcal{T}_k$, we have

$$\|e_j\| \leqq \gamma_7 \|e_{j-1}\|^{4/3}, \tag{4.24}$$

where $\gamma_7 = (\sqrt{\gamma_3}\,\gamma_6 \beta_2^2 / \beta_1)^{2/3}$.

Now since the sequence converges we can choose k_0 so that $\gamma_7 \|e_j\|^{1/3} \leqq (r/2\beta)^{1/\tau}$ for all $j \geqq k_0 - 1$, where r is as in (4.20) and, without loss of generality, may be assumed to satisfy $r/2\beta < r/2 < 1$. Therefore, for any $k > k_0$ such that $k \in \mathcal{K}'$, we have, using Lemma 4.4, (4.24), and (4.23),

$$\begin{aligned} \|e_k\| &= \|e_{k_0-1}\| \prod_{j=k_0}^{k} \frac{\|e_j\|}{\|e_{j-1}\|} \\ &= \|e_{k_0-1}\| \prod_{j \in [k_0,k] - \mathcal{T}_k} \frac{\|e_j\|}{\|e_{j-1}\|} \prod_{j \in \mathcal{T}_k \cap [k_0,k]} \frac{\|e_j\|}{\|e_{j-1}\|} \\ &\leqq \|e_{k_0-1}\| \beta^k \prod_{j \in \mathcal{T}_k \cap [k_0,k]} (\gamma_7 \|e_{j-1}\|^{1/3}) \\ &\leqq \|e_{k_0-1}\| \beta^k \left(\frac{r}{2\beta} \right)^{1/\tau |\mathcal{T}_k \cap [k_0,k]|} \\ &\leqq \|e_{k_0-1}\| \beta^k \left(\frac{r}{2\beta} \right)^{k - k_0/\tau} \\ &\leqq \|e_{k_0-1}\| \left(\frac{r}{2\beta} \right)^{-k_0/\tau} \left(\frac{r}{2} \right)^k. \end{aligned}$$

For k sufficiently large this violates the assumption (4.20) for Case 2. Thus the convergence must be R-superlinear. □

Although we have assumed in Theorem 4.1 that both sequences $\{x_k\}$ and $\{\lambda_k\}$ are convergent, it is interesting to note that if λ_k is given by the least-squares multiplier estimate (3.16), then convergence of $\{\lambda_k\}$ follows from convergence of $\{x_k\}$, so the assumption of multiplier convergence is not needed in Theorem 4.1.

Theorem 4.1 establishes an R-superlinear rate of convergence, and we do not now see any way to strengthen the result to show Q-superlinear convergence for our algorithm. However, if we instead choose ρ_k to be fixed and sufficiently large, we can prove the following.

COROLLARY 4.1. *Consider a modification of Algorithm* 4.1, *where in Step* 4, *for sufficiently large k, ρ_k is chosen to be equal to a constant greater than $\hat{\rho}$ satisfying* (3.3) *for all large k. Then if $x_k \to x_*$ and $\lambda_k \to \lambda_*$, it follows that $x_k \to x_*$ Q-superlinearly. That is,*

$$\frac{\|x_{k+1}-x_*\|}{\|x_k-x_*\|} \to 0. \tag{4.25}$$

Proof. Note that as long as ρ_k satisfies (3.3) and is bounded, then Lemma 4.3 holds, and thus Theorem 4.1 holds also. Then by Theorem 4.1 we still have R-superlinear convergence, which implies

$$\sum_{k=0}^{\infty} \|x_k-x_*\| < \infty. \tag{4.26}$$

However, the modified algorithm is equivalent, for large k, to the fixed ρ version of SALSA analyzed by Tapia in [25]. It then follows from (4.26) and Theorem 7.2 of [25] that convergence of x_k is Q-superlinear. □

A corollary in the unconstrained case. It is interesting to note that if we apply Theorem 4.1 in the case of unconstrained optimization, it implies a new result about the convergence of the unconstrained BFGS method.

COROLLARY 4.2. *If x_* is a local minimizer of the function $f(x)$ such that $\nabla^2 f(x_*)$ is nonsingular, and the sequence $\{x_k\}$ generated by the* BFGS *method with step-length* 1 *converges to x_*, then the convergence is Q-superlinear.*

Proof. Note that Algorithm 4.1 applied to a problem with no constraints is simply the BFGS method. By Theorem 4.1, if the iterates converge to the solution, they do so R-superlinearly. This of course implies that

$$\sum_{k=0}^{\infty} \|x_k-x_*\| < \infty. \tag{4.27}$$

By Theorems 3.4 and 3.5 of Dennis and Moré [5], this implies that the sequence converges Q-superlinearly. □

Recently, Griewank [11] has shown us an alternative proof of Corollary 4.2 using techniques developed in [10].

5. Numerical experiments. The algorithms described in § 3 have been programmed and tested on a SUN 3/50 Workstation in double precision FORTRAN with a machine epsilon of about 2×10^{-16}. The tolerance for the stopping criterion was chosen as tol $= 10^{-6}$ and the allowed maximum iteration number was mxiter $= 100$. All the other parameters used in the algorithm are as specified in § 3. In particular, $\nu = \beta_1 = 0.01$. We now give some details about our numerical experiments that are not stated in the description of the algorithms.

5.1. Experiment description. In our implementation, we always set the initial Hessian approximation to a scalar multiple of the identity matrix. A pre-update scaling proposed by Oren and Spedicato [17] for use with the BFGS secant method for unconstrained optimization has been adapted to both algorithms SALSA and PDA to give this scalar. Following Shanno and Phua [23] we perform the pre-update scaling only at the first iteration.

In the unconstrained case the scaling factor is chosen so that the spectrum of the initial approximation B_0 overlaps the spectrum of the true Hessian of the objective near x_0. Now in SALSA, we are approximating the Hessian of the augmented Lagrangian and in PDA we are approximating the Hessian of the standard Lagrangian. These facts indicate that we should set $B_0^L = \eta I$ where $\eta > 0$ is a pre-update scaling factor that for SALSA would naturally be given by

$$\eta = y_0^{ST} s_0 / s_0^T s_0. \tag{5.1}$$

For PDA it is appropriate to use

$$\eta = \begin{cases} y_0^{\ell T} s_0 / s_0^T s_0, & y_0^{\ell T} s_0 > 0 \\ 1, & \text{otherwise.} \end{cases} \tag{5.2}$$

Observe that according to our construction of SALSA, the factor η given by (5.1) will always be positive and therefore SALSA will always take advantage of the pre-updating scaling. However, from (5.2) we see that this is not the case for the factor based on PDA. In order to ensure that any differences between the numerical performance of SALSA and PDA were not due to this difference in pre-update scaling, we used choice (5.2) for both SALSA and PDA in our experiments. This decision puts SALSA at a slight disadvantage, as numerical experimentation showed that choice (5.1) led to slightly better performance for SALSA than did choice (5.2).

As can be seen from the algorithm description, the algorithms SALSA and PDA are forced to terminate in the following three situations:

1. B_k^L is numerically indefinite. This is the situation when the Cholesky factorization of $Z_k^T B_k Z_k$ cannot be carried out or $s_k^T B_k s_k \leqq 0$;
2. the number of back-trackings in the line search exceeds 10;
3. the number of iterations exceeds mxiter.

All three of these cases will be called *irregular terminations*, in contrast to the regular termination that occurs when the stopping criterion is satisfied. In addition, the algorithms are stopped if a matrix A_k is found to be numerically rank deficient. However, this situation only occurred once in the entire sequence of experiments.

A set of 44 test problems has been chosen from Hock and Schittkowski [14] and Schittkowski [22]. A precise description of these problems can be found in the above two references. All the problems are numbered as in these references. Problems with numbers less than 200 (29 problems) are from Hock and Schittkowski [14] and the rest (15 problems) are from Schittkowski [22]. For those of the problems having inequality constraints, only the constraints active at the solution are included. Linearly constrained problems have been excluded from our test set.

Most of the test problems are so well conditioned that the identity matrix is often too good an approximation matrix to really test the robustness of an algorithm. In order to test the robustness of algorithms SALSA and PDA, from each given standard test problem we construct four scaled variants. We first define a diagonal matrix D_q by

$$D_{ii}=1+\left(1-\frac{i-1}{n-1}\right)(10^{-q}-1), \qquad i=1,2,\cdots,n, \tag{5.3}$$

where $q\in\mathbf{R}^n$ is a control parameter. In our tests, for each given objective function f and constraint function h, we solve the following five problems

$$\begin{aligned}&\text{minimize} && f(D_qx),\\ &\text{subject to} && h(D_qx)=0,\end{aligned} \tag{5.4}$$

for $q=0,1,2,3,4$. Obviously, $q=0$ corresponds to the original problem and $q>0$ to the scaled variants. If the Hessian matrix of a function $f(x)$ is $H(x)$, then after the diagonal scaling, the Hessian of $f(D_qx)$ is $D_qH(x)D_q$. Since the condition number of D_q^2 is 10^{2q} and if $H(x_*)$ is well conditioned, then for q large, in general, $D_qH(x_*)D_q$ will be relatively ill conditioned compared with $H(x_*)$.

The starting points x_0 are chosen as

$$x_0=x_s+(\gamma-1)(x_s-x_*), \tag{5.5}$$

where x_s are the standard starting points given in [14] and [22]. However, for Problems 12, 316–322, 336, and 338, we use $x_0=(10^{-4},\cdots,10^{-4})$ instead of the given $x_0=0$ because $A(0)$ has zero columns and therefore is not of full rank. It is easy to see that

$$\|x_0-x_*\|=|\gamma|\,\|x_s-x_*\|.$$

The number γ is thus used to control the distance $\|x_0-x_*\|$ and was given different values as described in § 5.2. For each problem, we let the integer q vary from 0 to 4. The total number of test cases is 220.

In the sequel, by one function evaluation we mean an evaluation of the $(m+1)$-vector $[f(x), h(x)]$. Similarly, one gradient evaluation represents an evaluation of the $n\times(m+1)$-matrix $[g(x)A(x)]$. Since the algorithms require only one gradient evaluation per iteration, the number of iterations needed for a run is always one less than the number of gradient evaluations, because iterations are counted from 0.

5.2. Numerical results. It is interesting to see how the two updating methods, SALSA and PDA, behave locally without a line search. After deactivating the line search subroutine as well as the pre-update scaling (because without a line search the information obtained from the first iteration is usually unreliable), we ran both SALSA and PDA on the 220 test cases, always using step-length one and starting from the standard starting points x_s given in [14] and [22] (i.e., we set $\gamma=1$ in (5.5)). It turns out that the standard starting points are fairly close to the solutions because for all the problems at least one of the two algorithms converged for at least one value of q. We will call this test (220 test cases) the local test.

We also tested SALSA and PDA with the line search procedure described in § 3.5 and with the pre-update scaling (5.2) on the same set of test problems. As already mentioned, the standard starting points as given in [14] and [22] are generally fairly close to the solutions. In order to test the algorithms in a realistic global environment, we set $\gamma=10$ for the starting points defined in (5.5) but with a few exceptions. Because for all the q-values both algorithms failed to converge for Problem 72, we still set $\gamma=1$ for this problem. We ran the two algorithms with the above prescribed starting points

TABLE 1
Average numbers of function and gradient evaluations.

Local test				Global test			
$q=0$		$q=1,2,3,4$		$q=0$		$q=1,2,3,4$	
SALSA	PDA	SALSA	PDA	SALSA	PDA	SALSA	PDA
23	21	32	31	24:29	27:29	28:33	30:37

and with the line search subroutine on the 220 test cases for the pre-update scaling (5.2). We will call this test (220 test cases) the global test.

Detailed information on both the local and the global tests that used the pre-update scaling (5.2) can be found in Tables 3-6 in the Appendix. In Table 1, we list the average numbers of function and gradient evaluations required by SALSA and PDA. To distinguish the standard test problems with its scaled variants, we present the results for $q=0$ (standard) and for $q>0$ (scaled) separately. The average number for each category is taken over all test cases in that category for which both SALSA and PDA converged. For the local test, since the number of function evaluations is always equal to the number of gradient evaluations, only one number is given for each category. For the global test, in each category the average number of function evaluations is given, followed by the average number of gradient evaluations separated with a colon. The rest of the table should be self-explanatory.

As one can see from Table 1, the numbers of function and gradient evaluations required by SALSA and PDA are comparable for test cases where both algorithms converged. Therefore, we infer, based on our numerical experiments, that as far as efficiency is concerned, SALSA and PDA appear comparable.

However, we observe that SALSA has displayed a somewhat higher degree of robustness. This can be seen from Table 2, where the irregular termination behavior of SALSA and PDA is summarized. As can be seen from the table, for the total number of 440 test cases, PDA had more irregular terminations than SALSA did (59 versus 42). However, since most of PDA's irregular terminations occurred in the local test, it does seem that the line search and the scalings helped to narrow the gap in robustness between SALSA and PDA.

We close this section by providing some additional observations obtained from our numerical tests. Among all the updates made by SALSA in our tests, the back-up strategy was used about 24 percent of the time. Of course, the choice of β_1 in (3.9) affects how often the back-up strategy is used and a decrease in the value of β_1 will

TABLE 2
Number of irregular terminations.

Local test				Global test			
$q=0$		$q=1,2,3,4$		$q=0$		$q=1,2,3,4$	
SALSA	PDA	SALSA	PDA	SALSA	PDA	SALSA	PDA
4	9	20	26	2	6	16	18

result in less usage of the back-up strategy. As a comparison, we also ran SALSA using Powell's damped BFGS update as a back-up strategy instead of the one described in § 3.3. Very similar results were obtained, though Powell's damped BFGS update, as the back-up strategy, was used slightly more often, and the number of function evaluations was slightly increased.

6. Concluding remarks. SALSA appears to have certain theoretical advantages over PDA. On the one hand, if a value for the augmentation parameter happens to be picked up that is greater than the threshold value, under standard assumptions, it will have local and Q-superlinear convergence. Local convergence has not yet been established for Powell's damped BFGS method. On the other hand, if the augmentation parameter happens to be smaller than the threshold value, we have established, under much weaker and more realistic assumptions than those that were assumed by Powell, that SALSA will, if it converges, converge at an R-superlinear rate, as has been proved for Powell's damped BFGS method. As an immediate corollary, we have that if the BFGS secant method in unconstrained optimization converges, it gives Q-superlinear convergence.

Our numerical experiments have shown that for a fairly large set of test problems the overall numerical performance of SALSA was moderately better than that of PDA in terms of robustness as measured by the number of irregular terminations. The higher degree of robustness of SALSA is likely due to the fact that B_k is not involved in y_k^S, but is involved in y_k^P (see (3.1) and (1.10)).

Based on the established convergence results and our computational experiments, we have been led to the conclusion that in addition to its strong theoretical properties, the structured augmented Lagrangian BFGS secant method, if properly implemented, also performs experimentally at least as well as Powell's damped BFGS secant method.

Appendix: Tables. Detailed information on our numerical experiments is given here.

In Tables 3-6, the problems are numbered after Hock and Schittkowski [14] and Schittkowski [22] and are specified in the first column, along with the corresponding numbers of variables and constraints ($n:m$).

For each value of q listed in the tables are $ng:nf$, the numbers of gradient and function evaluations, respectively, as well as the final values of $\|\nabla \ell(x_k, \lambda_k)\|^2$ when the algorithms terminate.

The irregular terminations are indicated by boxes around the values of $\|\nabla \ell(x_k, \lambda_k)\|_2$ that are greater than tol$=10^{-6}$. The symbols "Inf" and "NaN" in the tables stand for "Infinity" and for "Not a Number" under the IEEE floating point standard as implemented in the operating system SunOS 4.0.3. Basically, both indicate that a floating point overflow has occurred.

The three types of irregular terminations as listed in § 5 can be distinguished as follows. If the number of gradient evaluations is 101, then the algorithm was stopped because the maximum number of iterations was exceeded. If a pair $ng:nf$ is followed by an asterisk "*," then the algorithm was stopped because the maximum number of back-tracking steps in the line search was exceeded. Otherwise, the irregular terminations were due to the numerical indefiniteness of the Hessian approximation matrix. For Problem 72 in the local test of Powell's Damped BFGS Algorithm (PDA), the blank entry indicates that the algorithm was terminated because the matrix A_k was found to be rank deficient.

TABLE 3

SALSA *without line search.*

Prob.#	$q = 0$		$q = 1$		$q = 2$		$q = 3$		$q = 4$	
$(n : m)$	$ng : nf$	$\|\|\nabla\ell\|\|$	$ng : nf$	$\|\|\nabla\ell\|\|$	$ng : nf$	$\|\|\nabla\ell\|\|$	$ng : nf$	$\|\|\nabla\ell\|\|$	$ng : nf$	$\|\|\nabla\ell\|\|$
6 (2:1)	9:9	.2D-09	17:17	.4D-08	21:21	.7D-10	27:27	.5D-06	31:31	.2D-07
7 (2:1)	15:15	.2D-07	14:14	.6D-09	17:17	.3D-06	18:18	.2D-06	21:21	.1D-07
10 (2:1)	12:12	.1D-06	12:12	.1D-06	15:15	.3D-07	15:15	.3D-07	16:16	.8D-06
11 (2:1)	8:8	.2D-07	12:12	.1D-07	17:17	.4D-06	23:23	.3D-10	25:25	.8D-09
12 (2:1)	51:51	.6D-09	46:46	.5D-06	35:35	.1D-06	33:33	.2D-06	25:25	.8D-06
26 (3:1)	34:34	.5D-06	31:31	.5D-06	36:36	.1D-05	43:43	.8D-06	47:47	.6D-06
27 (3:1)	34:34	.6D-06	22:22	.5D-07	8:8	.9D-06	4:4	.2D-06	4:4	.2D-08
29 (3:1)	13:13	.3D-08	14:14	.4D-07	31:31	.7D-06	34:34	.2D-07	41:41	.4D-07
39 (4:2)	13:13	.8D-06	14:14	.2D-06	19:19	.4D-06	24:24	.9D-06	10:10	.5D-07
40 (4:3)	8:8	.5D-07	33:33	.9D-06	6:6	.2D-07	6:6	.4D-11	6:6	.3D-11
43 (4:2)	12:12	.4D-07	17:17	.3D-07	39:39	.5D-07	57:57	.2D-06	88:88	.2D-06
46 (5:2)	101:101	$\boxed{\text{.2D+00}}$	16:16	$\boxed{\text{.4D+60}}$	29:29	.1D-06	38:38	.2D-06	38:38	.3D-06
47 (5:3)	68:68	.4D-07	23:23	.3D-07	36:36	.8D-06	52:52	.1D-06	79:79	.7D-06
56 (7:4)	14:14	.5D-06	101:101	$\boxed{\text{.3D+02}}$	10:10	.5D-07	11:11	.9D-06	12:12	.5D-06
60 (3:1)	11:11	.8D-06	19:19	.3D-06	24:24	.7D-07	39:39	.2D-08	37:37	.6D-06
61 (3:2)	11:11	.3D-11	11:11	.5D-09	14:14	.6D-06	15:15	.2D-07	16:16	.2D-06
63 (3:2)	8:8	.6D-06	13:13	.4D-08	14:14	.2D-06	15:15	.9D-07	15:15	.8D-06
65 (3:1)	10:10	.7D-07	27:27	.2D-07	28:28	.3D-07	37:37	.8D-08	40:40	.4D-07
66 (3:2)	7:7	.3D-06	27:27	.9D-07	101:101	$\boxed{\text{NaN}}$	101:101	$\boxed{\text{NaN}}$	101:101	$\boxed{\text{NaN}}$
71 (4:3)	6:6	.4D-07	14:14	.5D-06	19:19	.3D-07	40:40	.2D-07	63:63	.2D-06
72 (4:2)	21:21	.2D-06	31:31	.3D-06	101:101	$\boxed{\text{NaN}}$	50:50	$\boxed{\text{.7D+00}}$	43:43	$\boxed{\text{.7D+00}}$
77 (5:2)	43:43	.4D-06	101:101	$\boxed{\text{.1D-01}}$	31:31	.1D-06	37:37	.4D-06	43:43	.2D-06
78 (5:3)	8:8	.4D-07	51:51	.2D-06	32:32	.3D-06	29:29	.1D-06	8:8	.6D-07
79 (5:3)	11:11	.1D-06	15:15	.8D-08	49:49	.3D-07	83:83	.2D-07	75:75	.5D-08
80 (5:3)	7:7	.8D-08	33:33	.4D-07	32:32	.2D-06	39:39	.1D-08	6:6	.8D-07
81 (5:3)	10:10	.6D-08	22:22	.2D-06	101:101	$\boxed{\text{NaN}}$	101:101	$\boxed{\text{.1D+00}}$	6:6	.8D-07
93 (6:2)	33:33	.3D-06	6:6	$\boxed{\text{.1D+28}}$	4:4	$\boxed{\text{.2D+44}}$	24:24	$\boxed{\text{.2D+01}}$	29:29	$\boxed{\text{.4D+07}}$
100 (7:2)	51:51	.3D-06	28:28	.5D-06	37:37	.2D-06	47:47	.1D-06	52:52	.6D-06
104 (8:4)	27:27	.4D-06	101:101	$\boxed{\text{.5D+09}}$	101:101	$\boxed{\text{.5D+02}}$	101:101	$\boxed{\text{.5D+02}}$	101:101	$\boxed{\text{.5D+05}}$
106 (8:6)	43:43	$\boxed{\text{.1D+15}}$	29:29	$\boxed{\text{.2D+09}}$	89:89	$\boxed{\text{.8D+08}}$	61:61	.3D-06	61:61	.6D-06
216 (2:1)	25:25	.1D-06	9:9	.2D-07	22:22	.1D-09	29:29	.5D-07	38:38	.7D-07
219 (4:2)	18:18	.2D-07	22:22	.2D-07	25:25	.2D-06	30:30	.6D-07	17:17	.2D-06
316 (2:1)	60:60	.2D-08	60:60	.4D-07	41:41	.4D-08	37:37	.4D-08	27:27	.3D-06
317 (2:1)	59:59	.2D-09	54:54	.3D-07	48:48	.4D-07	36:36	.1D-07	28:28	.5D-11
318 (2:1)	62:62	.4D-06	60:60	.3D-08	43:43	.8D-07	59:59	.6D-09	28:28	.3D-11
319 (2:1)	59:59	.9D-11	59:59	.6D-08	53:53	.8D-09	44:44	.2D-07	33:33	.1D-06
320 (2:1)	49:49	.2D-06	70:70	.3D-07	53:53	.3D-10	43:43	.2D-06	37:37	.1D-07
321 (2:1)	44:44	.8D-06	71:71	.4D-08	53:53	.3D-10	46:46	.7D-06	44:44	.2D-08
322 (2:1)	25:25	.3D-11	39:39	.4D-06	64:64	.4D-08	50:50	.3D-06	51:51	.6D-08
335 (3:2)	25:25	.7D-08	28:28	.3D-07	28:28	.3D-06	37:37	.4D-07	46:46	.2D-06
336 (3:2)	48:48	.2D-08	75:75	.7D-07	76:76	.2D-06	72:72	.1D-05	72:72	.8D-09
338 (3:2)	56:56	.4D-06	43:43	.2D-09	36:36	.2D-07	36:36	.5D-08	30:30	.1D-07
355 (4:1)	7:7	$\boxed{\text{.3D+47}}$	41:41	.2D-07	101:101	$\boxed{\text{.3D+00}}$	24:24	$\boxed{\text{.3D+38}}$	59:59	$\boxed{\text{.3D+68}}$
373 (9:6)	101:101	$\boxed{\text{.2D+04}}$	14:14	.5D-06	21:21	.1D-07	26:26	.8D-07	32:32	.2D-06
375 (10:9)	26:26	.4D-10	16:16	.2D-07	20:20	.4D-06	15:15	.1D-07	16:16	.1D-09

TABLE 4

PDA *without line search.*

Prob.#	$q = 0$		$q = 1$		$q = 2$		$q = 3$		$q = 4$	
$(n : m)$	$ng : nf$	$\|\|\nabla\ell\|\|$	$ng : nf$	$\|\|\nabla\ell\|\|$	$ng : nf$	$\|\|\nabla\ell\|\|$	$ng : nf$	$\|\|\nabla\ell\|\|$	$ng : nf$	$\|\|\nabla\ell\|\|$
6 (2:1)	13:13	.5D-09	14:14	.3D-08	18:18	.1D-11	21:21	.8D-08	26:26	.2D-13
7 (2:1)	15:15	.4D-08	21:21	.2D-10	47:47	.1D-06	52:52	.2D-07	46:46	.2D-08
10 (2:1)	12:12	.1D-06	13:13	.2D-06	17:17	.8D-10	20:20	.1D-08	22:22	.3D-08
11 (2:1)	8:8	.2D-07	13:13	.5D-07	18:18	.3D-06	24:24	.8D-08	32:32	.1D-06
12 (2:1)	21:21	.2D-11	24:24	.2D-10	39:39	.2D-09	28:28	.9D-06	30:30	.2D-06
26 (3:1)	34:34	.5D-06	31:31	.5D-06	35:35	.5D-06	43:43	.8D-06	46:46	.9D-06
27 (3:1)	31:31	.3D-07	22:22	.5D-07	12:12	.5D-06	4:4	.2D-06	4:4	.2D-08
29 (3:1)	49:49	$\boxed{\text{.2D+109}}$	13:13	.9D-07	25:25	.2D-07	29:29	.4D-07	35:35	.2D-07
39 (4:2)	13:13	.8D-06	14:14	.2D-06	16:16	.8D-06	18:18	.2D-08	10:10	.4D-07
40 (4:3)	8:8	.2D-07	39:39	.4D-08	6:6	.2D-07	6:6	.4D-11	6:6	.3D-11
43 (4:2)	12:12	.4D-07	16:16	.5D-08	44:44	.4D-08	63:63	.1D-06	87:87	.6D-06
46 (5:2)	62:62	.6D-06	89:89	.5D-06	23:23	$\boxed{\text{Inf}}$	39:39	.9D-06	43:43	.8D-07
47 (5:3)	54:54	.1D-06	29:29	.9D-07	65:65	.4D-07	64:64	.2D-06	87:87	.3D-06
56 (7:4)	101:101	$\boxed{\text{NaN}}$	7:7	.7D-06	10:10	.5D-07	12:12	.8D-06	13:13	.8D-06
60 (3:1)	11:11	.8D-06	20:20	.7D-07	28:28	.2D-08	27:27	.1D-06	37:37	.3D-06
61 (3:2)	11:11	.3D-11	15:15	.8D-06	19:19	.4D-08	22:22	.3D-07	24:24	.6D-07
63 (3:2)	8:8	.6D-06	15:15	.5D-09	21:21	.2D-09	19:19	.4D-07	22:22	.5D-08
65 (3:1)	14:14	.1D-06	22:22	.8D-06	32:32	.4D-06	98:98	.3D-07	95:95	.3D-06
66 (3:2)	7:7	.3D-06	28:28	.5D-08	101:101	$\boxed{\text{.2D+07}}$	75:75	.7D-08	101:101	$\boxed{\text{NaN}}$
71 (4:3)	6:6	.4D-07	15:15	.6D-09	23:23	.2D-06	51:51	.6D-06	73:73	.5D-09
72 (4:2)	21:21	.4D-06	37:37	.1D-06	101:101	$\boxed{\text{NaN}}$	—	□	43:43	$\boxed{\text{.7D+00}}$
77 (5:2)	44:44	.3D-06	17:17	.8D-06	27:27	.3D-06	31:31	.2D-06	39:39	.7D-07
78 (5:3)	8:8	.4D-07	22:22	.8D-06	14:14	$\boxed{\text{.1D+49}}$	54:54	.5D-06	8:8	.1D-06
79 (5:3)	11:11	.1D-06	20:20	.2D-06	42:42	.9D-07	84:84	.1D-06	101:101	$\boxed{\text{.3D+02}}$
80 (5:3)	7:7	.8D-08	20:20	.1D-06	35:35	.5D-07	101:101	$\boxed{\text{NaN}}$	6:6	.8D-07
81 (5:3)	9:9	.8D-08	27:27	.3D-07	54:54	.7D-07	92:92	.4D-06	6:6	.8D-07
93 (6:2)	35:35	.3D-06	101:101	$\boxed{\text{NaN}}$	5:5	$\boxed{\text{.1D+47}}$	34:34	$\boxed{\text{.4D-01}}$	44:44	$\boxed{\text{.1D+08}}$
100 (7:2)	50:50	.5D-06	28:28	.5D-06	37:37	.2D-06	47:47	.1D-06	60:60	.3D-06
104 (8:4)	27:27	.4D-06	92:92	$\boxed{\text{.6D+14}}$	84:84	$\boxed{\text{.2D-01}}$	101:101	$\boxed{\text{.3D+00}}$	101:101	$\boxed{\text{.9D+01}}$
106 (8:6)	29:29	.5D-08	32:32	.1D-08	57:57	$\boxed{\text{.6D+12}}$	51:51	$\boxed{\text{.2D+14}}$	101:101	$\boxed{\text{.2D+21}}$
216 (2:1)	26:26	.1D-09	9:9	.2D-07	26:26	.4D-11	37:37	.3D-09	50:50	.5D-09
219 (4:2)	18:18	.2D-07	20:20	.9D-08	23:23	.3D-06	35:35	.1D-06	18:18	.2D-06
316 (2:1)	37:37	.3D-06	40:40	.2D-07	29:29	.3D-10	31:31	.1D-07	32:32	.2D-06
317 (2:1)	17:17	$\boxed{\text{.6D+02}}$	37:37	.5D-06	28:28	.5D-06	31:31	.2D-09	32:32	.2D-07
318 (2:1)	17:17	$\boxed{\text{.1D+03}}$	33:33	.9D-08	28:28	.6D-07	30:30	.2D-07	32:32	.2D-07
319 (2:1)	18:18	$\boxed{\text{.8D+02}}$	22:22	$\boxed{\text{.2D+03}}$	29:29	.7D-07	30:30	.1D-07	31:31	.3D-07
320 (2:1)	15:15	$\boxed{\text{.1D+03}}$	22:22	$\boxed{\text{.4D+02}}$	34:34	.1D-05	29:29	.1D-08	30:30	.2D-07
321 (2:1)	15:15	$\boxed{\text{.2D+02}}$	37:37	.7D-07	41:41	.2D-07	28:28	.1D-07	29:29	.4D-06
322 (2:1)	55:55	.1D-05	45:45	.8D-12	39:39	.1D-07	35:35	.4D-08	27:27	.5D-08
335 (3:2)	25:25	.1D-07	32:32	.1D-07	28:28	.3D-07	39:39	.3D-07	51:51	.3D-06
336 (3:2)	32:32	.8D-10	54:54	.2D-06	50:50	.2D-07	65:65	.1D-07	82:82	.5D-08
338 (3:2)	19:19	$\boxed{\text{.1D+03}}$	13:13	$\boxed{\text{.2D+06}}$	15:15	$\boxed{\text{.1D+02}}$	16:16	$\boxed{\text{.2D+01}}$	18:18	$\boxed{\text{.1D-02}}$
355 (4:1)	6:6	$\boxed{\text{.7D+106}}$	6:6	$\boxed{\text{.2D+84}}$	23:23	$\boxed{\text{.1D+71}}$	35:35	$\boxed{\text{.6D+61}}$	22:22	$\boxed{\text{.1D+60}}$
373 (9:6)	25:25	.3D-06	14:14	.5D-06	21:21	.1D-07	26:26	.2D-06	30:30	.2D-06
375 (10:9)	14:14	.2D-07	15:15	.1D-07	16:16	.4D-06	22:22	.3D-07	20:20	.5D-06

TABLE 5

SALSA *with line search.*

Prob.#	$q = 0$		$q = 1$		$q = 2$		$q = 3$		$q = 4$	
$(n : m)$	$ng : nf$	$\|\|\nabla\ell\|\|$	$ng : nf$	$\|\|\nabla\ell\|\|$	$ng : nf$	$\|\|\nabla\ell\|\|$	$ng : nf$	$\|\|\nabla\ell\|\|$	$ng : nf$	$\|\|\nabla\ell\|\|$
6 (2:1)	13:19	.2D-07	15:16	.3D-07	26:30	.6D-09	29:31	.8D-08	38:43	.6D-06
7 (2:1)	18:20	.8D-08	20:21	.5D-07	26:28	.2D-09	24:24	.8D-06	28:32	.4D-07
10 (2:1)	17:19	.2D-10	18:20	.1D-05	17:17	.1D-06	19:19	.3D-07	19:19	.3D-06
11 (2:1)	11:12	.3D-07	15:18	.2D-06	21:26	.3D-09	21:26	.1D-07	30:34	.1D-08
12 (2:1)	23:31	.9D-08	23:28	.2D-09	20:27	.8D-11	38:63	.2D-08	30:43	.3D-07
26 (3:1)	32:34	.9D-06	31:31	.8D-06	34:35	.7D-06	47:48	.8D-06	53:54	.6D-06
27 (3:1)	51:63	.2D-06	31:39	.6D-07	61:72	.1D-07	17:17	.8D-06	4:4	.2D-07
29 (3:1)	27:32	.3D-07	15:15	.3D-07	19:22	.5D-09	23:25	.7D-06	29:32	.3D-06
39 (4:2)	18:20	.3D-06	25:27	.5D-07	22:22	.4D-06	28:28	.5D-07	14:14	.6D-06
40 (4:3)	10:11	.1D-06	22:23	.9D-06	7:7	.8D-08	7:7	.7D-12	7:7	.5D-12
43 (4:2)	26:37	.2D-06	29:42	.8D-06	37:52	.8D-06	47:69	.4D-08	55:81	.3D-06
46 (5:2)	101:116	$\boxed{.7D-02}$	25:29	.7D-06	41:47	.3D-06	60:70	.2D-06	89:99	.9D-06
47 (5:3)	57:61	.1D-06	22:23	.1D-07	40:44	.4D-06	36:38	.2D-06	69:76	.8D-06
56 (7:4)	20:27	.9D-07	40:45	.3D-07	41:80	$\boxed{.7D-05}$	12:12	.2D-07	15:15	.5D-06
60 (3:1)	23:25	.4D-06	25:26	.8D-07	31:33	.1D-06	35:36	.1D-05	41:42	.4D-06
61 (3:2)	25:36	.2D-10	17:20	.8D-07	14:16	.1D-06	16:17	.1D-06	17:18	.2D-09
63 (3:2)	9:9	.6D-09	10:10	.2D-08	11:12	.4D-07	12:13	.2D-07	13:14	.1D-07
65 (3:1)	24:34	.3D-06	44:65	.7D-06	29:35	.7D-06	31:36	.7D-08	34:39	.2D-07
66 (3:2)	8:8	.2D-09	18:27	.3D-07	6:7	$\boxed{.8D+00}$	6:21*	$\boxed{.8D+00}$	8:37*	$\boxed{.8D+00}$
71 (4:3)	9:10	.6D-07	10:10	.4D-08	11:12	.3D-06	12:15	.7D-09	16:23	.3D-06
72 (4:2)	26:27	.3D-06	31:31	.3D-06	32:35	$\boxed{.8D+00}$	89:114	$\boxed{.7D+00}$	39:42	$\boxed{.7D+00}$
77 (5:2)	78:90	.8D-06	24:26	.2D-06	29:30	.4D-06	34:35	.9D-06	39:40	.3D-08
78 (5:3)	28:39	.1D-06	11:12	.6D-06	41:43	.6D-06	54:77	.6D-06	8:8	.3D-06
79 (5:3)	17:18	.9D-07	20:21	.1D-07	22:23	.6D-07	40:49	.5D-07	59:72	.8D-07
80 (5:3)	9:9	.7D-06	18:18	.4D-06	32:36	.6D-07	65:100	.6D-06	6:6	.6D-06
81 (5:3)	19:22	.4D-06	24:24	.9D-07	42:48	.3D-06	93:103	.4D-06	6:6	.6D-06
93 (6:2)	30:35	.2D-06	84:108	.5D-07	72:108	.4D-07	101:205	$\boxed{.3D+01}$	37:172	$\boxed{.2D+04}$
100 (7:2)	82:93	.1D-06	53:60	.2D-06	55:61	.1D-06	68:76	.4D-06	75:86	.5D-07
104 (8:4)	25:26	.8D-06	28:31	.1D-06	33:38	.8D-06	14:30*	$\boxed{.2D-01}$	26:48*	$\boxed{.1D+01}$
106 (8:6)	40:42	.1D-08	40:63*	$\boxed{.2D+14}$	37:52	$\boxed{.4D+16}$	60:65	.9D-06	80:98	.3D-08
216 (2:1)	17:20	.4D-10	16:18	.1D-06	14:14	.4D-07	13:16	.2D-07	13:21	.1D-05
219 (4:2)	39:49	.9D-08	47:59	.4D-07	36:37	.3D-06	43:45	.5D-06	50:52	.5D-07
316 (2:1)	25:28	.9D-07	23:29	.2D-06	16:18	.3D-08	17:19	.2D-08	17:19	.7D-09
317 (2:1)	14:21	.2D-09	42:59	.3D-06	16:17	.2D-06	16:17	.5D-07	16:17	.2D-07
318 (2:1)	19:22	.4D-07	21:25	.2D-07	17:18	.1D-07	16:17	.1D-06	16:17	.8D-08
319 (2:1)	22:33	.2D-07	31:45	.1D-05	19:20	.3D-08	18:18	.2D-07	17:17	.4D-06
320 (2:1)	13:15	.1D-05	22:33	.1D-06	25:29	.2D-07	19:19	.9D-06	20:21	.3D-07
321 (2:1)	15:16	.5D-06	16:19	.1D-09	26:35	.5D-06	25:26	.3D-06	22:24	.6D-06
322 (2:1)	19:22	.2D-06	19:23	.1D-12	14:16	.3D-06	17:18	.6D-07	21:21	.3D-08
335 (3:2)	23:30	.2D-07	79:176	$\boxed{.3D-01}$	42:56	.3D-07	28:36	.6D-09	22:27	.6D-06
336 (3:2)	20:27	.8D-07	21:25	.2D-07	29:45	.1D-05	31:52	.1D-07	19:57*	$\boxed{.3D-02}$
338 (3:2)	9:9	.1D-07	15:17	.2D-06	20:28	.2D-07	22:33	.4D-06	30:46	.4D-06
355 (4:1)	84:112	.9D-06	32:39	.8D-07	93:166	.1D-06	101:150	$\boxed{.6D+00}$	101:188	$\boxed{.1D-01}$
373 (9:6)	101:206	$\boxed{.3D+01}$	101:203	$\boxed{.4D+01}$	34:41	.6D-06	42:51	.2D-07	48:58	.4D-06
375 (10:9)	11:11	.5D-06	15:16	.5D-07	16:17	.3D-06	16:16	.6D-06	18:18	.8D-10

TABLE 6

PDA *with line search.*

Prob.#	$q = 0$		$q = 1$		$q = 2$		$q = 3$		$q = 4$	
$(n : m)$	$ng : nf$	$\|\nabla \ell\|$	$ng : nf$	$\|\nabla \ell\|$	$ng : nf$	$\|\nabla \ell\|$	$ng : nf$	$\|\nabla \ell\|$	$ng : nf$	$\|\nabla \ell\|$
6 (2:1)	13:16	.3D-10	14:17	$\boxed{\text{.5D-05}}$	25:25	.1D-13	29:34	.4D-12	32:33	.3D-12
7 (2:1)	17:21	.5D-06	35:42	.3D-09	34:38	.2D-06	36:41	.3D-07	35:39	.6D-06
10 (2:1)	17:19	.2D-10	18:20	.1D-06	18:18	.6D-06	23:23	.2D-09	25:25	.9D-09
11 (2:1)	11:12	.3D-07	16:18	.7D-08	23:28	.5D-07	28:32	.3D-10	34:39	.1D-07
12 (2:1)	23:28	.2D-10	23:29	.5D-07	23:29	.5D-07	27:35	.9D-09	32:41	.3D-10
26 (3:1)	32:34	.9D-06	31:31	.8D-06	36:37	.6D-06	47:48	.8D-06	53:54	.6D-06
27 (3:1)	46:55	.2D-07	50:69	.2D-06	101:132	$\boxed{\text{.2D-01}}$	18:18	.3D-06	4:4	.2D-07
29 (3:1)	31:51	$\boxed{\text{.8D+08}}$	15:15	.3D-07	21:23	.9D-06	22:25	.2D-06	28:31	.4D-06
39 (4:2)	18:20	.3D-06	20:20	.4D-06	31:32	.2D-07	28:29	.4D-07	19:20	.4D-06
40 (4:3)	10:11	.1D-06	22:22	.9D-06	7:7	.8D-08	7:7	.7D-12	7:7	.5D-12
43 (4:2)	26:39	.3D-07	29:42	.5D-07	41:59	.1D-06	56:83	.3D-06	62:92	.3D-06
46 (5:2)	101:126	$\boxed{\text{.1D-03}}$	26:30	.4D-06	101:134	$\boxed{\text{.1D-05}}$	99:133	.8D-06	101:135	$\boxed{\text{.2D-05}}$
47 (5:3)	58:62	.6D-06	22:23	.1D-07	41:47	.3D-06	36:38	.2D-06	72:94	.9D-06
56 (7:4)	18:25	.9D-06	33:36	.5D-07	48:58*	$\boxed{\text{.3D+23}}$	12:12	.2D-07	15:15	.5D-06
60 (3:1)	24:27	.4D-07	23:24	.1D-07	32:33	.6D-07	36:38	.2D-06	41:42	.4D-06
61 (3:2)	18:21	.2D-06	15:16	.5D-06	17:18	.3D-10	20:20	.9D-08	21:22	.8D-08
63 (3:2)	9:9	.6D-09	10:10	.2D-08	12:12	.6D-06	15:15	.3D-08	18:18	.1D-07
65 (3:1)	25:37	.8D-06	27:42	.8D-09	29:35	.2D-07	36:44	.7D-06	35:49	.3D-06
66 (3:2)	8:8	.2D-09	18:24	.8D-06	26:35	.3D-06	36:43	.6D-06	36:50	.1D-07
71 (4:3)	9:10	.6D-07	10:10	.4D-08	12:13	.2D-08	16:19	.3D-09	29:38	.2D-06
72 (4:2)	26:27	.8D-06	32:33	.2D-06	35:62*	$\boxed{\text{.7D+00}}$	48:72*	$\boxed{\text{.3D+00}}$	43:43	$\boxed{\text{.7D+00}}$
77 (5:2)	80:89	.2D-06	23:24	.1D-06	29:30	.4D-06	34:35	.9D-06	38:39	.5D-07
78 (5:3)	24:32	.2D-08	11:12	.6D-06	29:34	.9D-07	67:102	.1D-07	8:8	.3D-06
79 (5:3)	17:18	.9D-07	20:21	.2D-07	21:22	.5D-07	48:54	.2D-07	48:57	.8D-07
80 (5:3)	9:9	.7D-06	18:18	.4D-06	27:28	.2D-06	62:95	.3D-07	6:6	.6D-06
81 (5:3)	18:20	.7D-08	23:25	.6D-07	33:34	.3D-06	90:94	.7D-06	6:6	.6D-06
93 (6:2)	29:34	.8D-07	87:111	.3D-07	33:41	.6D-06	59:74	.1D-06	62:138	$\boxed{\text{.2D+06}}$
100 (7:2)	83:96	.2D-06	53:60	.8D-06	54:59	.2D-06	61:66	.9D-06	66:70	.2D-06
104 (8:4)	25:26	.8D-06	28:31	.1D-06	36:40	.5D-07	14:30*	$\boxed{\text{.2D-01}}$	24:40*	$\boxed{\text{.8D+00}}$
106 (8:6)	42:45	$\boxed{\text{.2D+19}}$	32:42	$\boxed{\text{.2D+16}}$	45:55	$\boxed{\text{.4D+09}}$	63:68	.1D-05	64:92	$\boxed{\text{.6D+21}}$
216 (2:1)	17:20	.4D-10	16:18	.1D-06	14:14	.4D-07	17:19	.2D-09	27:35	.5D-06
219 (4:2)	49:57	.2D-07	80:115	.6D-08	52:60	.6D-06	46:85*	$\boxed{\text{.4D+00}}$	47:62	.2D-07
316 (2:1)	33:63	.3D-06	23:29	.2D-06	23:30	.8D-07	26:31	.2D-07	27:34	.2D-07
317 (2:1)	16:23	.2D-06	30:48	.3D-06	24:38	.2D-07	24:28	.4D-06	33:48	.3D-07
318 (2:1)	14:18	.5D-06	19:24	.4D-07	22:28	.9D-06	25:31	.8D-07	29:46	.2D-09
319 (2:1)	17:23	.7D-07	18:22	.3D-06	26:40	.2D-06	25:36	.1D-07	27:36	.2D-07
320 (2:1)	16:20	.6D-06	15:17	.6D-07	29:43	.1D-07	24:33	.2D-07	29:40	.9D-11
321 (2:1)	17:20	.2D-06	16:19	.1D-09	26:31	.8D-09	32:52	.1D-06	29:43	.2D-06
322 (2:1)	12:14	$\boxed{\text{.4D+01}}$	20:26	.4D-07	14:16	.3D-06	26:37	.5D-06	29:43	.2D-08
335 (3:2)	25:31	.3D-06	42:64	.8D-09	35:45	.5D-09	36:41	.1D-06	26:31	.4D-07
336 (3:2)	21:25	.4D-08	21:28	.3D-07	36:66	$\boxed{\text{.5D-02}}$	24:34	.9D-07	26:44	.1D-09
338 (3:2)	9:9	.1D-07	15:16	.1D-09	23:29	.1D-08	28:41	.6D-07	12:22	.3D-06
355 (4:1)	101:136	$\boxed{\text{.2D-01}}$	35:42	.4D-07	93:166	.1D-06	101:230	$\boxed{\text{.7D-01}}$	101:155	$\boxed{\text{.1D-01}}$
373 (9:6)	101:223	$\boxed{\text{.3D+01}}$	101:199	$\boxed{\text{.4D+01}}$	34:41	.6D-06	42:52	.4D-06	48:58	.4D-06
375 (10:9)	11:11	.5D-06	17:17	.7D-06	17:17	.5D-06	20:20	.3D-07	22:22	.4D-06

REFERENCES

[1] R. H. BYRD AND J. NOCEDAL, *An analysis of reduced Hessian methods for constrained optimization*, Math. Programming, 49 (1991), pp. 285-323.

[2] ———, *A tool for the analysis of quasi-Newton methods with application to unconstrained optimization*, SIAM J. Numer. Anal., 26 (1989), pp. 727-739.

[3] R. M. CHAMBERLAIN, *Some examples of cycling in variable metric method for constrained optimization*, Math. Programming, 16 (1979), pp. 378-383.

[4] T. F. COLEMAN AND A. R. CONN, *On the local convergence of a quasi-Newton method for the nonlinear programming problem*, SIAM J. Numer. Anal., 21 (1984), pp. 755-769.

[5] J. E. DENNIS, Jr. AND J. J. MORÉ, *A characterization of superlinear convergence and its application to quasi-Newton methods*, Math. Comp., 28 (1974), pp. 549-560.

[6] P. FENYES, *Partitioned quasi-Newton methods for nonlinear equality constrained optimization*, Ph.D. thesis, Department of Computer Science, Cornell University, Ithaca, NY, 1987.

[7] R. FONTECILLA, *Local convergence of secant methods for nonlinear constrained optimization*, SIAM J. Numer. Anal., 25 (1988), pp. 692-712.

[8] R. FONTECILLA, T. STEIHAUG, AND R. A. TAPIA, *A convergence theory for a class of quasi-Newton methods for constrained optimization*, SIAM J. Numer. Anal., 24 (1987), pp. 1133-1151.

[9] S. T. GLAD, *Properties of updating methods for the multipliers in augmented Lagrangians*, 28 (1979), pp. 135-156.

[10] A. GRIEWANK, *The global convergence of partitioned* BFGS *on semi-smooth problems with convex decompositions*, Report ANL/MCS-TM-105, Mathematics and Computer Science Division, Argonne National Laboratory, Argonne, IL, 1987.

[11] A. GRIEWANK, Private communication, 1989.

[12] S.-P. HAN, *Superlinear convergent variable metric algorithms for general nonlinear programming*, Math. Programming, 11 (1976), pp. 263-282.

[13] ———, *A globally convergent method for nonlinear programming*, J. Optim. Theory Appl., 22 (1977), pp. 297-309.

[14] W. HOCK AND K. SCHITTKOWSKI, *Test examples for nonlinear programming codes*, Lecture Notes in Econom. and Math. Systems 187, Springer-Verlag, Berlin, 1981.

[15] H. J. MARTÍNEZ, *A numerical investigation on the* BFGS *update*, Master's thesis, Department of Mathemetical Sciences, Rice University, Houston, TX, 1986.

[16] J. NOCEDAL AND M. OVERTON, *Projected Hessian updating algorithms for nonlinear constrained optimization*, SIAM J. Numer. Anal., 22 (1985), pp. 821-850.

[17] S. S. OREN AND E. SPEDICATO, *Optimal conditioning of self-scaling variable metric algorithms*, Math. Programming, 10 (1976), pp. 70-90.

[18] M. J. D. POWELL, *The convergence of variable metric method for nonlinearly constrained optimization calculation*, in Nonlinear Programming 3, O. Mangasarian, R. Meyer, and S. Robinson, eds., Academic Press, New York, 1978.

[19] ———, *A fast algorithm for nonlinearly constrained optimization calculation*, in Numerical Analysis Proceedings, Dundee 1977, G. A. Watson, ed., Springer-Verlag, Berlin, 1978.

[20] ———, *Extensions to subroutine* VFO2AD, in System Modeling and Optimization, Lecture Notes in Control and Inform. Sci. 38, R. F. Drenick and F. Kozin, eds., Springer-Verlag, New York, 1982.

[21] ———, *The performance of two subroutines for constrained optimization on some difficult test problems*, in Numerical Optimization 1984, P. T. Boggs, R. Byrd, and R. Schnabel, eds., Society for Industrial and Applied Mathematics, Philadelphia, 1985.

[22] K. Schittkowski, *More test examples for nonlinear programming codes*, Lecture Notes in Econom. and Math. Systems 282, Springer-Verlag, Berlin, 1987.

[23] D. F. SHANNO AND K. H. PHUA, *Matrix conditioning and nonlinear optimization*, Math. Programming, 14 (1978), pp. 145-160.

[24] R. TAPIA, *Diagonalized multiplier methods and quasi-Newton methods for constrained optimization*, J. Optim. Theory Appl., 22 (1977), pp. 135-194.

[25] ———, *On secant updates for use in general constrained optimization*, Math. Comp., 51 (1988), pp. 181-202.

SIAM J. OPTIMIZATION
Vol. 2, No. 2, pp. 242–263, May 1992

004

A STRUCTURE-EXPLOITING ALGORITHM FOR NONLINEAR MINIMAX PROBLEMS*

ANDREW R. CONN† AND YUYING LI‡

Abstract. In this paper, some basic concepts are generalised which characterise the best linear Chebyshev approximation in one variable to general nonlinear minimax problems. A new method for solving a nonlinear minimax problem is presented, which exploits the structure and characterisation of the solution whenever possible. The algorithm is globally convergent with a superlinear convergence rate. Numerical results indicate the efficacy of the new method.

Key words. nonlinear Chebyshev approximation

AMS(MOS) subject classifications. 41A50, 65D99, 65F20, 65K05

1. Introduction. We want to solve a *discrete nonlinear minimax problem*, which is written as

$$\min_{x\in\Re^n}\max_{i\in M} f_i(x), \tag{1.1}$$

where M is a finite index set. This is equivalent to finding the minimum value for the *maximum function* $\psi(x) = \max_{i\in M} f_i(x)$.

It is clear that a discrete Chebyshev problem

$$\min_{x\in\Re^n}\max_{1\le i\le m} |f_i(x)|, \tag{1.2}$$

which is a major class of discrete minimax problems, could be regarded as a special case of a general minimax problem (1.1) with

$$M = \{1, 2, \cdots, m, m+1, \cdots, 2m\}, \quad f_{i+m}(x) = -f_i(x), \quad i = 1, \cdots, m.$$

For simplicity, we describe our algorithm mainly in terms of the discrete Chebyshev problem (1.2) written in the form of (1.1). The extensions required for the general problem (1.1) are mentioned. In this paper, we are content to find a local minimum of (1.1) and we assume that a local minimum for (1.1) always exists. We also assume that each $f_i(x)$ is twice continuously differentiable.

Numerical methods for the discrete nonlinear Chebyshev/minimax problem are less prolific than for the linear problem. It is well known that the maximum function, $\psi(x) = \max_{i\in M} f_i(x)$, is not differentiable at kinks that arise whenever $f_i(x) = f_j(x)$, $i, j \in M$, $i \neq j$. Therefore, traditional gradient-type methods cannot be applied directly.

The existing methods are essentially based on successive linear programming or nonlinear programming techniques applied to an equivalent nonlinear programming problem. Examples include [1], [13], [20], [22], [23], [24], [27], [29], [30], [36], and [38].

* Received by the editors February 15, 1990; accepted for publication (in revised form) May 17, 1991.

† Thomas J. Watson Research Center, P.O.Box 218, Yorktown Heights, New York 10598. The research of this author was supported in part by Natural Sciences and Engineering Research Council of Canada grant A8639.

‡ Computer Science Department, Cornell University, Upson Hall, Ithaca, New York 14853. The research of this author was partially supported by the United States Army Research Office through the Mathematical Science Institute, Cornell University.

The classical Chebyshev theory provides us with characterisations for the best linear Chebyshev approximation. These properties uniquely determine a solution in many instances and thus requiring approximations with these special features is likely to result in a more efficient technique. Indeed, such has been the experience with classical Remez algorithms for best continuous/discrete linear Chebyshev approximation (see, for example, [35]).

In this paper, we first generalise the characterisation of the best linear Chebyshev approximation to a solution of a *nonlinear minimax problem.* The generalisation is useful computationally because we can force the approximate solutions to have these properties and thus expedite the solution-finding process. This is particularly beneficial for those problems arising from the discretisation of continuous approximation problems.

In developing our algorithm, we determine a suitable descent direction based on the structure of a solution (which consists of functions in general that are not necessarily close to the current maximum function). The new approach proposed is different from the existing methods in that one attempts to use the structure and characterisation of a solution of the minimax problem explicitly.

For clarity and brevity, we omit the proofs of some theorems. The interested reader is referred to [26] for details.

2. Structure of solutions to minimax problems.

The continuous Chebyshev approximation problem on an interval $[a,b]$ can be described as

$$\min_{x\in\Re^n}\max_{t\in[a,b]}|f(t)-\phi(x,t)|,$$

where $f(t)$ and $\phi(x,t)$ are given functions.

Assume $\phi(x,t)=\sum_{i=1}^n x_i\phi_i(t)$. It is well known that, under the Haar condition, the absolute error function $|f(t)-\sum_{i=1}^n x_i\phi_i(t)|$ of the best linear Chebyshev approximation achieves the maximum value on $n+1$ points with the signs of the errors alternating [31]. Any ordered $n+1$ distinct points have been termed a *reference* and an approximation with the errors alternating signs on a reference has been called a *reference function* [35]. If a reference function has the same magnitude of errors on the reference, it is further called a *levelled* reference function.

The famous Remez algorithm finds the best Chebyshev approximation by constructing levelled reference functions at each step until a levelled reference function with the maximum error is obtained. For discrete linear Chebyshev approximations, the concept of reference and reference function has proven to be useful in developing computationally efficient algorithms (e.g., [4] and [6]).

Under some conditions, $|f(t)-\phi(x,t)|$ of the best nonlinear Chebyshev approximation achieves the maximum value at k points with the signs of the errors alternating [32]. Since the conditions are rather restrictive and k is not known a priori, there seems to be no computational algorithm that attempts to exploit the structure of the solution for a nonlinear Chebyshev problem.

In this section, we introduce the concepts of cadre and reference set for nonlinear minimax problems. They are generalisations of the corresponding concepts for linear Chebyshev problems.

DEFINITION 2.1. The vector set $\mathcal{C}=\{\nabla f_{i_j}\}_{j=0}^l$ is called a *cadre* if and only if:

1. $\text{rank}([\nabla f_{i_0},\cdots,\nabla f_{i_l}])=l$;
2. for any $\{\nabla f_{j_1},\cdots,\nabla f_{j_l}\}\subset\mathcal{C}$, $\text{rank}([\nabla f_{j_1},\cdots,\nabla f_{j_l}])=l$.

This term was used by Descloux [18] to describe a linear Chebyshev solution when the Haar condition is not satisfied. A cadre can be equivalently defined by the following lemma.

LEMMA 2.2. *$\mathcal{C} = \{\nabla f_{i_j}\}_{j=0}^{l}$ is a cadre if and only if* $\operatorname{rank}(\mathcal{C}) = l$ *and there exist multipliers $\{\lambda_i\}$ such that*

$$\sum_{j=0}^{l} \lambda_j \nabla f_{i_j} = 0 \quad \textit{and} \quad \lambda_j \neq 0, \qquad j = 0, \cdots, l. \tag{2.1}$$

We refer to $\{\lambda_j\}$, normalised by $\sum_{j=0}^{l} \lambda_j = 1$, if $\sum_{j=0}^{l} \lambda_j \neq 0$ and $\lambda_0 = 1$ otherwise, as cadre multipliers*. The relation* (2.1) *is also called the* characteristic relation (*cf.* [31]).

Cadre multipliers are different from the Lagrangian multipliers used in optimization. The Lagrangian multipliers are usually associated with a stationary point and, under certain nondegeneracy assumptions, the nonzero multipliers are associated only with activities (see the following page for a definition of this term and the term ϵ-active). The cadre multipliers, however, are defined for any cadre and the functions in a cadre are not necessarily ϵ-active. Hence, we deliberately use the term cadre multipliers instead of just multipliers in order to differentiate them from the Lagrangian multipliers.

DEFINITION 2.3. The functions $\{f_{i_j}(x)\}_{j=0}^{l}$ are said to be locally forming a *reference set* of a minimax problem (1.1) if $\mathcal{C} = \{\nabla f_{i_j}\}_{j=0}^{l}$ is a cadre such that

1. The cadre multipliers $\{\lambda_j\}_{j=0}^{l}$ satisfy $\lambda_j > 0$, $j = 0, \cdots, l$;
2. The functions $\{f_{i_j}(x)\}_{j=0}^{l}$ all have the same sign.

The reference set is further called a *levelled reference set* if the value of each function is the same, viz.,

$$f_{i_j}(x) = f_{i_k}(x) \qquad \text{for any } i_j, i_k \in \mathcal{C}.$$

From the optimality conditions of (1.2) (e.g., [37]), we obtain an equivalent characterisation for a local minimum of (1.2) that relates to the structure of the best linear Chebyshev approximation.

THEOREM 2.4. *Suppose x^* is a local minimum for a minimax problem* (1.1). *Then, there exists a set of $l+1$ functions $\{f_{i_j}(x)\}_{j=0}^{l}$, which is a levelled reference set at x^* on the cadre $\mathcal{C} = \{\nabla f_{i_j}(x^*)\}_{j=0}^{l}$ with the maximum deviation.*

A reference set is a generalisation of the alternating sign property of a best Chebyshev approximation. Our experience with the numerical methods for linear l_∞ problems [6] suggests that it is very important to exploit computationally the above properties of a solution. The algorithm proposed in this paper is developed under this principle.

3. The model algorithm. The proposed algorithm is a descent method with a line search. The special features of the suggested algorithm, however, are that the search directions always decrease the maximum function and attempt to enforce the characterisation of a solution at the same time. Since a levelled reference set with the maximum deviation characterises a solution to a minimax problem, we attempt to compute the solution by constructing approximate solutions with such properties.

Assume $\mathcal{W} = \{i_0, i_1, \cdots, i_l\}$ is an index set and all the functions in $\mathcal{W}$ form a reference set that is not levelled. Denote

$$A = [\nabla f_{i_0} - \nabla f_{i_1}, \cdots, \nabla f_{i_0} - \nabla f_{i_l}],$$
$$\Phi(x)^T = [f_{i_0}(x) - f_{i_1}(x), f_{i_0}(x) - f_{i_2}(x), \cdots f_{i_0}(x) - f_{i_l}(x)],$$

and $i_0 \in \mathcal{A}(x, 0)$. Here, $\mathcal{A}(x, 0)$ denotes the indices of the *active functions*, which are the functions achieving the maximum value at the current point x. In other words, $\mathcal{A}(x, 0) = \{i \in M | \psi(x) = f_i(x)\}$. More generally, we define the set of ϵ-active functions $\mathcal{A}(x, \epsilon)$ to be the set of functions that achieve the maximum deviation within a tolerance of ϵ, a small positive constant that may be reduced by the algorithm. That is, $\mathcal{A}(x, \epsilon) = \{i \in M | \psi(x) - f_i(x) \leq \epsilon\}$.

From the following two lemmas, it is possible to determine descent directions that attempt to construct a levelled reference set in the neighbourhood of a cadre or reference set.

LEMMA 3.1. *Suppose the functions in $\mathcal{W}$ form a reference set that includes all the current active functions. Then, the direction defined from $\mathcal{W}$ by*

$$v = -A(A^T A)^{-1}\Phi(x) \tag{3.1}$$

is a descent direction for all the active functions provided the reference set is not levelled.

In [29], a similar result, that the vertical direction v is a descent direction when the Lagrangian multipliers are nonnegative, is stated.

If a unit step along v is taken, $\Phi(x) + A^T v = 0$. Thus the functions in $\mathcal{W}$ would all have the same value as the representative function, a function chosen from $\mathcal{A}(x, \epsilon)$ at the start of the iteration, up to first order.

LEMMA 3.2. *Suppose $\mathcal{C} = \{\nabla f_{i_0}, \nabla f_{i_1}, \cdots, \nabla f_{i_l}\}$ is a nonreference set cadre with cadre multipliers $\{\lambda_j\}_{j=0}^{l}$ summing to one and $f_{i_0}(x)$ achieves the current maximum deviation for* (1.1). *Then, the direction v defined on $\mathcal{W} = \{i_0, i_1, \cdots, i_l\}$ by*

$$[\nabla f_{i_0} - \sigma_0 \sigma_j \nabla f_{i_j}]^T v = -(f_{i_0} - \sigma_0 \sigma_j f_{i_j}), \quad i_j \in \mathcal{W}, i_j \neq i_0, \ \sigma_j = \operatorname{sgn}(f_{i_j}), \tag{3.2}$$

decreases all the active functions, assuming $\mathcal{W}$ includes all the active functions at x.

If $\{f_{i_j}\}_0^l$ are linear functions at $x+v$, $\{f_{i_j}\}_0^l$ form a reference set. Thus, whenever the f_{i_j}'s do not constitute a reference set, moving along v, which is defined by (3.2),

$$[\nabla f_\mu - \sigma_0 \sigma_j \nabla f_{i_j}]^T v = -(f_\mu - \sigma_0 \sigma_j f_{i_j}), \qquad i_j \in \mathcal{W}, \quad i_j \neq i_0,$$

attempts to construct such a set.

We build up cadres using the concept of working sets. A working set is a function index set that includes all the indices of the current maximum functions. We emphasize, however, that the working set $\mathcal{W}$ is not generally an active set. In §5, we describe the details of setting up a working set.

The search direction is determined from the working set. If a cadre has not been located, in addition to decreasing the maximum function, the search direction is constructed to level the functions in the working set, when this is possible. The motivation behind this levelling comes from the fact that the structure of the solution requires the error curve to be levelled on the extreme points.

The suggested model algorithm is now outlined.

MODEL ALGORITHM

```
Step 1. Suppose an initial point $x^0$ is given. Set $k \leftarrow 0$.
Step 2. [Set up a working set]
        The new working set $\mathcal{W}^k$ is determined. Check if there is a cadre $\mathcal{C}^k$
        whose indices form a subset of $\mathcal{W}^k$. If there is no such cadre, go to
        Step 4.
Step 3. [Construct a levelled reference set]
        Check reference set conditions. If the cadre corresponds to a reference
        set, compute a descent direction by levelling the reference set. Otherwise,
        find a descent direction that attempts to construct a reference
        set. Go to Step 5.
Step 4. [Descend and level]
        A search direction $d^k$ is found that decreases all the $\epsilon$-active functions
        and levels the functions in the working set $\mathcal{W}^k$, if possible.
Step 5. [Line search]
        A line search is performed on $\psi(x)$ along the direction $d^k$
            $x^{k+1} \leftarrow x^k + \lambda^k d^k; \quad k \leftarrow k+1.$
Step 6. [Termination]
        If optimal, stop. Otherwise, go to Step 2.
```

Step 3 of the model algorithm is one of the major parts in which the characterisation of the solution is exploited. From (3.1) and (3.2), we can compute a descent direction when a cadre is located (see also §6). Next, we discuss how to identify cadres (§4), how to construct a working set (§5), and how to compute a search direction when there is no cadre (§6). We also present details of the computation, including handling degeneracy (§8).

4. Identifying cadres. Given a set of functions $\{f_{i_0}, \cdots, f_{i_l}\}$, we discuss whether there exists a cadre within this set. We divide cadres into two types, depending upon whether

$$\sum_{j=0}^{l} \lambda_j = 1 \quad \text{or} \quad \sum_{j=0}^{l} \lambda_j = 0,$$

where $\{\lambda_j\}_{j=0}^{l}$ are cadre multipliers. The cadre that defines a reference set always belongs to the first type.

It is straightforward to prove the following lemma.

LEMMA 4.1. *Suppose $\{\nabla f_{i_0} - \nabla f_{i_1}, \cdots, \nabla f_{i_0} - \nabla f_{i_l}\}$ are linearly independent. Then, the rank of the vector set $\{\nabla f_{i_0}, \nabla f_{i_1}, \cdots, \nabla f_{i_l}\}$ is at least l.*

The following lemma gives, under certain assumptions, necessary and sufficient conditions for the existence of a cadre with the sum of cadre multipliers being zero.

LEMMA 4.2. *Suppose $A = [\nabla f_{i_0} - \nabla f_{i_1}, \cdots, \nabla f_{i_0} - \nabla f_{i_{l-1}}]$ is of full rank and that $Z^T \nabla f_{i_0} \neq 0$, where the columns of Z form a basis for the null space of A^T. Then, there exists a cadre $\mathcal{C} \subseteq \{\nabla f_{i_0}, \nabla f_{i_1}, \cdots, \nabla f_{i_l}\}$ with cadre multipliers summing to zero if and only if $[\nabla f_{i_0} - \nabla f_{i_1}, \cdots, \nabla f_{i_0} - \nabla f_{i_l}]$ is rank deficient.*

Proof. Suppose $\mathcal{C} = \{\nabla f_{k_0}, \cdots, \nabla f_{k_\nu}\}$ is a cadre and $\{k_0, k_1, \cdots, k_\nu\} \subseteq \{i_0, i_1, \cdots, i_l\}$ with

$$\sum_{j=0}^{\nu} \lambda_j \nabla f_{k_j} = 0, \quad \sum_{j=0}^{\nu} \lambda_j = 0, \quad \lambda_j \neq 0, \quad j = 0, \cdots, \nu.$$

Then it is obvious that

$$\sum_{j=0}^{\nu} \lambda_j (\nabla f_{i_0} - \nabla f_{k_j}) = 0. \tag{4.1}$$

From (4.1) and the assumption that $\{\nabla f_{i_0} - \nabla f_{i_1}, \cdots, \nabla f_{i_0} - \nabla f_{i_{l-1}}\}$ are linearly independent, we know that $i_l \in \{k_0, \cdots, k_\nu\}$. Hence, $\lambda_l \neq 0$ and we have

$$(\nabla f_{i_0} - \nabla f_{i_l}) = \sum_{j=1}^{l-1} \hat{\lambda}_j (\nabla f_{i_0} - \nabla f_{i_j}),$$

after padding with zeros if necessary. On the other hand, if we assume that $\{\nabla f_{i_0} - \nabla f_{i_1}, \cdots, \nabla f_{i_0} - \nabla f_{i_{l-1}}\}$ are linearly independent and $\{\nabla f_{i_0} - \nabla f_{i_1}, \cdots, \nabla f_{i_0} - \nabla f_{i_{l-1}}, \nabla f_{i_0} - \nabla f_{i_l}\}$ are linearly dependent, we have

$$\nabla f_{i_0} - \nabla f_{i_l} = \sum_{j=1}^{l-1} \hat{\lambda}_j (\nabla f_{i_0} - \nabla f_{i_j}). \tag{4.2}$$

From Lemma 4.1 and the assumption that A is full rank, we have that

$$\operatorname{rank}(\{\nabla f_{i_0}, \nabla f_{i_1}, \cdots, \nabla f_{i_{l-1}}\}) \geq l - 1.$$

Moreover, from $Z^T \nabla f_{i_0} \neq 0$, and the argument that follows, we can conclude that

$$\operatorname{rank}(\{\nabla f_{i_0}, \nabla f_{i_1}, \cdots, \nabla f_{i_{l-1}}\}) = l. \tag{4.3}$$

The above is true because, if $\{\nabla f_{i_0}, \nabla f_{i_1}, \cdots, \nabla f_{i_{l-1}}\}$ are linearly dependent, then there exist $\{\lambda_j\}$ that are not all zero such that

$$\sum_{j=0}^{l-1} \lambda_j \nabla f_{i_j} = 0.$$

If $\sum_{j=0}^{l-1} \lambda_j \neq 0$, without loss of generality, we can assume $\sum_{j=0}^{l-1} \lambda_j = 1$. Thus $\lambda_0 = 1 - \sum_{j=1}^{l-1} \lambda_j$. Hence

$$\nabla f_{i_0} = \sum_{j=1}^{l-1} \lambda_j (\nabla f_{i_0} - \nabla f_{i_j}).$$

We conclude that $Z^T \nabla f_{i_0} = 0$, which is a contradiction.

If $\sum_{j=0}^{l-1} \lambda_j = 0$, we have $\lambda_0 = -\sum_{j=1}^{l-1} \lambda_j$. Hence

$$\sum_{j=1}^{l-1} \lambda_j (\nabla f_{i_0} - \nabla f_{i_j}) = 0,$$

which is again a contradiction to the assumption that A is full rank.

Thus, using (4.2), we obtain

$$\sum_{j=0}^{l} \hat{\lambda}_j \nabla f_{i_j} = 0 \quad \text{and} \quad \sum_{j=0}^{l} \hat{\lambda}_j = 0, \tag{4.4}$$

where $\hat{\lambda}_0 = 1 - \sum_{j=1}^{l-1} \hat{\lambda}_j, \hat{\lambda}_l = -1$.

Define $\mathcal{C} = \{ \nabla f_{i_j} \mid \hat{\lambda}_j \neq 0 \ , \ j = 0, \cdots, \nu\}$. Using (4.4),

$$\mathrm{rank}(\mathcal{C}) \leq |\mathcal{C}| - 1.$$

From (4.3), we know that

$$\mathrm{rank}(\mathcal{C}) \geq |\mathcal{C}| - 1.$$

Hence

$$\mathrm{rank}(\mathcal{C}) = |\mathcal{C}| - 1. \tag{4.5}$$

Moreover,

$$\hat{\lambda}_j \neq 0, \quad \nabla f_{i_j} \in \mathcal{C} \quad \text{with} \sum_{\nabla f_{i_j} \in C} \hat{\lambda}_j = 0.$$

Using Lemma 2.2, $\mathcal{C}$ is a cadre with the sum of the cadre multipliers being zero. □

Now, we present a lemma that tells us how to identify cadres with cadre multipliers summing to one.

LEMMA 4.3. *Suppose $\{\nabla f_{i_0} - \nabla f_{i_1}, \cdots, \nabla f_{i_0} - \nabla f_{i_l}\}$ are linearly independent. Then there exists a cadre $\mathcal{C} \subseteq \{\nabla f_{i_0}, \nabla f_{i_1}, \cdots, \nabla f_{i_l}\}$ with cadre multipliers summing to one if and only if the orthogonal projected gradient, $Z^T \nabla f_{i_0}$, is zero, where*

$$A = [\nabla f_{i_0} - \nabla f_{i_1}, \cdots, \nabla f_{i_0} - \nabla f_{i_l}], \qquad Z^T A = 0.$$

Proof. Since $\{\nabla f_{i_0} - \nabla f_{i_1}, \cdots, \nabla f_{i_0} - \nabla f_{i_l}\}$ are linearly independent, using Lemma 4.1,

$$\mathrm{rank}(\{\nabla f_{i_0}, \nabla f_{i_1}, \cdots, \nabla f_{i_l}\}) \geq l. \tag{4.6}$$

The orthogonal projection of ∇f_{i_0} on the null space of A^T is $Z^T \nabla f_{i_0}$. The vector $Z^T \nabla f_{i_0}$ is zero if and only if there exist $\{\lambda_j\}_{j=0}^{l}$ such that

$$\lambda_0 \nabla f_{i_0} + \sum_{j=1}^{l} \lambda_j \nabla f_{i_j} = 0, \qquad \sum_{j=0}^{l} \lambda_j = 1. \tag{4.7}$$

Suppose (4.7) is satisfied. From (4.6) and (4.7), $\mathrm{rank}(\{\nabla f_{i_0}, \nabla f_{i_1}, \cdots, \nabla f_{i_l}\}) = l$. Let $\mathcal{C} = \{\nabla f_{i_j} \mid \lambda_j \neq 0, j = 0, 1, \cdots, l\}$. Then, as in the argument for (4.5), $\mathcal{C}$ has rank $|\mathcal{C}| - 1$. From Lemma 2.2, $\mathcal{C}$ is a cadre. Moreover, the sum of the cadre multipliers is one.

On the other hand, if there is a cadre $\mathcal{C} \subseteq \{\nabla f_{i_0}, \nabla f_{i_1}, \cdots, \nabla f_{i_l}\}$ with cadre multipliers summing to one, then, following Lemma 2.2, there exist $\{\lambda_j\}$ such that (4.7) holds and then, $Z^T \nabla f_{i_0} = 0$. □

Lemmas 4.2 and 4.3 together enable us to determine whether there exists a cadre.

5. Establishment of the working set. A working set is a function index set, which is used to determine the current descent direction. Since we want the search direction to decrease all the ϵ-active functions, this working set $\mathcal{W}^k$ is chosen to include all the ϵ-active functions at the current point x^k. Nonetheless, there is flexibility in constructing such a set. We have chosen to build up the working set by selecting the functions that determine the maximum function through several iterations. This is motivated by the fact that it is the extreme points that are important in determining the best approximation in the Chebyshev sense. Thus we require that

$$\mathcal{W}^k \subseteq \mathcal{W}^{k-1} \cup \mathcal{A}(x^k, \epsilon), \qquad \mathcal{W}^0 = \emptyset. \tag{5.1}$$

Moreover, the current ϵ-active functions are given priority over the functions in the old working set when forming the new working set.

However, since adjustment of the functions in the working set is necessary when the current working set is not approaching a reference set (essentially to account for the alternating sign property) we use $\hat{\mathcal{W}}^k$ to denote the set after possible modification and the rules for changing the set will be described precisely later. Thus, we more correctly require

$$\mathcal{W}^k \subseteq \hat{\mathcal{W}}^{k-1} \cup \mathcal{A}(x^k, \epsilon). \tag{5.2}$$

Assume, at the kth iteration, that a *representative function* $f_\mu(x)$, which can be any function $f_\mu(x)$ such that $\mu \in \mathcal{A}(x^k, \epsilon)$, is selected. Suppose $\mathcal{W}^k = \{\mu, i_1, \cdots, i_l\}$. The following Jacobian matrix corresponding to $\mathcal{W}^k$,

$$A^k = [\nabla f_\mu - \nabla f_{i_1}, \cdots, \nabla f_\mu - \nabla f_{i_l}], \tag{5.3}$$

is required *numerically* to have full rank. More specifically, our implementation accounts for this numerical rank. Conceptually it is equivalent to having some tolerance on the smallest singular value of A^k.

In implementation, we consider the projected gradient $Z^T \nabla f_\mu$ numerically zero if

$$\|Z^T \nabla f_\mu(x^k)\| \leq \tau_c^k,$$

where the columns of Z are an orthonormal basis for the null space of ${A^k}^T$ and τ_c^k is a small positive constant. Hence, if we identify cadres according to Lemma 4.3, we have a *near* cadre.

Since we need the QR decomposition (see, for example, [21, Chap. 6]) of the matrix A^k in computing the direction (see § 6), we build up the current working set $\mathcal{W}^k$ as follows.

CONSTRUCT $\mathcal{W}^k$:

Step 1. Set $Q \leftarrow I_{n \times n}$, $\mathcal{W}^k \leftarrow \{\mu\}$, where $\mu \in \mathcal{A}(x^k, \epsilon)$. $t \leftarrow 0$. $\hat{\mathcal{A}} = \mathcal{A}(x^k, \epsilon)$.

Step 2. If $\hat{\mathcal{A}} \setminus \mathcal{W}^k = \emptyset$, go to Step 3. Otherwise, let Q_2 be the last $n - t$ columns of Q and $j \in \hat{\mathcal{A}} \setminus \mathcal{W}^k$. If $\|Q_2{}^T(\nabla f_\mu - \nabla f_j)\| \leq \tau_0$, set $\hat{\mathcal{A}} = \hat{\mathcal{A}} \setminus \{j\}$, go to Step 2. Otherwise, go to Step 4.

Step 3. If $\hat{\mathcal{W}}^{k-1} \setminus \mathcal{W}^k = \emptyset$, stop. Otherwise, let Q_2 be the last $n - t$ columns of Q. If $\|Q_2^T \nabla f_\mu\| \leq \tau_c^k$, stop. Otherwise choose $j \in \hat{\mathcal{W}}^{k-1} \setminus \mathcal{W}^k$. If $\|Q_2{}^T(\nabla f_\mu - \nabla f_j)\| \leq \tau_0$, set $\hat{\mathcal{W}}^{k-1} = \hat{\mathcal{W}}^{k-1} \setminus \{j\}$ and go to Step 3. Otherwise, continue.

Step 4. Let $a = \nabla f_\mu - \nabla f_j$. Add the column a to A^k and update Q and R accordingly. Set:

$$A^k \leftarrow [A^k, a], \quad \mathcal{W}^k \leftarrow \mathcal{W}^k \cup \{j\}, \quad t \leftarrow t + 1.$$

Go to Step 2.

Thus the working set is the largest subset of $\hat{\mathcal{W}}^{k-1} \cup \mathcal{A}(x^k, \epsilon)$ (largest in the sense of the corresponding Jacobian matrix A^k being full rank), where the indices of the current ϵ-active functions have been entered preferentially.

Following the procedure of constructing a working set, it is clear that, if the current point is nondegenerate (a current point x^k is degenerate when there is a cadre $\mathcal{C} = \{\nabla f_{i_0}, \nabla f_{i_1}, \cdots, \nabla f_{i_l}\}$ such that $\{i_0, i_1, \cdots, i_l\} \subset \mathcal{A}(x^k, 0)$) and there is no cadre with cadre multipliers summing to zero, the Jacobian corresponding to all the ϵ-active functions is of full rank. Therefore

$$\mathcal{A}(x^k, \epsilon) \subseteq \mathcal{W}^k.$$

Moreover, if $\|Z^T \nabla f_\mu\| \le \tau_c^k$, where $Z = Q_2$ for some Q, then a cadre (or a near cadre) with cadre multipliers summing to one is found.

6. Determining the search direction. Assume the working set at the current point x_c is

$$\mathcal{W}(x_c) = \{i_0, \cdots, i_l\} \quad \text{and} \quad \mu = i_0.$$

The desired search direction, in addition to being one of descent, attempts to enforce the characterisation of a solution.

Before a cadre with multipliers adding to one is located, we would like the search direction to decrease all the active functions and level all the functions in the working set, if possible. It is clear that $d = x - x_c$, where x attempts to solve

$$\begin{aligned} &\min_{x \in \Re^n} f_\mu(x) \\ &\text{subject to} \\ &f_\mu(x) - f_{i_j}(x) = 0, \qquad i_j \in \mathcal{W}(x_c), \end{aligned} \tag{6.1}$$

in the required direction. Note that μ is in fact a function of x_c and we use it to denote the current representative function as long as no confusion arises.

Dropping the subscript on x_c to simplify the description, one may approximate (6.1) as follows:

$$\begin{aligned} &\min_{d \in \Re^n} \nabla f_\mu(x)^T d + \tfrac{1}{2} d^T G d \\ &\text{subject to} \\ &\Phi(x) + A^T d = 0, \end{aligned} \tag{6.2}$$

where

$$\begin{aligned} A &= [\nabla f_\mu(x) - \nabla f_{i_1}(x), \nabla f_\mu(x) - \nabla f_{i_2}(x), \cdots, \nabla f_\mu(x) - \nabla f_{i_l}(x)], \\ \Phi(x) &= [f_\mu(x) - f_{i_1}(x), f_\mu(x) - f_{i_2}(x), \cdots, f_\mu(x) - f_{i_l}(x)]^T, \end{aligned}$$

and G is a matrix such that $Z^T G Z$ is positive definite, where the columns of Z form an orthonormal basis for the null space of A^T.

When close to a stationary point, $Z^T G Z$ is chosen to contain the curvature information of the functions in the working set in the null space of A^T (see §7 for details).

From the construction of the working set $\mathcal{W}(x)$, we know that A is of full rank. Following [11], the solution to (6.2) may be written as

$$\begin{aligned} d &= \hat{h} + v, \\ \hat{h} &= -Z(Z^TGZ)^{-1}Z^T(\nabla f_\mu(x) + Gv), \\ v &= -A(A^TA)^{-1}\Phi(x). \end{aligned}$$

It has been suggested in [11] that one could ignore the computation of Z^TGv altogether without significantly effecting the rate of convergence. In this case, an approximate solution to (6.2) can be written as

$$d = h + v,$$

where

$$\begin{aligned} h &= -ZB^{-1}Z^T(\nabla f_\mu(x)), \\ v &= -A(A^TA)^{-1}\Phi(x), \end{aligned} \tag{6.3}$$

and

$$B = Z^TGZ.$$

It is clear that h is in the null space of A^T while v is in the range space of A. The direction in the null space of A^T will be called the *horizontal direction* and the direction in the range space of A will be called the *vertical direction.* We also point out that, given $\mathcal{W}$, Z, and B, the value of h and v is independent of the choice of μ (see [15] for details).

We now prove that a nonzero horizontal direction h is a descending direction for all the functions in $\mathcal{W}$.

LEMMA 6.1. *Assume $\mathcal{W}$ is the working set that defines the search direction. Assume further that B is positive definite and that there is no cadre $\mathcal{C} = \{\nabla f_{i_0}, \cdots, \nabla f_{i_l}\}$, with the cadre multipliers summing to one, such that $\{i_0, \cdots, i_l\} \subseteq \mathcal{W}$. Then the horizontal direction decreases all the functions in $\mathcal{W}$ equally (up to the first order); otherwise (i.e., there exists a cadre with the cadre multipliers summing to one), the horizontal direction h defined from $\mathcal{W}$ is zero.*

Proof. The horizontal direction defined in (6.3) is

$$h = -ZB^{-1}Z^T(\nabla f_\mu(x)), \qquad \mu = i_0,$$

where $Z^TZ = I_{n-l}$, $A^TZ = 0$. Since B is positive definite and

$$h^T\nabla f_\mu(x) = -(Z^T\nabla f_\mu(x))^TB^{-1}(Z^T\nabla f_\mu(x)),$$

it follows that

$$h^T\nabla f_\mu(x) < 0 \quad \text{iff } Z^T\nabla f_\mu \neq 0.$$

Since there is no cadre $\mathcal{C} = \{\nabla f_{i_0}, \cdots, \nabla f_{i_l}\}$ with the cadre multipliers summing to one such that $\{i_0, \cdots, i_l\} \subseteq \mathcal{W}$, we have, from the definition of $\mathcal{W}$ and Lemma 4.3, $Z^T\nabla f_\mu \neq 0$ and h is a descent direction for the representative function $f_\mu(x)$.

Furthermore, since

$$A^Th = 0 \quad \text{and} \quad \nabla f_{i_j}{}^T h = \nabla f_\mu{}^T h, \quad i_j \in \mathcal{W},$$

any function in the working set $\mathcal{W}$ will be decreased by the same amount (up to first order) as the representative function f_μ.

On the other hand, assuming there exists a cadre with cadre multipliers summing to one, by Lemma 4.3, the result follows. □

In conclusion, the horizontal direction h is a projection of the negative gradient of the representative function onto the null space of A^T. It is always a descent direction as long as $\mathcal{W}$ is not a cadre with cadre multipliers summing to one. As a descent direction, it decreases the functions in the working set by the same amount (up to first order). The horizontal direction h defined on the cadre with the cadre multipliers summing to one is always zero.

No cadre. When a cadre is not located, vertical directions are descent directions in most cases.

Whenever this is the situation, we perform the levelling process, i.e., set the search direction $d = v + h$. In the case in which the vertical direction is ascending, the vertical direction is discarded and the horizontal direction alone is taken as the search direction; specifically, we define

$$d^k = \begin{cases} h^k + v^k & \text{if } \nabla {f_\mu}^T v^k < 0, \\ h^k & \text{otherwise.} \end{cases} \tag{6.4}$$

Our numerical experience shows that an ascent vertical direction is a rare occurrence. This may be explained by the fact that the working set is constructed to approach a reference set. In the event that ascent does occur, we consider this as an indication that the working set is not approaching a reference set. This may be caused by some function, which will eventually not be maximum, being included in $\mathcal{W}^k$. Thus the next working set will not always include all the functions of the current working set; instead, we define

$$\begin{aligned} &\hat{\mathcal{W}}^k \leftarrow \mathcal{W}^k \setminus I^+, \quad \text{if } \nabla f_\mu^T v^k \geq 0, \text{ where} \\ &I^+ = \begin{cases} \{j_0\} & \text{if } \mathcal{A}(x^k, \epsilon) \subset \mathcal{W}^k \text{ and } f_\mu - f_{j_0} = \max_{j \in \mathcal{W}^k}(f_\mu - f_j); \\ \emptyset & \text{otherwise.} \end{cases} \end{aligned} \tag{6.5}$$

A cadre is located. If there exists a cadre with multipliers summing to zero, the cadre does not correspond to a reference set. In this case, although v corresponds to levelling, we emphasize decreasing the maximum function. In particular, it is not necessarily desirable to level functions that do not correspond to a reference set. Thus we simply take $d^k = h^k$. (Note that $h \neq 0$, since there is no cadre with cadre multipliers summing to one.)

If the functions in the working set, $\mathcal{W}^k$, form a (near) reference set, the vertical direction v^k defined by (6.3) attempts to level the functions in the working set while the horizontal direction h^k (again defined by (6.3)—$h^k = 0$ only if $\mathcal{W}^k$ contains an *exact* cadre with cadre multipliers summing to one) makes the gradients approach an *exact* cadre. From Lemma 3.1, v^k is a descent direction. Thus $d^k = h^k + v^k$ is a descent direction (note that h^k is a descent direction).

Suppose a cadre with multipliers summing to one has been located within the working set. Then the vertical direction v defined by (3.2) is a descent direction for the maximum function. Moreover, we can write (3.2) as

$$\begin{aligned} &\hat{A}v = -\hat{\Phi} \quad \text{where} \\ &\hat{A} = [\nabla f_\mu - \sigma_0\sigma_1\nabla f_{i_1}, \cdots, \nabla f_\mu - \sigma_0\sigma_l\nabla f_{i_l}], \\ &\hat{\Phi} = [f_\mu - \sigma_0\sigma_1 f_{i_1}, \cdots, f_\mu - \sigma_0\sigma_l f_{i_l}]^T. \end{aligned} \tag{6.6}$$

We also modify the working set for the next iteration as follows. The cadre multipliers associated with the functions in the working set are used to construct the working set for the next iteration. The functions with positive multipliers are considered to be the functions which should be in the working set, i.e., the correct functions. For the functions with negative multipliers, we would like to put its negative function into the working set. However, because of nonlinearity and the fact that the cadre and reference set are both local properties, we prefer not to do so. Instead, the functions with negative multipliers are simply deleted from the working set, since the functions corresponding to negative multipliers will no longer remain ϵ-active when the direction v is taken and the multipliers sum to one. Thus we define

$$\hat{\mathcal{W}}^k \leftarrow \mathcal{W}^k \setminus \{ i_j \mid \lambda_j < 0 \}. \tag{6.7}$$

The multipliers are thus used as a means to construct the working set and more than one function may be removed.

If the functions in the working set are all active and the multipliers sum to one, moving along the vertical direction initially decreases all the functions with the negative multipliers faster (up to first order) than those with positive multipliers. This comes from the following lemma (for the proof, see [14]).

LEMMA 6.2. *Suppose $\mathcal{W} = \{\mu, i_1, \cdots, i_l\}$ consists only of indices of the currently active functions. Assume further that $C = \{\nabla f_\mu, \nabla f_{i_1}, \cdots, \nabla f_{i_l}\}$ is a cadre. Assume the direction v is determined from $\mathcal{W}$, as in* (3.2). *Then:*

1. *all the active functions with negative multipliers will be decreased more rapidly than all the other active functions, if the cadre multipliers sum to one, i.e., $\sum_{j=0}^{l} \lambda_j = 1$;*
2. *all the active functions are decreased equally* (*up to first order*) *provided the cadre multipliers sum to zero, i.e., $\sum_{j=0}^{l} \lambda_j = 0$.*

This corresponds to (possibly multiple) dropping of active functions for the equivalent nonlinear programming problem.

Now, consider a general nonlinear minimax problem written as

$$\min_{x \in \Re^n} \max_{i \in \{1, \cdots, m\}} f_i(x).$$

The search direction can be computed in exactly the same way except that the reference set, after a cadre has been located, could not be established as before. Since there exists no negative function of a given function, the vertical direction that determines which active functions should be dropped is not defined. Thus we now discuss how the definition of the vertical direction is modified for the general minimax problem.

If the current maximum deviation $\psi(x^k)$ is positive, we assume that for any given $f_i(x)$, there exists an imaginary $f_{i+m}(x) = -f_i(x)$. The working set $\mathcal{W}^k$ is chosen such that

$$-\psi(x^k) < f_{i_j}(x^k) \le \psi(x^k) \quad \text{for any } i_j \in \mathcal{W}^k.$$

Hence locally we can treat the problem as a Chebyshev problem and the vertical direction, defined as for the Chebyshev problem, is a descent direction.

If the current maximum deviation $\psi(x)$ is nonpositive, we define a descent direction in a way similar to a general nonlinear programming approach [13]. In this case, if there exists some cadre multiplier that is negative, we simply remove the corresponding single function from the working set and update the projection matrix and recompute the search direction from the new projector. Under the assumption of linear independence, this will give a descent direction [13].

7. Approximation of the Hessian. In order to obtain a horizontal descent direction at each iteration B^k, an $(n-l)\times(n-l)$ matrix is assumed to be sufficiently positive definite.

For problems whose solutions are on a smooth valley, i.e., the number of active functions is less than $n+1$, the second-order information from the nonlinear active functions becomes significant for the fast final convergence of the algorithm. When close to x^*, B^k should be a good approximation to the projected Lagrangian Hessian ${Z^k}^T G^k Z^k$, where $G^k = \sum_{j=0}^{l} \lambda_j^k \nabla^2 f_{i_j}(x^k)$, the columns of Z^k form a basis for the null space of ${A^k}^T$, and ${\lambda_j}^k$ is an approximation to the Lagrangian multipliers (which are defined by the first-order optimality conditions of the equivalent nonlinear programming problem; see, for example, [13] or [38]).

If we assume the second-order sufficiency conditions hold at x^* and let λ^k be a good approximation to the cadre multipliers λ^* at *a solution* x^* (which are equal to the Lagrangian multipliers at a solution), then the matrix ${Z^k}^T G^k Z^k$, for x^k sufficiently close to x^*, is positive definite, as follows from continuity arguments.

A first-order method, for example, of [10], solves the problem whose solution is at a vertex (i.e., with $n+1$ linear independent activities) with a fast asymptotic rate of convergence since, once the correct activities are determined, one is merely using Newton's method (or a quasi-Newton method) to determine the unique intersection of these activities, with the corresponding quadratic (or superlinear) rate of convergence. First-order directions are usually good descent directions when one is far away from a stationary point and the computation of a first-order direction is cheaper than a second-order direction.

We choose to use the first-order direction if it gives a good improvement in the sense of constructing reference sets. Computationally, we consider that the first-order direction fails to improve the establishment of reference sets when the working set has not been changed for γ consecutive iterations (this may be a result of having the correct set but in this case it is reasonable to want to accelerate convergence by using a second-order direction). We arbitrarily set $\gamma = 3$ in our implementation. When failure occurs, we use the second-order information of the representative function or of all the functions in the working set, depending on how close we are to a stationary point of the subproblem.

Let *ibase* denote the number of consecutive iterations for which the working set remains unchanged. Suppose ρ is a small positive constant used to measure the closeness to a stationary point. The matrix G^k may be set up as follows:

$$(7.1) \qquad G^k \begin{cases} \approx \nabla^2 f_\mu(x^k) & \text{if } \textit{ibase} \geq \gamma \text{ and } \|{Z^k}^T \nabla f_\mu\| > \rho, \\ \approx \sum_{j=0}^{l} \lambda_j^k \nabla^2 f_{i_j}(x^k) & \text{if } \textit{ibase} \geq \gamma \text{ and } \|{Z^k}^T \nabla f_\mu\| \leq \rho, \\ = I & \text{otherwise,} \end{cases}$$

where ${\lambda_j}^k$ is an approximation to the Lagrangian multipliers. We note that when $\|{Z^k}^T \nabla f_\mu\| \leq \rho$, it is reasonable to expect a suitable approximation to the Lagrangian multipliers.

Also, when $G^k = I$, the search direction is a first-order direction.

In our algorithm, however, we use a quasi-Newton method to update an approximation to the projected Hessian matrix B^k. Suppose Z^k is the orthogonal matrix such that ${Z^k}^T A^k = 0$, where A^k is defined as in (5.3). In the implementation, we have used the extended BFGS updating given below. B^k is initialised to be ${Z^k}^T G^k Z^k$

when necessary, where G^k is approximated according to (7.1) by finite differences. The extended BFGS updating follows:

$$B^{k+1} = B^k - \frac{1}{s_r^{k^T} B^k s_r^k} B^k s_r^k s_r^{k^T} B^k + \frac{1}{y_r^{k^T} s_r^k},$$

where

$$\begin{aligned} s_r^k &= Z^{k+1^T}(x^{k+1} - x^k), \\ y_r^k &= Z^{k+1^T} \nabla f_\mu(x^{k+1}) - Z^{k^T} \nabla f_\mu(x^k). \end{aligned}$$

Assume B^k is positive definite. Then B^{k+1} remains positive definite if $s_r^{k^T} y_r^k > 0$. For unconstrained minimization, this condition is ensured by a line search. For constrained minimization, however, it cannot be satisfied in general. We have chosen to skip the update if the above condition is not satisfied.

8. Degeneracy. For a discrete Chebyshev problem, degeneracy handling is an important part of a useful algorithm. This is because, for example, in the linear case, it is not unusual for many residuals to achieve the maximum deviation. In this section, we discuss the handling of degeneracy in our algorithm.

We define a current point x^k to be degenerate when there is a cadre $\mathcal{C} = \{\nabla f_\mu, \nabla f_{i_1}, \cdots, \nabla f_{i_l}\}$ such that $\{\mu, i_1, \cdots, i_l\} \subset \mathcal{A}(x^k, 0)$.

Denote

$$\mathcal{W}^k = \{\mu, i_1, \cdots, i_l\}, \qquad A^k = [\nabla f_\mu - \nabla f_{i_1}, \cdots, \nabla f_\mu - \nabla f_{i_l}].$$

If x^k is a degenerate point, the following difficulty may occur. There is more than one cadre $\mathcal{C} = \{\nabla f_\mu, \nabla f_{i_1}, \cdots, \nabla f_{i_l}\}$ satisfying $\mathcal{W}^k \subset \mathcal{A}(x^k, 0)$. Thus it may not be possible to define a search direction such that it decreases the functions in all the cadres, although we know how to define a descending direction on one cadre.

If we consider the cadres that correspond to subsets of active functions, then there can be three types of degenerate points:

Type A. There only exist cadres with cadre multipliers summing to zero;

Type B. There exists a unique cadre and its cadre multipliers sum to one;

Type C. There exists more than one cadre and at least one with cadre multipliers summing to one.

A point x^* is a stationary point if and only if there exists at least one reference set consisting of active functions only.

We identify cadres by a tolerance of τ^k; the (numerical) degeneracy identified depends on the tightness of ${\tau_c}^k$. Thus when degeneracy is encountered, we reduce it by

$$\tau_c^{k+1} \leftarrow \frac{{\tau_c}^k}{2}. \tag{8.1}$$

Numerically, the degeneracy of Type A can only occur when $\|Z^{k^T} \nabla f_\mu^k\| > \tau_c^k$ and $\mathcal{W}^k \subset \mathcal{A}(x^k, \epsilon)$ [15]. For the degenerate points of Type A, there cannot be any reference set consisting of only the active functions. This is because, for any reference set, each of the corresponding cadre multipliers is positive and the sum of them is one. Thus the current point cannot be optimal. For this type of degeneracy, the horizontal direction h defined on the current working set decreases all the ϵ-active functions, up to first order, by the same amount.

Degeneracy of Type B occurs when $\|Z^{k^T}\nabla f_\mu^k\| \le \tau_c^k$, $\mathcal{A}(x^k,\epsilon) = \mathcal{W}^k$, and there exists zero multipiers [15]. For the degenerate points of Type B, it is possible that a reference set exists within the active set. If there is such a reference set, then the current point is already a stationary point. Otherwise, since there exists a unique cadre, the vertical direction v defined on the cadre by (3.2) attempts to construct a levelled reference set. Moreover, other maximum functions not in the cadre can also be decreased at the same time.

If $\|Z^{k^T}\nabla f_\mu^k\| \le \tau_c^k$ and $\mathcal{W}^k \subset \mathcal{A}(x^k,\epsilon)$, degeneracy of Type C occurs [15]. For the degenerate points of Type C, we do not know how to determine a descent direction without additional computation. Following a similar approach to [7] and [17], we solve the least squares problem:

$$\begin{aligned} &\min_{\theta\in\Re^{l+1}} \left\| \sum_{j=0}^{l} \theta_j \nabla f_{i_j} \right\|_2 \\ &\text{subject to} \\ &\sum_{j=0}^{l} \theta_j = 1, \quad \theta_j \ge 0, \quad j = 0,\cdots,l, \quad \mu = i_0. \end{aligned} \tag{8.2}$$

Assume $\{\lambda_j^k\}$ is the solution to (8.2). Analogous to the proof in [7], d^k defined by

$$d^k = -\sum_j \lambda_j^k \nabla f_{i_j} \tag{8.3}$$

is a descent direction unless $d^k = 0$, in which case we are optimal. Moreover, it is not difficult to prove that (8.2) can be solved via a least squares problem with only simple nonnegativity constraints [15].

9. Summary of the algorithm. Now we give a more detailed description of the algorithm.

Initialization: Suppose an initial point x^0 is given. Set $k \leftarrow 1$, $\epsilon \leftarrow \epsilon_0$, $\hat{\mathcal{W}}^0 \leftarrow \emptyset$.

Step 1. [QR decomposition]
Find the working set $\mathcal{W}^k \subseteq \hat{\mathcal{W}}^{k-1} \cup \mathcal{A}(x^k,\epsilon)$, Jacobian A^k, and its QR decomposition. Assume the columns of Z^k form a basis for the null space of A^{k^T}.
If $\mathcal{A}(x^k,\epsilon) \subseteq \mathcal{W}^k$ and $\|Z^{k^T}\nabla f_\mu\| \le \tau_c^k$, go to Step 2;
If $\mathcal{A}(x^k,\epsilon) \subseteq \mathcal{W}^k$ and $\|Z^{k^T}\nabla f_\mu\| > \tau_c^k$, go to Step 3;
Set $\epsilon \leftarrow 0.1\epsilon$;
If $\mathcal{A}(x^k,\epsilon) \not\subseteq \mathcal{W}^k$ and $\|Z^{k^T}\nabla f_\mu\| > \tau_c^k$, go to Step 4;
If $\mathcal{A}(x^k,\epsilon) \not\subseteq \mathcal{W}^k$ and $\|Z^{k^T}\nabla f_\mu\| \le \tau_c^k$, go to Step 5;
We note that the first and last instances imply that we have a cadre of type 1 ($\sum_{j=0}^{l}\lambda_j = 1$), and the third implies that we have a cadre of type 0 ($\sum_{j=0}^{l}\lambda_j = 0$).

Step 2. [Cadre "found" with $\sum_{i\in\mathcal{C}}\lambda_i = 1$]
If $\mathcal{W}^k$ is a reference set, obtain $B^k = Z^{k^T} G^k Z^k$, where G^k is defined as in (7.1); Compute the horizontal direction h^k and the vertical

direction v^k from (6.3); Set the search direction $d^k = h^k + v^k$ and $\hat{\mathcal{W}}^k \leftarrow \mathcal{W}^k$.

Otherwise, compute the vertical direction according to (6.6) and set $\hat{\mathcal{W}}^k$ using (6.7). Modify τ_c^k by (8.1) if degeneracy is encountered. Set $d^k = v^k$. Go to Step 6.

Step 3. [Cadre not found]

Obtain B^k as an approximation to $Z^{k^T} G^k Z^k$, where G^k is defined as in (7.1). Compute the horizontal direction h^k and the vertical direction v^k from (6.3). Compute the search direction d^k using (6.4). Set up $\hat{\mathcal{W}}^k$ according to (6.5). Go to Step 6.

Step 4. [Cadre "found" with $\sum_{i \in \mathcal{C}} \lambda_i = 0$]

Compute $d^k = -Z^k Z^{k^T} \nabla f_\mu^k$. $\hat{\mathcal{W}}^k \leftarrow \mathcal{W}^k$. Modify τ_c^k by (8.1) if degeneracy is encountered. Go to Step 6.

Step 5. [More than one cadre and at least one with $\sum_{i \in \mathcal{C}} \lambda_i = 1$]

Compute the search direction d^k using (8.3). Obtain $\hat{\mathcal{W}}^k$ from (6.7). Modify τ_c^k by (8.1).

Step 6. [Line search]

Perform a safeguarded line search. Set $k \leftarrow k + 1$. If $\|d^k\|_2 < \tau_s$ and $\mathcal{W}^k$ includes a levelled reference set, stop. Otherwise, go to Step 1. □

We use quotes around "found" to emphasize that τ_c^k is nonzero. The safeguards and details of the line search are given in [15].

10. Numerical testing. In this section, we compare the new algorithm with four other typical methods: [8], [13], [23], and [38].

The numerical results are for both minimax problems and discrete Chebyshev problems, all written in the form:

$$\min_{x \in \Re^n} \max_{i \in M} f_i(x). \tag{10.1}$$

The method of Conn. The method of [13] basically applies the active set strategy of nonlinear programming to the equivalent form of a minimax problem. It is a globally convergent algorithm with a superlinear convergence rate.

At each iteration, an equality-constrained quadratic programming subproblem is solved to determine the search direction. The subproblem is established upon all the current ϵ-active functions. The finite difference of the derivatives is used to approximate the second-order information.

This approach essentially corresponds to the sequential equality-constrained quadratic programming (EQP) approach for nonlinear programming problems, using projected Hessians. However, once the search direction is determined, the line search is done directly on the nondifferentiable maximum function $\psi(x)$.

Although there have been relatively fewer numerical results for general nonlinear minimax problems than for linear problems, to date, the available numerical results seem to indicate that the following method [23], which is a combination of a linear programming (LP) approach and a quasi-Newton method for a nonlinear system of equations, works well on most types of minimax problems.

The method of Hald and Madsen. At each iteration of the first stage, the method of [23] requires an exact solution to a constrained linear minimax problem

$$\begin{aligned}&\min_{d\in\Re^n}\max_{i\in[M]}\ \{f_i(x^k)+\nabla f_i(x^k)^T d\}\\ &\text{subject to}\\ &\|d\|_\infty \leq \Lambda^k\end{aligned}$$

in order to find the search direction. A trust region method has been incorporated to ensure convergence.

If a solution is suspected of going through a smooth valley, i.e., the number of active functions at the solution is less than $n+1$, a switch to a second stage is made. Then a nonlinear system of equations established by the Kuhn–Tucker conditions for the active functions is solved by some quasi-Newton method.

The entire Lagrangian Hessian is approximated by some modified secant updates. It is possible for the maximum $\psi(x)$ to be increased. A return to the first stage might be necessary.

Under certain conditions, the method of [23] is globally convergent with a quadratic or superlinear final convergence, depending upon whether a Newton or a quasi-Newton method is involved.

The first stage of the method essentially corresponds to a sequential linear programming approach (SLP), stabilized via a trust region, for nonlinear programming problems.

The method of Womersley and Fletcher. The method of [38] is similar to that of [13]. It is a descent method which uses an active set strategy, a nonsmooth line search, and a quasi-Newton approximation to the projected Hessian of the Lagrangian function.

Global convergence of the algorithm has been proved. Under certain conditions, superlinear convergence occurs.

Like that of [13], this method could be considered as belonging to the class of sequential equality-constrained quadratic programming (EQP) approaches.

The method of Charalambous. In the approach of [8], the original minimax problem is defined as a modified least pth objective function which under certain conditions have the same optimum as the original problem.

10.1. Computational costs comparison. At each iteration, the methods of [13] and [38] and the new algorithm require the computation of a search direction obtained by solving an equality-constrained quadratic programming (EQP) or an equality-constrained linear programming (ELP). Comparable line searches have been used in the methods of Conn and Womersley and Charalambous and Fletcher, whereas Hald and Madsen used the trust region method. For our new algorithm, determining a cadre and dropping one function in the working set, when a nonreference set cadre is found, requires no extra work compared with the methods of [13] and [38]. When more than one function in the working set is dropped, an equivalent number of QR updates are required. Since these functions should be dropped and function evaluation typically is more expensive than a single QR update, in general, this extra work is well justified. The amount of computation per iteration required by the above three methods (i.e., [13] and [38] and the new method) is roughly the same.

The amount of work required by each iteration of [8] is roughly the same as performing a quasi-Newton step for an unconstrained function.

At each iteration of [23], in stage one, a linear programming problem of size at least $n \times |M|$ is solved up to optimality. At each iteration of stage two, if it is ever entered, the computation required is similar to the methods of [13] and [38]. However, in general, most of the iterations are spent in stage one.

Loosely speaking, comparison of computational costs of one iteration of the new algorithm and that of [23] is similar to the comparison between one iteration of EQP and IQP methods.

A solution of EQP can be obtained by solving two linear systems of equations. The size of each linear system is at most n. A solution for IQP, however, usually requires iterative methods (i.e., inner iterations). Although the number of iterations are bounded by the number of unknowns and constraints, it is potentially very large and it could even become prohibitive for a discretised Chebyshev problem because the number of the constraints of its associated IQP can be much larger than those of the usual nonlinear programming problems.

Therefore, considering the amount of work required per iteration, the method of [23] is considerably more expensive than the others.

For nonlinear programming problems, the advantage of the IQP approach compared to EQP, however, has been the iterative search for the correct active set. Likewise, one would expect that the advantage of the method of [23] over that of [13] and [38] and the new algorithm is similar to that of the successive IQP method over the successive EQP approach for nonlinear programming problems; namely, it can identify the correct active set faster. This advantage probably is the case for the methods of [13] and [38]. The new algorithm, however, is not a pure active set method. It can also identify the correct active set quickly. It achieves this not by an iterative search but by recognising the structure of the optimum and constructively building up the reference set. Through exploiting the structure of the Chebyshev problem and minimax problem, we are able to retain the advantages of both the EQP approach and the IQP approach.

Finally, we remark that for a degenerate point of Type A or B, there is no extra work required compared with that for a nondegenerate point. For a degenerate point of Type C, we must solve a least squares problem with nonnegativity constraints.

10.2. Numerical results. We present some limited numerical results in this section.

For our numerical testing, the initial parameters required by the algorithm are set as

$$\tau_c^0 = 0.05, \quad \tau_0 = 10^{-12}, \quad \tau_s = \tfrac{1}{2}10^{-5}, \quad \rho = 0.5, \quad \epsilon_0 = 0.1.$$

The algorithm terminates when the following three conditions are satisfied:

1. $\|d^k\|_2 \le \tau_s$;
2. $\mathcal{W}^k \subseteq \mathcal{A}(x^k, \epsilon)$;
3. $\lambda_j^k \ge 0$, for all $j \in \mathcal{W}^k$.

Thus, at termination, there exists, approximately, a levelled reference set with the maximum deviation.

The test problems include both nonlinear minimax problems and nonlinear Chebyshev problems.

We implicitly write a nonlinear Chebyshev problem

$$\min_{1 \le i \le m} |f_i(x)|$$

in the general minimax form

$$\min_{1 \le i \le 2m} f_i(x),$$

where $f_{i+m}(x) = -f_i(x)$, for $i = 1, \cdots, m$.

Consider the following nonlinear programming problem:

$$\begin{aligned} &\min_{x \in \Re^n} F(x) \\ &\text{subject to} \\ &g_i(x) \geq 0, \qquad i = 2, \cdots, m, \end{aligned}$$

and the minimax problem:

$$\begin{aligned} &\min_{x \in \Re^n} \max_{1 \leq i \leq m} f_i(x) \\ &\text{subject to} \\ &f_1(x) = F(x), \\ &f_i(x) = F(x) - \alpha_i g_i(x), \qquad 2 \leq i \leq m, \end{aligned}$$

where

$$\alpha_i \geq 0, \qquad 2 \leq i \leq m.$$

It is straightforward to show that for sufficiently large α_i, the optimum of the minimax problem coincides with that of the nonlinear programming problem (see [2]).

We have tested some nonlinear programming problems through the above transformation. The α parameter is set as

$$\alpha_i = 10.0, \qquad 2 \leq i \leq m,$$

which we know, a priori, is sufficiently large.

We have listed the results for the following minimax testing problems (their references are also indicated): Charalambous and Bandler 1 and Charalambous and Bandler 2 [9], Freudenstein and Roth [36], Colville problem 2 [12], Barrodale, Powell, and Roberts [5], Wong 1, Wong 2, and Wong 3 [10], Rosen and Suzuki [33], Rosenbrock [34], Transmission Problems [3], Davidon [16], Enzyme [25], El Attar [19], Hettich [36], Bard [36], Watson [36], and Osborne [36]. The starting points used are the same as that specified in the references.

The results for the problems Davidon, Enzyme, El Attar, and Hettich, under the column [23] are taken from [28], which describes essentially the same method as that of Hald and Madsen.

In Table 10.1, we report the number of function evaluations required by our new algorithm under the column NM. For each problem, we have used the nomenclature of the cited reference. The results of other methods, using a comparable stopping tolerance, are listed for comparison where available.

The column under the column *nact* indicates the number of maximum functions at the solution.

The Rosenbrock problem is degenerate at the solution. The Watson problem is degenerate at the starting point. The Watson problem with $n = 20$ is also degenerate at the solution obtained. For the other test problems, numerical degeneracy does not occur.

The reported results use extended BFGS updates. Similar results were obtained using exact derivatives. From the limited numerical results, we observe that, compared with [8], [13], and [38], the overall number of function evaluations required by the new

TABLE 10.1
Number of function evaluations: BFGS *updates.*

Problems	n	m	nact	NM	HM	CN	WF	CL
					[23]	[13]	[38]	[8]
Charalambous & Bandler 1	2	3	2	11	11[a]	18	12	
Charalambous & Bandler 2	2	3	3	6	11[a]	8	6	
Freudenstein & Roth	2	2	2	11	15[a]			
Colville 2	15	21	12	49	41[a]	275	80	413
Barrodale, Powell et al.	5	21	5	21	10		38	
Wong1.1	7	5	3	25	23	106	53	107
Wong1.2	7	5	3	33	29	77	37	
Wong2	10	9	7	24	27[a]			120
Wong3	20	18	13	33	49[a]			318
Rosen & Suzuki	4	4	3	12	18	64	37	66
Rosenbrock	2	2	4	31	21			
Transmission 1	6	11	4	52	21	67		78
Transmission 2	6	11	4	25	46	80		
Davidon	4	20	3	20	27			
Enzyme	4	22	5	11	18			
El Attar	6	51	7	25	12			
Hettich	4	5	4	11	19[b]			
Bard	3	15	3	10	9[a]			
Madsen	2	3	2	17	13[a]			
Watson6	6	31	7	24	12[a]			
Watson20	20	31	39	22	39[a]			
Osborne	5	33	5	10	31[a]			

[a] The results are obtained by using the codes in [23].

[b] The algorithm stopped because of roundoff error without obtaining a solution.

algorithm is much less. We also recall that the amount of computation per iteration required by all but [8] to determine the search direction and stepsize are comparable. If one considers in more detail the number of function evaluations required and the size and complexity of the matrices being updated it would appear that the new method is more efficient than [8]. Hence, the new method appears to be more efficient than that of [8], [13], and [38].

The only method that seems to be competitive with the new algorithm is that of [23]. The number of function evaluations required by these two methods is comparable. However, we recall that the amount of remaining computation required per iteration demanded by the method of [23] is significantly more than the proposed method. Thus our new method still appears to be preferable.

We have also tested our new algorithm on a real application problem. The problem has 80 functions, in terms of a general minimax problem, with 40 variables. The number of activities at the solution is 39 (out of 80). Our algorithm solved it successfully in 50 function evaluations while the method of [23] failed to locate a solution.

11. Summary. The algorithm presented is a globally convergent algorithm with superlinear convergence rate [26]. It has been developed based on the principle that a minimax problem, in particular the Chebyshev problem, has special properties that can be computationally exploited in both the linear and nonlinear cases.

In this paper, we generalise the characterisation for a best linear Chebyshev approximation to nonlinear minimax problems. These generalisations are implementable computationally. We then present an algorithm which profits from this exploitation.

Typically, the algorithm attempts to find a cadre by focusing on the functions that have achieved maximum values through iterations, i.e., functions in working sets. These functions are then levelled by vertical directions whenever possible. If a reference set has been located, it is then levelled by vertical directions (which are descent directions) and thus a solution is quickly determined. If, however, the cadre does not correspond to a reference set, a descent direction is then defined as an attempt to construct one. Since, at any solution, there exists a *levelled* reference set with the *maximum* value, it is clear that the computational procedure is meaningful and we believe our numerical results indicate its promise.

We point out that it is possible for the Maratos effect to occur for the new algorithm as presently implemented. However, we have not experienced this effect during our numerical testing. Moreover, the algorithm can be slightly modified to guarantee that there is no Maratos effect. One only needs to reevaluate the functions at the point $x^k + h^k$ and compute the vertical direction using the updated values when one is close to a stationary point (see [15] for more details).

Finally, we point out that the algorithm can be extended to solve the constrained minimax problem (see [15] for more details).

REFERENCES

[1] D. H. Anderson and M. R. Osborne, *Discrete, nonlinear approximations in polyhedral norms: A Levenberg-like algorithm*, Numer. Math., 28 (1977), pp. 167–170.

[2] J. W. Bandler and C. Charalambous, *Nonlinear programming using minimax techniques*, J. Optim. Theory Appl., 13 (1974), pp. 607–619.

[3] J. W. Bandler and P. A. McDobald, *Optimization of microwave networks by razor research*, IEEE Trans. Microwave Theory Tech., 17 (1969), pp. 552–562.

[4] I. Barrodale and C. Phillips, *An improved algorithm for discrete Chebychev linear approximation*, in Proc. 4th Manitoba Conference on Numer. Math., University of Manitoba, Winnipeg, Canada, 1974, pp. 177–190.

[5] I. Barrodale, M. J. D. Powell, and F. D. K. Roberts, *The differential correction algorithm for rational l_∞-approximation*, SIAM J. Numer. Anal., 7 (1972), pp. 493–504.

[6] R. H. Bartels, A. R. Conn, and Y. Li, *Primal methods are better than dual methods for solving overdetermined linear systems in the l_∞ sense?*, SIAM J. Numer. Anal., 26 (1989), pp. 693–726.

[7] S. Busovača, *Handling degeneracy in a nonlinear l_1 algorithm*, Tech. Report CS-85-34, University of Waterloo, Department of Computer Science, University of Waterloo, Waterloo, Ontario, Canada, 1985.

[8] C. Charalambous, *Acceleration of the least pth algorithm for minimax optimization*, Math. Programming, 17 (1979), pp. 270–297.

[9] C. Charalambous and J. W. Bandler, *Nonlinear minimax optimisation as a sequence of least pth optimization with finite values of p*, Internat. J. System Sci., 7 (1976), pp. 377–394.

[10] C. Charalambous and A. R. Conn, *An efficient method to solve the minimax problem directly*, SIAM J. Numer. Anal., 15 (1978), pp. 162–187.

[11] T. F. Coleman and A. R. Conn, *Nonlinear programming via an exact penalty function: Asymptotic analysis*, Math. Programming, 24 (1982), pp. 123–136.

[12] A. R. Colville, *A comparative study on nonlinear programming codes*, Tech. Report 320-2949, IBM Thomas J. Watson Research Center, Yorktown Heights, NY, 1968.

[13] A. R. Conn, *An efficient second order method to solve the (constrained) minimax problem*, Tech. Report CORR 79-5, Department of Combinatorics and Optimization, University of Waterloo, Waterloo, Ontario, Canada, 1979.

[14] A. R. Conn and Y. Li, *Structure and characterization of discrete Chebyshev problems*, Tech. Report CS-88-39, Department of Computer Science, University of Waterloo, Waterloo, Ontario, Canada, 1988.

[15] ———, *An approach to nonlinear l_∞ approximation*, in Proc. Fifth Mexican Workshop on Numerical Analysis, J. P. Hennart, ed., Society for Industrial and Applied Mathematics, Philadelphia, PA, 1991, pp. 346–365.

[16] W. C. DAVIDON, *A new least squares algorithm*, J. Optim. Theory Appl., 18 (1976), pp. 187–198.

[17] A. DAX, *A note on optimality conditions for the Euclidean multifacility location problem*, Math. Programming, 36 (1986), pp. 72–80.

[18] J. DESCLOUX, *Dégéneresence dans les approximations de Tschebysheff linéaries et discrètes*, Numer. Math., 3 (1961), pp. 180–187.

[19] EL-ATTAR, M. VIDYASAGAR, AND S. R. K. DUTTA, *An algorithm for l_1-approximation*, SIAM J. Numer. Anal., 16 (1979), pp. 70–86.

[20] R. FLETCHER, *A model algorithm for composite nondifferentiable optimization problems*, Math. Programming Stud., 17 (1982), pp. 67–76.

[21] G. H. GOLUB AND C. F. VAN LOAN, *Matrix Computations*, The Johns Hopkins University Press, Baltimore, MD, 1983.

[22] N. I. M. GOULD, *On solving three classes of nonlinear programming problems via simple differentiable penalty functions*, J. Optim. Theory Appl., 56 (1988), pp. 89–126.

[23] J. HALD AND K. MADSEN, *Combined* LP *and Quasi-Newton methods for minimax optimization*, Math. Programming, 20 (1981), pp. 49–62.

[24] S. P. HAN, *Variable metric methods for minimizing a class of nondifferentiable functions*, Math. Programming, 20 (1981), pp. 1–13.

[25] J. S. KOWALIK AND M. R. OSBORNE, *Methods for Unconstrained Optimization*, American Elsevier, New York, 1968.

[26] Y. LI, *An efficient algorithm for nonlinear minimax problems*, Ph.D. thesis, Computer Science Department, University of Waterloo, Waterloo, Ontario, Canada, 1988.

[27] K. MADSEN, *Minimax solution of nonlinear equations without calculating derivatives*, Math. Programming Stud., 3 (1975), pp. 110–126.

[28] ———, *Minimization of nonlinear approximation functions*, Ph.D. thesis, Technical University of Denmark, Lyngby, Denmark, 1985.

[29] W. MURRAY AND M. L. OVERTON, *A projected Lagrangian algorithm for nonlinear minimax optimization*, SIAM. J. Sci. Statist. Comput., 1 (1980), pp. 345–370.

[30] M. R. OSBORNE AND G. A. WATSON, *An algorithm for minimax approximation in the nonlinear case*, Comput. J., 12 (1968), pp. 63–68.

[31] M. J. D. POWELL, *Approximation Theory and Methods*, Cambridge University Press, Cambridge, U.K., 1981.

[32] J. R. RICE, *The Approximation of Functions* 2, *Nonlinear Theory*, Addison–Wesley, Reading, MA, 1969.

[33] J. B. ROSEN AND S. SUZUKI, *Construction of the nonlinear programming test problems*, Comm. ACM, 8 (1965), p. 113.

[34] H. H. ROSENBROCK, *An automatic method for finding the greatest or least value of a function*, Comput. J., 3 (1960), pp. 175–184.

[35] E. L. STIEFEL, *Numerical methods of Tchebycheff approximation*, in Proceedings of a Symposium, Mathematics Research Center, University of Wisconsin Press, Madison, WI, 1959.

[36] G. A. WATSON, *The minimax solution of an overdetermined system of non-linear equations*, J. Inst. Math. Appl., 23 (1979), pp. 167–180.

[37] R. S. WOMERSLEY, *Local properties of algorithms for minimizing nonsmooth composite functions*, Math. Programming, 32 (1985), pp. 69–89.

[38] R. S. WOMERSLEY AND R. FLETCHER, *An algorithm for composite nonsmooth optimization problems*, J. Optim. Theory Appl., 48 (1986), pp. 493–523.

SIAM J. OPTIMIZATION
Vol. 2, No. 2, pp. 264–303, May 1992

OPTIMAL DISTRIBUTION OF LARVICIDE IN RUNNING WATERS*

ALAIN CHALIFOUR[†] AND MICHEL C. DELFOUR[‡]

Abstract. In some regions of the world black flies are the vector of serious endemic diseases such as onchocerciasis or river blindness. The object of this paper is to present theoretical and numerical solutions to the control of black fly larvae in running waters. The problem is modelled by a diffusion-transport partial differential equation with impulse controls and state constraint. We present the solutions for the one-dimensional case.

Key words. health problems, optimization, river, larvicide, control, onchocerciasis

AMS(MOS) subject classifications. 49A22, 49D37

1. Introduction.

Black flies (*Simulium damnosum*) are known not only for being a nuisance in causing economic losses in different areas of human activities, but also for transmission of pathogens and parasites to man and animals. In some areas, black flies are vectors of a filarial worm (*Onchocerca volvulus*) that causes a serious endemic disease whose final stage is known as river blindness. "Onchocerciasis, or river blindness, is one of the major endemic, parasitic diseases which in addition to causing untold human suffering is a major obstacle to socioeconomic development. It is found in the Americas, in the southwestern part of the Arabian peninsula, and in East, Central, and West Africa. It is estimated that between 20 and 30 million people are infected by onchocerciasis throughout the world" (see OCP [13]).

The strategy chosen was to break the chain of transmission by destroying the vector at its most vulnerable state, that is, the larval state. To control black fly larvae in running waters, special products have been developed with targeted toxic effects. Helicopters are used to periodically spray the rivers at prescribed sites over very large geographical areas. To reduce the costs of operations, it is important to determine the amounts of product and locations of the injection sites to minimize the total quantity of sprayed larvicide while maintaining a given level of mortality along the river to be treated.

Black fly larvae are found at specific breeding sites in rivers. Since we have to control the mortality level over long distances, we have to take into account the transport, the diffusion, and the decay of the larvicide. The behaviour of the concentration of larvicide along the river can be modelled by a diffusion-transport partial differential equation. Biologists have established that for the types of larvicides used the rate of mortality is proportional to the "dose": the time integral of the concentration up to

*Received by the editors November 5, 1990; accepted for publication (in revised form) June 24, 1991.

[†] Département de Mathématiques et d'Informatique, Université du Québec à Trois-Rivières, C.P. 500, Trois-Rivières, Québec, Canada, G9A 5H7.

[‡] Centre de Recherches Mathématiques et Département de Mathématiques et de Statistique, Université de Montréal, C.P. 6128 A, Montréal, Québec, Canada, H3C 3J7. The research of this author was supported in part by a Killam fellowship from Canada Sciences Council, National Sciences and Engineering Research Council of Canada grants OGP-8730 and INF-7939, and a "Fonds pour la Formation de Chercheurs et l'Aide à la Recherche (FCAR)" grant from the "Ministère de l'Education du Québec."

infinity. In this problem the state variable is the spacial distribution of the dose, which is the solution of a partial differential equation over the river. The injections of insecticide appear as impulse controls at points along the river (one-dimensional model) or along lines corresponding to the paths followed by the helicopter (two-dimensional model). The one-dimensional model is used to treat a segment of river while the two-dimensional model is used for a complex site.

The control problem consists of finding the best locations to inject the larvicide while maintaining a minimum dose and minimizing the total amount of larvicide sprayed in the river. So it is an impulse control problem for partial differential equations with state constraint. On top of this there is an interesting logistic problem, which consists of scheduling helicopters and managing fuel and larvicide caches over a large network of rivers.

The object of this paper is to present a modelization of the problem and a numerical solution of the optimal control problem in dimension one. A mixed discontinuous finite element formulation (Lesaint and Raviart [11] and Raviart [14]) has been used to solve the diffusion-transport equations. A special technique has also been developed to get around the combinatorial problem that naturally arises from the discretization of the impulse control problem. Numerical results are presented using real data.

The modelization of this type of problem is related to the problem of the localization of industrial sites or plants along rivers where the objective is to minimize ecological damages to the environment. For more details the reader is referred to the book and papers of G. I. Marchuk [12] and his team.

2. One-dimensional equations and susceptibility model.

2.1. Equations for the concentration. First consider a one-dimensional river. Denote by $c(x,t)$ (kg/m^3) the concentration of larvicide at time $t > 0$ (second) and at point x (meter) downstream of the origin 0. It is the solution of the diffusion-transport equation (Taylor [15], Aris [1], Khalig [10], Marchuk [12])

$$\frac{\partial c}{\partial t}+V(x)\frac{\partial c}{\partial x}-\frac{\partial}{\partial x}\left(E(x)\frac{\partial c}{\partial x}\right)+R(x)c = \sum_{i=1}^{N}\frac{m_i}{A(x_i)}\delta_i(t)\delta_i(x), \quad x \in \mathbb{R}, \quad t > 0, \tag{2.1}$$

where $V(x)$ is the "mean velocity" of the water in m/ sec; $A(x) > 0$ is the cross-sectional area of the river in m^2; $E(x) \geq 0$ is the dispersion coefficient in m^2/ sec; $R(x) \geq 0$ is the loss coefficient in sec^{-1}; $N > 0$ is the number of sites to be visited; $x_1, x_2, \cdots, x_N$ are the locations of the sites such that

$$x_0 \overset{\text{def}}{=} 0 \leq x_1 \leq x_2 \leq \cdots \leq x_N \leq x_{N+1} \overset{\text{def}}{=} L \tag{2.2}$$

for some finite $L > 0$; $t_1, t_2, \cdots, t_N$ are the times at which the sites are visited:

$$t_0 \overset{\text{def}}{=} 0 \leq t_1 \leq t_2 \leq \cdots \leq t_N \leq t_{N+1} \overset{\text{def}}{=} T \tag{2.3}$$

for some finite $T > 0$; and $\delta_i(t)$ and $\delta_i(x)$ are the Dirac delta functions at $t = t_i$ and $x = x_i$, respectively. In each site x_i the helicopter sprays m_i kg of product. We assume that the mixing of the product is instantaneous and produces a uniform concentration in the vertical section $A(x_i)$ of the river at x_i. The initial condition is

$$c(x,0) = c^0(x), \quad x \in \mathbb{R}. \tag{2.4}$$

2.2. Susceptibility model. The biological aspects of this work rest on experimental data and observations. They are used to construct and validate what the biologist calls a "susceptibility model," which provides the relationship between the mortality rate P (a percentage) of larvae and concentrations of larvicide over periods of time. Laboratory and field experiments have established a direct relationship between the mortality rate of larvae and the dose $u(x)$, which is the time-integral of the concentration of larvicide over $t > 0$:

$$u(x) = \int_0^\infty c(x,t)\,dt \tag{2.5}$$

for the biological larvicide B.t.i. (see Guillet, Escaffre, Prud'hom, and Bakayoko [8] and Guillet, Hougard, Doannio, and Escaffre [9]). So to obtain a given mortality rate of P percent, it is sufficient to specify a minimum level u_P for the dose and to require that at each point of the river

$$u(x) \geq u_P, \qquad x \in [0,L]. \tag{2.6}$$

The total mass $M[a,b]$ of larvicide going through the segment $[a,b]$ is given by

$$M[a,b] = \int_a^b u(x)A(x)V(x)\,dx = \int_a^b u(x)Q(x)\,dx \tag{2.7}$$

where

$$Q(x) = V(x)A(x) \tag{2.8}$$

is the flow in $\mathrm{m}^3/\sec$ at the point x.

2.3. Equations for the dose. Assuming that the concentration $c(x,t)$ in each point x goes to 0 as t goes to ∞, we obtain the following equation for the dose $u(x)$:

$$-\frac{d}{dx}\left(E(x)\frac{du}{dx}\right) + V(x)\frac{du}{dx} + R(x)u = c^0(x) + \sum_{i=1}^{N} \frac{m_i}{A(x_i)}\delta_i(x), \qquad x \in \mathbb{R}. \tag{2.9}$$

The dose is a cumulating variable over time. Therefore, the injection times, t_i's, completely disappear in the above equation. This considerably simplifies the formulation of the initial problem and yields a new time-independent problem on $\mathbb{R}$. In practice, for a fixed segment of river, all injections take place within a fixed time framework $[0,T]$ of at most 24 hours. Yet the amount of time it takes to build up the dose is a function of the various parameters entering into the equations. When the flow is too slow in a specific area, this area receives special attention. Such sites usually become "compulsory sites." In extreme cases complementary surface treatment is required. All this is taken into account when partitioning the river into sections.

2.3.1. The problem without diffusion. When there is no diffusion (that is, $E = 0$), we assume that the boundary condition

$$c(0,t) = c_0(t), \qquad t > 0, \tag{2.10}$$

is given at $x = 0$. Equation (2.9) reduces to

$$V(x)\frac{du}{dx} + R(x)u = c^0(x) + \sum_{i=1}^{N} \frac{m_i}{A(x_i)}\delta_i(x), \qquad x \in \mathbb{R}, \tag{2.11}$$

with the boundary condition

$$u(0) = u_0 \overset{\text{def}}{=} \int_0^\infty c_0(t)\, dt. \tag{2.12}$$

For $R \in L^2(0,L)$, $V \in H^1(0,L)$, and $V^{-1} \in L^\infty(0,L)$, then, $V^{-1} \in H^1(0,L)$ and system (2.11)–(2.12) has a unique solution given by

$$\begin{aligned} u(x) = {} & u_0\, e^{-\int_0^x R(y)/V(y)\, dy} + \int_0^x e^{-\int_z^x R(y)/V(y)\, dy}\, \frac{c^0(z)}{V(z)}\, dz \\ & + \sum_{\substack{i=1 \\ x \geq x_i}} e^{-\int_{x_i}^x R(y)/V(y)\, dy}\, \frac{m_i}{A(x_i)V(x_i)}. \end{aligned} \tag{2.13}$$

It will also be convenient to give a variational formulation of system (2.11)–(2.12) on a fixed segment $[0,L]$, $L > 0$ finite, and consider more general right-hand sides. For this purpose we introduce the following continuous bilinear form:

$$\begin{aligned} & b_0 : (L^2(0,L) \times \mathbb{R}) \times H^1(0,L) \to \mathbb{R}, \\ & b_0((u,u_L),v) = \int_0^L u\left[-\frac{d}{dx}(Vv) + Rv\right] dx + u_L V(L)v(L). \end{aligned} \tag{2.14}$$

THEOREM 2.1. *Assume that $R \in L^2(0,L)$, $V \in H^1(0,L)$, and $V^{-1} \in L^\infty(0,L)$. Then for any ℓ in $H^1(0,L)'$, the variational equation*

$$\exists (u,u_L) \in L^2(0,L) \times \mathbb{R}, \quad \forall v \in H^1(0,L), \quad b_0((u,u_L),v) = \langle \ell, v\rangle_{H^1} \tag{2.15}$$

has a unique solution. When ℓ is of the form

$$\langle \ell, v\rangle = \int_0^L c^0(x)v(x)\, dx + \sum_{i=1}^{N} \frac{m_i}{A(x_i)}v(x_i) + u_0 V(0)v(0), \tag{2.16}$$

the function u in (2.15) coincides with the solution of (2.11)–(2.12) almost everywhere and u_L is the right-hand side value $u(L^+)$ of the piecewise smooth function u.

Proof. See the appendix for the proof. □

2.3.2. The problem with diffusion. When $E(x) > 0$, (2.9) has a unique solution in the Sobolev space $H^1(\mathbb{R})$ under the following conditions.

THEOREM 2.2. *Assume that c^0 in $L^2(\mathbb{R})$, E and R in $L^\infty(\mathbb{R})$, and V in $W^{1,\infty}(\mathbb{R})$ verify the conditions*

$$\exists \alpha > 0, \quad \forall x \in \mathbb{R}, \quad R(x) - \frac{1}{2}\frac{dV}{dx}(x) \geq \alpha, \quad E(x) \geq \alpha. \tag{2.17}$$

The bilinear form

$$a(u,v) = \int_{\mathbb{R}} E(x)\frac{du}{dx}\frac{dv}{dx} + V(x)\frac{du}{dx}v + R(x)uv\, dx \tag{2.18}$$

is continuous and coercive on $H^1(\mathbb{R})$ and finding a solution to (2.9) *is equivalent to the following variational problem: to find u in $H^1(\mathbb{R})$ such that for all v in $H^1(\mathbb{R})$,*

$$a(u,v) = \int_{\mathbb{R}} c^0(x)v(x)\,dx + \sum_{i=1}^{N} \frac{m_i}{A(x_i)}v(x_i). \tag{2.19}$$

The variational equation (2.19) *has a unique solution in $H^1(\mathbb{R})$, which coincides with the solution of* (2.9).

Proof. For the proof, see the Appendix. □

This first existence result gives the dose for an infinite river. However, for operational and numerical reasons, the user is generally interested in simulating only a finite segment $[0, L]$, $L > 0$, of the river. To do that, we have to specify in $x = 0$ and $x = L$ "transparent conditions" that do not perturb the physics of the problem too much. Thus, in addition to the hypotheses of Theorem 2.2, we now assume that E and R are continuous and that

$$\begin{aligned} &E(x) = E(0), \quad V(x) = V(0), \quad R(x) = R(0) \geq 0, \quad \forall x \leq 0,\\ &E(x) = E(L), \quad V(x) = V(L), \quad R(x) = R(L) \geq 0, \quad \forall x \geq L. \end{aligned} \tag{2.20}$$

So, upstream of $x = 0$ and downstream of $x = L$ the river is uniform with constant parameters E, V, and R. Then on the part $]-\infty, 0]$ (respectively, $[L, \infty[$), we have the asymptotic condition $u(-\infty) = 0$ (respectively, $u(+\infty) = 0$) and it is easy to verify that at $x = 0$ (respectively, $x = L$) the following identity holds:

$$\begin{aligned} -E(0)\frac{du}{dx}(0) + \beta(0)u(0) \;&=\; C_0\\ \left(\text{respectively, } E(L)\frac{du}{dx}(L) + \beta(L)u(L) \;\right.&=\left.\; C_L\right) \end{aligned} \tag{2.21}$$

where

$$\begin{aligned} C_0 &= \int_{-\infty}^{0} e^{\frac{1}{2E(0)}\left[\sqrt{V(0)^2+4E(0)R(0)}-V(0)\right]y}c^0(y)\,dy,\\ C_L &= \int_{L}^{\infty} e^{-\frac{1}{2E(L)}\left[\sqrt{V(L)^2+4E(L)R(L)}+V(L)\right](y-L)}c^0(y)\,dy \end{aligned} \tag{2.22}$$

and

$$\begin{aligned} \beta(0) &= \tfrac{1}{2}\left[\sqrt{V(0)^2 + 4E(0)R(0)} + V(0)\right],\\ \beta(L) &= \tfrac{1}{2}\left[\sqrt{V(L)^2 + 4E(L)R(L)} - V(L)\right]. \end{aligned} \tag{2.23}$$

The idea is to consider (2.9) on $]0, L[$ with the boundary conditions (2.21). The following theorem now gives the connection between the solution of (2.9) on $\mathbb{R}$ and (2.9) on $]0, L[$ with the boundary conditions (2.21).

THEOREM 2.3. *Assume that, in addition to the hypotheses of Theorem* 2.2, *E and R are continuous and verify assumptions* (2.20). *Then the boundary value problem*
(2.24)

$$\begin{aligned} &-\frac{d}{dx}\left(E(x)\frac{du}{dx}\right) + V(x)\frac{du}{dx} + R(x)u = c^0(x) + \sum_{i=1}^{N} \frac{m_i}{A(x_i)}\delta_i(x), \qquad x \in [0, L],\\ &-E(0)\frac{du}{dx}(0) + \beta(0)u(0) \;=\; C_0, \qquad E(L)\frac{du}{dx}(L) + \beta(L)u(L) \;=\; C_L \end{aligned}$$

is equivalent to the following variational problem: to find u in $H^1(0,L)$ such that for all v in $H^1(0,L)$

$$b(u,v) = \int_0^L c^0(x)v(x)\,dx + \sum_{i=1}^{N} \frac{m_i}{A(x_i)}v(x_i) + C_0 v(0) + C_L v(L), \tag{2.25}$$

where b is the coercive continuous bilinear form

$$\begin{aligned} b(u,v) = \int_0^L & E(x)\frac{du}{dx}\frac{dv}{dx} + V(x)\frac{du}{dx}v + R(x)uv\,dx \\ &+ \beta(L)u(L)v(L) + \beta(0)u(0)v(0). \end{aligned} \tag{2.26}$$

Problem (2.25) *has a unique solution in $H^1(0,L)$, which coincides with the restriction to* $[0,L]$ *of the solution of* (2.9) *on* $\mathbb{R}$.

Proof. For the proof, see the Appendix. □

Remark 2.1. In system (2.1)–(2.2) we have assumed that all the injection points x_i belong to $]0,L[$ and avoided the points $x = 0$ and $x = L$. However, this is not a limitation in the variational formulations (2.18) and (2.25). Such injections will not appear in (2.9), but rather in the boundary conditions (2.21):

$$\begin{aligned} -E(0)\frac{du}{dx}(0) + \beta(0)u(0) &= \frac{m_0}{A(0)} + C_0, \\ E(L)\frac{du}{dx}(L) + \beta(L)u(L) &= \frac{m_L}{A(L)} + C_L. \end{aligned} \tag{2.27}$$ □

3. Optimal control of the one-dimensional model. In this section we assume that the initial condition $c^0(x)$ is zero. So $C_0 = C_L = 0$ everywhere in §2.

3.1. Problem formulation. Consider a segment $[0,L]$, $L > 0$, of river to be treated in $N > 0$ ordered sites x_i's and assume that the dose verifies (2.9) with the appropriate boundary conditions: (2.12) without diffusion and (2.21) with diffusion.

The following three problems will be discussed.

PROBLEM 1. Given the number of sites N and the positions $\vec{x}_N = \{x_i : 1 \le i \le N\}$, find the masses $\vec{m}_N = \{m_i : m_i \ge 0,\ 1 \le i \le N\}$ that minimize the total mass of sprayed product

$$M(N, \vec{x}_N, \vec{m}_N) = \sum_{i=1}^{N} m_i \tag{3.1}$$

under the state constraint

$$u(x) \ge u_P, \qquad 0 \le x \le L, \tag{3.2}$$

where u is the solution of the state equation (2.9) with its appropriate boundary conditions. □

PROBLEM 2. Given the number of sites N, find the positions $\vec{x}_N = \{x_i : 1 \le i \le N\}$ and the masses $\vec{m}_N = \{m_i : m_i \ge 0,\ 1 \le i \le N\}$ that minimize the total mass of sprayed product (3.1) under the state constraint (3.2), where u is the solution of the state equation (2.9) with its appropriate boundary conditions. □

PROBLEM 3. Find the number of sites N, the positions $\vec{x}_N = \{x_i : 1 \leq i \leq N\}$, and the masses $\vec{m}_N = \{m_i : m_i \geq 0,\ 1 \leq i \leq N\}$ that minimize the total mass of sprayed product (3.1) under the state constraint (3.2), where u is the solution of the state equation (2.9) with its appropriate boundary conditions. □

If there is no upper bound on the number of sites to be treated, the optimal solution is not given by a finite vector of pairs (x_i, m_i) but a measure on $]0, L[$. There the masses m_i go to zero as the number N goes to infinity and it can also be verified that in each point

$$u(x) = u_P, \qquad 0 \leq x \leq L. \tag{3.3}$$

Of course, Problem 3 is an asymptotic version of the problem. In practice, the helicopter can only treat a finite number of injection sites and there is an upper bound on the number of sites and a lower bound on the amount of product that is sprayed at each site. Several other constraints are present: compulsory site, upper bound on the dose to minimize damages to the environment, etc. However, Problem 3 will provide a lower bound on the total amount of product necessary to treat a given segment of river.

3.2. Solution of the optimization problems.

In this section we introduce sets of hypotheses under which the optimization problems (Problems 1–3) have a solution. In some cases we even give the exact form of the solution, which will help in testing the numerical algorithm.

3.2.1. The case without diffusion. When $E = 0$, we have seen in §2.3.1 that the problem has an explicit solution given by (2.13). We assume that at time 0 there is no larvicide in the segment of river ($c^0 = 0$, $u_0 = 0$). We have the following explicit solution.

THEOREM 3.1. *Assume that the hypotheses of Theorem 2.1 are verified. Let $E = 0$, $V(x) > 0$, $R(x) \geq 0$, $A(x) > 0$ be continuous and*

$$\forall i = 1, \cdots, N, \qquad Q(x_{i+1}) \leq Q(x_i) e^{\int_{x_i}^{x_{i+1}} (R(y)/V(y))\, dy}. \tag{3.4}$$

Problem 1 has a solution only if $x_1 = 0$. The distribution of masses

$$\begin{aligned} \tilde{m}_1 &= u_P Q(x_1)\ e^{\int_{x_1}^{x_2} (R(y)/V(y))\, dy}, \qquad && x_1 = 0, \\ \tilde{m}_i &= u_P Q(x_i)\ [e^{\int_{x_i}^{x_{i+1}} (R(y)/V(y))\, dy} - 1], && 2 \leq i \leq N, \end{aligned} \tag{3.5}$$

is a minimizing solution of Problem 1. Moreover, the corresponding dose $\tilde{u}$ satisfies

$$\tilde{u}(x_i^-) = u_P, \qquad 1 \leq i \leq N+1, \tag{3.6}$$

and the total mass of sprayed larvicide is

$$M_N = \sum_{i=1}^{N} \tilde{m}_i = u_P \left\{ Q(0) + \sum_{i=1}^{N} Q(x_i) \left[e^{\int_{x_i}^{x_{i+1}} (R(y)/V(y))\, dy} - 1 \right] \right\}. \tag{3.7}$$

If conditions (3.4) are verified with strict inequalities, then the points x_i's are distinct, that is, $0 = x_1 < x_2 < \cdots < x_{N-1} < x_N < L$ and the solution (3.5) is the unique minimizing solution.

Proof. For the proof, see the Appendix. □

Hypothesis (3.4) is verified when the volume of water per second is not increasing as we go downstream. If there is an important new inflow of water at one point that significantly changes the concentration, then the problem has to be set up as a system of two or more segments of river with appropriate mixing conditions at the points of junction.

It is also possible to find the best positions of the N sites.

THEOREM 3.2. *Assume that, in addition to the assumptions of Theorem* 3.1, *V and A (and hence $Q = VA$) belong to $C^1([0, L])$ and*

$$(3.8) \qquad Q(y) < Q(x)e^{\int_x^y (R(z)/V(z))\,dz}, \quad \forall x, y, \quad 0 \le x < y \le L \quad \text{and} \quad Q(L) > 0.$$

Then the solution of Problem 2 *is given by $\tilde{m}$ and $\tilde{u}$ in Theorem* 3.1 *and the optimal positions of the N sites must verify the conditions*

$$0 = x_1 < x_2 < \cdots < x_{i-1} < x_i < \cdots < x_N < x_{N+1} = L$$

and the following set of equations

$$(3.9) \quad e_{i-1} - 1 = \frac{a(x_i)Q(x_i) - Q'(x_i)}{a(x_i)Q(x_{i-1})}(e_i - 1) + \frac{Q(x_i) - Q(x_{i-1})}{Q(x_{i-1})}, \qquad 2 \le i \le N,$$

$$(3.10) \qquad \prod_{i=1}^{N} e_i = e^{\int_0^L a(y)\,dy},$$

where

$$(3.11) \qquad a(y) = \frac{R(y)}{V(y)}, \quad e_i = e^{\int_{x_i}^{x_{i+1}} a(y)\,dy}, \quad 1 \le i \le N.$$

Proof. For the proof, see the Appendix. □

In the general case e_N will be one of the zeros of a polynomial of degree N such that $e_i > 1$, $1 \le i \le N$. When Q is constant the e_i's are constant and we obtain

$$(3.12) \qquad \int_{x_i}^{x_{i+1}} a(y)\,dy = \frac{1}{N}\int_0^L a(y)\,dy, \qquad 1 \le i \le N.$$

If, in addition, a is constant, then

$$(3.13) \qquad x_{i+1} - x_i = \frac{L}{N}, \qquad 1 \le i \le N.$$

Remark 3.1. For a fixed $N \ge 1$ the required total mass of product is

$$(3.14) \qquad M_N = \sum_{i=1}^{N} \tilde{m}_i = u_P \left\{ Q(0) + \sum_{i=1}^{N} Q(x_i)\left[e^{\int_{x_i}^{x_{i+1}} (R(y)/V(y))\,dy} - 1\right]\right\}$$

and necessarily for $N' > N$,

$$(3.15) \qquad \inf_{m', x'} M_{N'} \le \inf_{m, x} M_N.$$

Assume that as N goes to ∞, $\max\{x_{i+1}-x_i : 1 \leq i \leq N\}$ goes to zero. Then as N goes to ∞,

$$\begin{aligned} M_N &\simeq u_P \left\{ Q(0) + \sum_{i=1}^{N} Q(x_i) \frac{R(x_i)}{V(x_i)} (x_{i+1} - x_i) \right\} \\ &\simeq u_P \left\{ Q(0) + \int_0^L Q(x) \frac{R(x)}{V(x)}\, dx \right\} \overset{\text{def}}{=} M_\infty. \end{aligned} \tag{3.16}$$

This is a lower bound on the amount necessary to satisfy the constraint (2.6). □

Going back to formula (3.5) we obtain as N goes to ∞ and $|x_{i+1}-x_i| \to 0$,

$$\begin{aligned} m_i &\simeq u_P Q(x_i) \frac{R(x_i)}{V(x_i)} (x_{i+1} - x_i), \qquad i \geq 2, \\ m_1 &\simeq u_P \left[Q(0) \frac{R(0)}{V(0)} (x_2 - x_1) + Q(0) \right], \end{aligned}$$

and the optimal control becomes a measure

$$dm(x) = u_P \left[Q(x) \frac{R(x)}{V(x)} + Q(0) \delta_0 \right] dx, \tag{3.17}$$

which consists of a regular part over $[0,L]$ and a singular part, the Dirac delta function δ_0, at $x=0$. Recalling that for each N, $\tilde{u}$ is the solution and $\tilde{u}(x_i^-) = u_P$, then it is expected that as N goes to ∞,

$$u(x) = u_P \quad \text{for almost all } x \in [0,L]. \tag{3.18}$$

This essentially provides the solution to Problem 3 and we shall now make the above considerations more precise.

As suggested by the previous asymptotic estimate, it is wise to enlarge the space of controls for Problem 3 to the space of positive measures: that is, elements m in $C([0,L])'$ such that

$$\forall v \geq 0, \quad v \in C([0,L]), \quad \int_0^L v(x)\, dm(x) \geq 0. \tag{3.19}$$

The state (u, u_L) is now a solution of the variational problem: to find (u, u_L) in $L^2(0,L) \times \mathbb{R}$ such that for all v in $H^1(0,L)$,

$$b_0((u,u_L),v) = \int_0^L \frac{v(x)}{A(x)}\, dm(x), \tag{3.20}$$

where b_0 is the bilinear form (2.14). The associated cost function becomes

$$M(m) = \int_0^L dm(x). \tag{3.21}$$

Since $H^1(0,L) \subset C([0,L])$, problem (3.20) is well posed and has a unique solution (u, u_L) in $L^2(0,L) \times \mathbb{R}$ and Problem 3 can be formulated as follows:

$$\inf_{\substack{0 \le m \in C([0,L])' \\ u(x) \ge u_P, u_L \ge u_P}} \int_0^L dm(x). \tag{3.22}$$

The variable u_L can be viewed as the right-hand side value $u(L^+)$ of u. When u is continuous at $x = L$, $u_L = u(L)$. As such the constraint $u_L \ge u_P$ must be verified. For $A \in C([0,L])$ it is easy to verify that the solution $u(x) = u_P$ and $u_L = u_P$ corresponds to the measure m_P defined by

$$\int_0^L v(x)\, dm_P(x) = b_0((u_P, u_P), Av), \tag{3.23}$$

where

$$b_0((u_P, u_P), Av) = u_P \left[\int_0^L R(x) A(x) v(x)\, dx + V(0) A(0) v(0) \right]. \tag{3.24}$$

But $A \ge 0$ and $R \ge 0$. So for $V(0) \ge 0$, m_P is positive. Therefore, $(u, u_L) = (u_P, u_P)$ is a feasible solution that turns out to be minimizing under reasonable hypotheses on A.

THEOREM 3.3. *Assume that the hypotheses of Theorem* 2.1 *are verified,* $A \in C([0,L])$, $A \ge 0$, $V(0) \ge 0$, *and* $R \ge 0$, *and that*

$$\forall (u, u_L) \in L^2(0,L) \times \mathbb{R}, \quad u \ge 0, \quad u_L \ge 0 \Rightarrow b_0((u, u_L), A) \ge 0. \tag{3.25}$$

Then $(u, u_L) = (u_P, u_P)$ *is a minimizing solution of Problem* 3, *the distribution* m_P *of larvicide is given by the measure*

$$\begin{aligned} \int_0^L w\, dm_P &= b_0((u_P, u_P), Aw), \quad \forall w \in C([0,L]) \\ &= u_P \left\{ \int_0^L Q(x) \frac{R(x)}{V(x)} w(x)\, dx + Q(0) w(0) \right\}, \end{aligned} \tag{3.26}$$

and the total amount of product used by the expression

$$\int_0^L dm_P = b_0((u_P, u_P), A), \tag{3.27}$$

where

$$\begin{aligned} b_0((u_P, u_P), A) &= u_P \left\{ \int_0^L R(x) A(x)\, dx + V(0) A(0) \right\} \\ &= u_P \left\{ \int_0^L Q(x) \frac{R(x)}{V(x)}\, dx + Q(0) \right\}. \end{aligned} \tag{3.28}$$

If condition (3.25) *is verified with a strict inequality for all positive nonzero* (u, u_L) $(u \ge 0,\ u_L \ge 0,$ *and* $(u, u_L) \ne (0,0))$, *then the above solution is unique and condition* (3.8) *is verified.*

Proof. For the proof, see the Appendix. □

At this juncture it is useful to clarify the connection between the positivity condition on the bilinear form b_0 and previous conditions, such as (3.4) or (3.8), on the flow Q.

COROLLARY 1. *Assume that the hypotheses of Theorem* 3.3 *are verified. Then the following conditions are equivalent:*

$$\text{(i)}\begin{cases} \forall\ (u,u_L)\in L^2(0,L)\times\mathbb{R}, \quad u\geq 0,\ u_L\geq 0 \\ b_0((u,u_L),A)\geq 0; \end{cases} \tag{3.25}$$

$$\text{(ii)}\begin{cases} -\dfrac{d}{dx}(VA)+RA\ \geq\ 0 \quad \text{a.e. in } [0,L], \\ V(L)A(L)\geq 0; \end{cases} \tag{3.29}$$

$$\text{(iii)}\begin{cases} -\dfrac{d}{dx}Q+\dfrac{R}{V}Q\ \geq\ 0 \quad \text{a.e. in } [0,L], \\ Q(L)\geq 0; \end{cases} \tag{3.30}$$

$$\text{(iv)}\begin{cases} Q(y)\leq Q(x)e^{\int_x^y (R(z)/V(z))\,dz}, \quad \forall\ x,y, \quad 0\leq x\leq y\leq L, \\ Q(L)\geq 0. \end{cases} \tag{3.31}$$

□

COROLLARY 2. *Assume that the hypotheses of Theorem* 3.3 *are verified. Then the following conditions are equivalent:*

$$\text{(i)}\begin{cases} \forall\ (u,u_L)\in L^2(0,L)\times\mathbb{R}, \quad u\geq 0,\ u_L\geq 0, \quad (u,u_L)\neq(0,0), \\ b_0((u,u_L),A)> 0; \end{cases} \tag{3.25$'$}$$

$$\text{(ii)}\begin{cases} -\dfrac{d}{dx}(VA)+RA> 0 \quad \text{a.e. in } [0,L], \\ V(L)A(L)> 0; \end{cases} \tag{3.29$'$}$$

$$\text{(iii)}\begin{cases} -\dfrac{d}{dx}Q+\dfrac{R}{V}Q> 0 \quad \text{a.e. in } [0,L], \\ Q(L)> 0. \end{cases} \tag{3.30$'$}$$

Moreover, any of the above conditions implies

$$\text{(iv)}\begin{cases} Q(y)< Q(x)e^{\int_x^y (R(z)/V(z))\,dz}, \quad \forall\ x,y, \quad 0\leq x< y\leq L, \\ Q(L)> 0. \end{cases} \tag{3.8}$$

□

Remark 3.2. Condition (3.30) is verified for a river where there is no inflow of water:

$$Q(x)\ \geq\ 0, \quad \frac{dQ}{dx}(x)\ \leq\ 0, \quad \forall x\in[0,L].$$

This means that losses of water are permitted, but an important inflow of water at a point would have to be explicitly modelled by an appropriate balance condition. □

3.2.2. The case with diffusion. When E is not zero, we no longer have an explicit solution and Problems 1 and 2 are best solved numerically. Problem 3 can be solved by the method of the previous section. So we choose as the space of controls for Problem 3 the space of positive measures m in $C([0,L])'$. The state u is now the solution of the variational problem: to find u in $H^1(0,L)$ such that

$$b(u,v) = \int_0^L \frac{v(x)}{A(x)}\,dm(x), \quad \forall v \in H^1(0,L), \tag{3.32}$$

where b is the bilinear form (2.26). The associated cost function is expression (3.21). Problem 3 can be formulated as follows:

$$\inf_{\substack{0 \le m \in C([0,L])' \\ u(x) \ge u_P}} \int_0^L dm(x). \tag{3.33}$$

As in §3.2.1, it is easy to verify that the solution $u(x) = u_P$ corresponds to the measure m_P defined by

$$\int_0^L v(x)\,dm_P(x) = b(u_P, Av), \tag{3.34}$$

where

$$b(u_P, Av) = u_P\left\{\int_0^L R(x)A(x)v(x)\,dx + \beta(L)A(L)v(L) + \beta(0)A(0)v(0)\right\}. \tag{3.35}$$

If $R \ge 0$ and $A \ge 0$, it is positive for all $v \ge 0$. Therefore, $u = u_P$ is a feasible solution that turns out to be minimizing under reasonable hypotheses on A.

THEOREM 3.4. *Assume that, in addition to the hypotheses of Theorem* 2.3, *A belongs to $H^1(0,L)$ and that*

$$\forall w \in H^1(0,L), \quad w \ge 0 \Rightarrow b(w,A) \ge 0. \tag{3.36}$$

Then $u = u_P$ is a minimizing solution of Problem 3, *and the distribution m_P of larvicide is given by the positive measure*

$$\int_0^L w\,dm_P = b(u_P, Aw), \quad \forall w \in C([0,L]), \tag{3.37}$$

where

$$\begin{aligned} b(u_P, Aw) = u_P\bigg\{ &\int_0^L R(x)A(x)w(x)\,dx \\ &+ \frac{1}{2}[\sqrt{V(L)^2 + 4E(L)R(L)} - V(L)]A(L)w(L) \\ &+ \frac{1}{2}[\sqrt{V(0)^2 + 4E(0)R(0)} + V(0)]A(0)w(0)\bigg\}. \end{aligned} \tag{3.38}$$

The total amount of product used is given by the expression

$$\int_0^L dm_P = b(u_P, A), \tag{3.39}$$

and

$$\begin{aligned}\int_0^L dm_P = u_P\Bigg\{ &\int_0^L R(x)A(x)\,dx \\ &+\frac{1}{2}[\sqrt{V(L)^2+4E(L)R(L)}-V(L)]A(L) \\ &+\frac{1}{2}[\sqrt{V(0)^2+4E(0)R(0)}+V(0)]A(0)\Bigg\}.\end{aligned} \tag{3.40}$$

If condition (3.36) *is verified with a strict inequality for all* $w>0$, *the above solution is unique.*

Proof. For the proof, see the Appendix. □

Remark 3.3. The minimizing distribution m_P of larvicide contains a distributed term on $[0,L]$ and two impulses at $x=0$ and $x=L$. □

COROLLARY 3. (i) *If* A *belongs to* $H^1(0,L)$, *condition* (3.36) *can be restated as follows:*

$$\begin{aligned}&\forall w\in H^1(0,L),\quad w\geq 0,\\ &b(w,A)=\int_0^L\left[E(x)\frac{dA}{dx}(x)+V(x)A(x)\right]\frac{dw}{dx}+R(x)A(x)w\,dx\\ &\qquad+\beta(L)A(L)w(L)+\beta(0)A(0)w(0)\ \geq\ 0.\end{aligned} \tag{3.41}$$

If A *belongs to* $H^2(0,L)$, *and* E *and* V *to* $H^1(0,L)$, *condition* (3.36) *is equivalent to*

$$\begin{aligned}&-\frac{d}{dx}\left[E\frac{dA}{dx}+VA\right]+RA\ \geq\ 0\ \textit{ for almost all } x\in[0,L],\\ &E(L)A'(L)+\frac{1}{2}[\sqrt{V(L)^2+4E(L)R(L)}+V(L)]A(L)\geq 0,\\ &-E(0)A'(0)+\frac{1}{2}[\sqrt{V(0)^2+4E(0)R(0)}+V(0)]A(0)\geq 0.\end{aligned} \tag{3.42}$$

(ii) *If, in addition to the hypotheses of Theorem* 3.4, A *is constant and*

$$-\frac{dQ}{dx}+\frac{R}{V}Q\ \geq\ 0,\quad \forall x\in[0,L], \tag{3.43}$$

then (3.40) *is verified. In particular, if* A, E, R, *and* V *are positive constants,* (3.42) *is verified and*

$$\int_0^L dm_P=u_PA[LR+\sqrt{V^2+4ER}]. \tag{3.44}$$

□

Remark 3.4. Setting $E=0$ in (3.40), we recover expressions (3.27)–(3.28). □

4. Numerical approximation. In the remaining sections of this paper we assume that the initial concentration c^0 of larvicide in the river is zero and hence $C_0=C_L=0$.

4.1. Approximation of the state equation.

4.1.1. The case without diffusion. When $E=0$ we have the explicit solution (2.13) and there is no need for an approximation of system (2.11)–(2.12).

4.1.2. The case with diffusion. Problem 3 has an explicit solution. Problems 1 and 2 require a numerical method to solve the diffusion-transport equation (2.24) with impulse controls. When the transport dominates, it is well known that classical finite elements do not produce good results. Mixed finite elements combined with the technique of Lesaint and Raviart [11] produce better results by introducing the right amount of numerical dissipation. Their generalization to dimension 2 uses the elements of Raviart and Thomas (cf. Raviart [14]).

In a mixed method a separate approximation is used for the variable u and its derivative $\frac{du}{dx}$. This amounts to introducing a new variable p:

$$E^{-1}(x)p = \frac{du}{dx}. \tag{4.1}$$

Its substitution in the system of equations (2.24) yields the boundary conditions

$$-p(0) + \beta(0)u(0) = 0, \qquad p(L) + \beta(L)u(L) = 0 \tag{4.2}$$

and the equation

$$-\frac{dp}{dx} + V(x)\frac{du}{dx} + R(x)u = \sum_{i=1}^{N} \frac{m_i}{A(x_i)}\delta_i(x), \qquad x \in [0, L]. \tag{4.3}$$

For numerical reasons we keep the transport term $V\frac{du}{dx}$ in (4.3). This term will be treated "à la Lesaint and Raviart [11]" in order to introduce a sufficient amount of numerical dissipation.

To solve (4.1)–(4.3) numerically we could choose a partition of the interval $[0, L]$ such that the positions of the sites $\{x_i : 2 \leq i \leq N\}$ coincide with nodes of the partition. However, our objective is to find the best positions of the sites without making the partition an integral part of our optimization problem. Let $\{\xi_j : 0 \leq j \leq J\}$ be a uniform partition of the interval $[0, L]$,

$$\xi_j = jh, \qquad 0 \leq j \leq J, \qquad h = \frac{L}{J}, \tag{4.4}$$

for some integer $J > 0$ and consider the system of equations (4.1), (4.2), and

$$-\frac{dp}{dx} + V(x)\frac{du}{dx} + R(x)u = f + \sum_{j=0}^{J} c_j\delta_j \tag{4.5}$$

for some coefficients c_j in $\mathbb{R}$, Dirac delta functions δ_j at ξ_j, and an arbitrary function f in $C([0, L])$.

Again it is important to recall that we do not want to necessarily put the injection points at the discretization nodes $\{\xi_j\}$ and solve a combinatorial problem with a very fine partition and a large number of variables. So when the injection m_i occurs at a point x_i inside the interval $I_j = [\xi_{j-1}, \xi_j]$, we distribute the impulsion between the two adjacent nodes ξ_{j-1} and ξ_j by introducing a weighting function

$$B(\zeta) = \begin{cases} \hat{B}(\zeta), & |\zeta| \leq 1, \\ 0, & \text{otherwise}, \end{cases} \tag{4.6}$$

where, for instance, $\hat{B}(\zeta)$ can be chosen as

$$\hat{B}(\zeta) = \tfrac{1}{2}[1 + \cos \pi\zeta], \quad (1 - |\zeta|)^2(1 + 2|\zeta|), \quad \text{or} \quad 1 - |\zeta|. \tag{4.7}$$

The first two choices are continuously differentiable in $]-1, 1[$ and $\mathbb{R}$, respectively. To make the connection between the original equation (4.3) and the mesh-dependent formulation (4.5), assign the coefficients

$$c_j = \sum_{i=1}^{N} \frac{m_i}{A(x_i)} B\left(\frac{\xi_j - x_i}{h}\right), \qquad 0 \le j \le J. \tag{4.8}$$

Note that for each i,

$$\sum_{j=0}^{J} B\left(\frac{\xi_j - x_i}{h}\right) = 1.$$

With this type of formulation the state now continuously depends on the position of the injection point inside a given interval I_j.

We now give a discontinuous mesh-dependent formulation for the underlying problem (4.1)–(4.2)–(4.5). Since the approximation will be discontinuous at each node, we are naturally led to introduce traces $\{U_j : 0 \le j \le J\}$ at each node ξ_j. On each interval $I_j = [\xi_{j-1}, \xi_j]$ we write (4.1) in the following weak form:

$$\begin{gathered}\int_{I_j} \left\{ E^{-1}(x) p q + u \frac{dq}{dx} \right\} dx + U_{j-1} q(\xi_{j-1}^+) - U_j q(\xi_j^-) = 0 \\ \forall q \in H^1(I_j), \qquad 1 \le j \le J\end{gathered} \tag{4.9}$$

with boundary and jump conditions

$$-p(0^+) + \beta(0) U_0 = c_0, \qquad p(\xi_J^-) + \beta(L) U_J = c_J, \tag{4.10}$$

$$p(\xi_j^-) - p(\xi_j^+) = c_j, \qquad 1 \le j \le J - 1. \tag{4.11}$$

It is easy to verify that for the continuous problem

$$U_0 = u(\xi_0^+), \quad u(\xi_j^-) = U_j = u(\xi_j^+), \quad 1 \le j < J, \quad U_J = u(\xi_J^-),$$

and that the U_j's are indeed the traces of the function u at the discretization nodes ξ_j's. Equation (4.5) is also written in weak form but the term in u is treated "à la Lesaint and Raviart [11]":

$$\begin{aligned}&\int_{I_j} \left\{ \left[-\frac{dp}{dx} + R(x) u \right] w - u \frac{d}{dx}[V(x) w] \right\} dx \\ &\quad + u(\xi_j^-) V(\xi_j) w(\xi_j^-) - u(\xi_{j-1}^-) V(\xi_{j-1}) w(\xi_{j-1}^+) = \int_{I_j} f w \, dx, \quad \forall w \in H^1(I_j),\end{aligned} \tag{4.12}$$

where $u(0^-) \overset{\text{def}}{=} U_0$. So the solution (u, p) can now be discontinuous at each ξ_j and

$$u_{|I_j} \text{ and } p_{|I_j} \text{ in } H^1(I_j). \tag{4.13}$$

The above framework was introduced in order to be able to choose a mesh-dependent discontinuous approximation such that

$$u_{|I_j},\quad w_{|I_j} \in P^0(I_j),\quad p_{|I_j},\ q_{|I_j} \in P^1(I_j), \tag{4.14}$$

where $P^k(I_j)$ is the space of polynomials of degree less than or equal to $k \geq 0$ on I_j. When the c_j's are zero, there are no jumps and p is continuous at each node ξ_j. In dimension 2 this would correspond to the elements of Raviart and Thomas (cf. Raviart [14]) with a continuous normal trace. Here we have relaxed that continuity to a prescribed jump condition given by (4.11). This relaxation is incorporated in the global formulation by introducing Lagrange multipliers W_j's, $0 \leq j \leq J$, to specify the jumps. The complete weak formulation becomes: to find $\{u_j \in P^0(I_j),\ p_j \in P^1(I_j),\ 1 \leq j \leq J\}$ and $\{U_j \in \mathbb{R},\ 0 \leq j \leq J\}$ such that for all $\{w_j \in P^0(I_j),\ q_j \in P^1(I_j),\ 1 \leq j \leq J\}$ and $\{W_j \in \mathbb{R},\ 0 \leq j \leq J\}$,

$$\begin{aligned}
&\sum_{j=1}^{J}\left\{\int_{I_j}\left[E^{-1}(x)pq + u\frac{dq}{dx}\right]dx \ - U_j q(\xi_j^-) + U_{j-1}q(\xi_{j-1}^+)\right.\\
&\qquad + \int_{I_j}\left[-\frac{dp}{dx} + Ru\right]w - u\frac{d}{dx}[V(x)w] - fw\,dx\\
&\qquad \left. + \ u(\xi_j^-)V(\xi_j)w(\xi_j^-) - u(\xi_{j-1}^-)V(\xi_{j-1})w(\xi_{j-1}^+)\right\}\\
&\quad + \sum_{j=1}^{J-1}[p(\xi_j^-) - p(\xi_j^+) - c_j]W_j\\
&\quad + [-p(0^+) + \beta_0 U_0 - c_0]W_0 + [p(\xi_J^-) + \beta_J U_J - c_J]W_J = 0.
\end{aligned} \tag{4.15}$$

It is important to note that (4.15) not only provides a weak formulation of (4.1), (4.2), and (4.5), but also the weak formulation of the adjoint system of equations that will be used in the construction of the adjoint state: the elements u, p, and U will play the role of test functions and w, q, and W will be the adjoint state variables.

The complete discretization of the state equation is obtained by using the trapeze quadrature formula for the approximation of the integrals in (4.15). The step by step derivation of (4.16)–(4.19) can be found in the Appendix. The piecewise constant function u is represented by the sequence $\{u_j\ :\ 1 \leq j \leq N\}$ where $u_j \in \mathbb{R}$ is the constant value of u on the interval I_j. The linear functions $p_j \in P^1(I_j)$ are represented by their values $p(\xi_{j-1}^+)$ and $p(\xi_j^-)$ at the left and right boundaries of the interval I_j. Similar notation is used for the piecewise constant function w and the piecewise linear function q.

The final algorithm for the state yields a linear system of the form

$$\Lambda\vec{u} = h\vec{f} + C\vec{c} \tag{4.16}$$

for the $(J+2)$-vectors $\vec{u} = \{U_0, u_1, \cdots, u_J, U_J\}$ and $\vec{f} = \{0, f_1, \cdots, f_J, 0\}$, and the $(J+1)$-vector $\vec{c} = \{c_0, c_1, \cdots, c_J\}$, and all the other variables can be obtained by explicit formulae. In other words, we have reduced the size of the linear system (4.15) by isolating variables that can be computed from the vectors $\vec{u}$ and $\vec{c}$:

$$U_j = \frac{u_{j+1} + u_j}{2} + \frac{1}{4E_j}c_j, \qquad 1 \leq j \leq J-1, \tag{4.17a}$$

$$\begin{aligned}&p(\xi_0^+) = 2\frac{E_0}{h}(u_1 - U_0), \qquad p(\xi_J^-) = \frac{2E_J}{h}[U_J - u_J],\\ &p(\xi_j^-) = \frac{E_j}{h}[u_{j+1} - u_j] + \frac{c_j}{2}, \qquad p(\xi_j^+) = p(\xi_j^-) - c_j, \quad 1 \le j \le J-1.\end{aligned} \tag{4.17b}$$

Note that

$$\frac{1}{2}[p(\xi_j^+) + p(\xi_j^-)] = \frac{E_j}{h}[u_{j+1} - u_j], \qquad 1 \le j \le J-1.$$

The matrices Λ and C are specified by the following system of equations:

$$\begin{aligned}&\left[\frac{2E_0}{h} + \beta_0\right]U_0 - \frac{2E_0}{h}u_1 = c_0\\ &-\left[\frac{2E_0}{h} + V_0\right]U_0 + \left[\frac{2E_0 + E_1}{h} + hR_1 + V_0\right]u_1 - \frac{E_1}{h}u_2 = hf_1 + \frac{c_1}{2}\\ &\begin{cases} -\left[\frac{E_{j-1}}{h} + V_{j-1}\right]u_{j-1} + \left[\frac{E_{j-1}+E_j}{h} + hR_j + V_{j-1}\right]u_j - \frac{E_j}{h}u_{j+1}\\ = hf_j + \frac{c_{j-1}+c_j}{2}, \qquad 2 \le j \le J-1,\end{cases}\\ &\begin{cases} -\left[\frac{E_{J-1}}{h} + V_{J-1}\right]u_{J-1} + \left[\frac{E_{J-1}+2E_J}{h} + hR_J + V_{J-1}\right]u_J - \frac{2E_J}{h}U_J\\ = hf_J + \frac{c_{J-1}}{2},\end{cases}\\ &-\frac{2E_J}{h}u_J + \left[\frac{2E_J}{h} + \beta_J\right]U_J = c_J,\end{aligned} \tag{4.18}$$

where $\beta_J = \beta(L)$, $\beta_0 = \beta(0)$,

$$\begin{aligned}&E_j = E(\xi_j), \qquad V_j = V(\xi_j), \qquad 0 \le j \le J,\\ &R_j = \frac{R(\xi_{j-1})+R(\xi_j)}{2}, \quad f_j = \frac{f(\xi_{j-1})+f(\xi_j)}{2}, \quad 1 \le j \le J.\end{aligned} \tag{4.19}$$

The right-hand side of the final system (4.16) is precisely the Raviart–Thomas system (cf. Raviart [14]) where the impulsions appear on the left-hand side. A jump occurring at the point ξ_j is distributed between the two neighbouring intervals I_j and I_{j+1}.

4.2. Approximation of the optimization problem.

4.2.1. The case without diffusion. Again, everything here is explicit and can be solved directly. We shall use the optimal solution for $E = 0$ as an initial guess for the optimization problem with diffusion.

4.2.2. The case with diffusion. Problem 2 can now be formulated as a nonlinear optimization problem where the *optimization variables* are the N-vectors $\vec{m} = (m_1, m_2, \cdots, m_N)^{\mathrm{T}}$ and $\vec{x} = (x_1, x_2, \cdots, x_N)^{\mathrm{T}}$:

$$\text{Min} \sum_{i=1}^{N} m_i \tag{4.20}$$

subject to the *state equation*

$$\Lambda\vec{u} = C\vec{c}, \tag{4.21}$$

$$U_j = \frac{u_{j+1} + u_j}{2} + \frac{1}{4E_j}c_j, \qquad 1 \le j \le J-1, \tag{4.22}$$

where $\vec{u}$ is the $(J+2)$-vector $\vec{u} = (U_0, u_1, \cdots, u_{J-1}, U_J)^\top$, $\vec{c}$ is the $(J+1)$-vector $\vec{c} = (c_0, c_1, \cdots, c_J)^\top$ with components

$$c_j = \sum_{i=1}^{N} \frac{m_i}{A(x_i)} B\left(\frac{\xi_j - x_i}{h}\right), \qquad 0 \le j \le J, \tag{4.23}$$

the matrices Λ and C are specified by the set of equations (4.18), and the functions B by one of the functions (4.7). Note that $\vec{f} = 0$ in (4.16) since $f = 0$ in §2.3.2.

In addition, we have the following *constraints* on the *optimization variables*

$$m_i \ge 0, \qquad 1 \le i \le N, \tag{4.24}$$

and *constraints* on the *state variables*

$$U_j \ge u_P, \qquad 0 \le j \le J, \tag{4.25}$$

$$u_j \ge u_P, \qquad 1 \le j \le J. \tag{4.26}$$

The variables $p(\xi_j^-)$ and $p(\xi_j^+)$ do not enter into the (discretized) optimization problem and can be recovered from (4.17b)

$$\begin{aligned} p(\xi_0^+) &= 2\frac{E_0}{h}(u_1 - U_0), \qquad p(\xi_J^-) = \frac{2E_J}{h}[U_J - u_J], \\ p(\xi_j^-) &= \frac{E_j}{h}[u_{j+1} - u_j] + \frac{c_j}{2}, \quad p(\xi_j^+) = p(\xi_j^-) - c_j, \quad 1 \le j \le J-1. \end{aligned} \tag{4.27}$$

In the above formulation, the constraint on the dose becomes the constraint (4.25) on the traces and we have also required that the L^2-function u between two consecutive nodes verifies the constraint (4.26). It is readily seen from identity (4.22) that if the set of constraints (4.26) plus $U_0 \ge u_P$ and $U_J \ge u_P$ are satisfied, then the set of constraints (4.25) is also satisfied. However, the converse is not true and (4.25) is a weaker condition than (4.26) plus $U_0 \ge u_P$ and $U_J \ge u_P$. So there is a certain amount of flexibility in the way in which we wish to deal with the state constraints. We can either choose the *weak condition*

$$U_j \ge u_P, \qquad 0 \le j \le J, \tag{4.25}$$

or the *strong condition*

$$u_j \ge u_P, \quad 1 \le j \le J, \quad U_0 \ge u_P, \quad U_J \ge u_P. \tag{4.28}$$

However, for completeness in the subsequent computations of various derivatives, we shall keep the full constraints (4.25) and (4.26). Then the appropriate terms will be deleted in the final expressions depending on the choice between condition (4.25) and (4.28).

Finally, in the above analysis we have not used the constraints

$$0 \le x_i \le L, \qquad 1 \le i \le N, \tag{4.29}$$

since any amount of product used outside the interval $[0, L]$ will increase the objective function without increasing the dose. If for some reason our thinking was not correct, it will be interesting to discover how the algorithm can take advantage of the possibility of spraying larvicide outside of $[0, L]$.

As we shall see in §5 we have used a penalization technique for the inequality constraints combined with a simple descent method to solve this problem. We now complete our analysis by specifying the penalized objective function

$$\begin{aligned} M_\epsilon(\vec{m}_N, \vec{x}_N) = &\sum_{i=1}^{N} m_i + \frac{1}{2\epsilon_1}\sum_{i=1}^{N}([m_i]^-)^2 \\ &+ \frac{1}{2\epsilon_2}\sum_{j=0}^{J}([u_P - U_j]^+)^2 + \frac{1}{2\epsilon_3}\int_0^L ([u_P - u(x)]^+)^2\, dx, \end{aligned} \tag{4.30}$$

for $\epsilon_1 > 0$, $\epsilon_2 > 0$, $\epsilon_3 > 0$ and providing the gradient of M_ϵ with respect to the masses and positions.

At this juncture it is useful to restate the penalized nonlinear optimization problem, which we shall actually use in the numerical computations:

$$\begin{aligned} \text{Min} \Bigg\{ &\sum_{i=1}^{N} m_i + \frac{1}{2\epsilon_1}\sum_{i=1}^{N}([m_i]^-)^2 \\ &+ \frac{1}{2\epsilon_2}\sum_{j=0}^{J}([u_P - U_j]^+)^2 + \frac{h}{2\epsilon_3}\sum_{j=1}^{J}([u_P - u_j]^+)^2 \Bigg\}, \end{aligned} \tag{4.31}$$

subject to the *state equations*

$$\Lambda \vec{u} = C\vec{c} \tag{4.21}$$

$$U_j = \frac{u_{j+1} + u_j}{2} + \frac{1}{4E_j} c_j, \qquad 1 \le j \le J - 1, \tag{4.22}$$

where $\vec{u}$ is the $(J + 2)$-vector $\vec{u} = (U_0, u_1, \cdots, u_{J-1}, U_J)^\top$ and $\vec{c}$ is the $(J + 1)$-vector $\vec{c} = (c_0, c_1, \cdots, c_J)^\top$ with components

$$c_j = \sum_{i=1}^{N} \frac{m_i}{A(x_i)} B\left(\frac{\xi_j - x_i}{h}\right), \qquad 0 \le j \le J. \tag{4.23}$$

The partial derivatives of M_ϵ with respect to m_i and x_i can be obtained by using a Lagrangian method. The Lagrangian is defined as the sum of M_ϵ plus the weak formulation (4.15). In this definition the test functions appearing in (4.15) will play the role of the Lagrange multipliers. For convenience, we introduce the notation

$$Y = (u_1, \cdots, u_J, p_1, \cdots, p_J, U_0, \cdots, U_J), \tag{4.32}$$

$$P = (w_1, \cdots, w_J, q_1, \cdots, q_J, W_0, \cdots, W_J), \tag{4.33}$$

and denote the above Lagrangian by $L(\vec{m}, \vec{x}, Y, P)$. The penalized cost can now be expressed as follows:

$$M_\epsilon(\vec{m}, \vec{x}) = \inf_Y \sup_P L(\vec{m}, \vec{x}, Y, P) \tag{4.34}$$

and the partial derivatives are given by

$$\frac{\partial M_\epsilon}{\partial m_i}(\vec{m},\vec{x}) = \frac{\partial L}{\partial m_i}(\vec{m},\vec{x},\hat{Y},\hat{P}), \qquad \frac{\partial M_\epsilon}{\partial x_i}(\vec{m},\vec{x}) = \frac{\partial L}{\partial x_i}(\vec{m},\vec{x},\hat{Y},\hat{P}), \tag{4.35}$$

where the pair $(\hat{Y},\hat{P})$ is the solution of the saddle point equations of the convex-concave differentiable Lagrangian functional $(Y,P) \mapsto L(\vec{m},\vec{x},Y,P)$. Partial derivatives with respect to the P variables yield the set of equations (4.15). So $\hat{Y}$ coincides with the solution of the variational equation (4.15) for the state. Partial derivatives with respect to the Y variables yield a set of variational equations for the *adjoint state* or *adjoint variables* $\hat{P}$. Its solution is unique. For theoretical justification of the techniques the reader is referred to Correa and Seeger [4] and for applications to the computation of shape derivatives, to Delfour and Zolésio [5], [6]. The detailed computations can be found in the Appendix. As in the case of the state equations, some of the adjoint variables can be explicitly expressed in terms of a shorter vector. More specifically, the adjoint variables are given by the following equations:

$$\Lambda^* \vec{w} = C\vec{d} + h\vec{g}, \tag{4.36}$$

with the vectors $\vec{w} = (\hat{W}_0, \hat{w}_1, \cdots, \hat{w}_J, \hat{W}_J)$, $\vec{g} = (0, g_1, \cdots, g_{J-1}, g_J, 0)$, and $\vec{d} = (d_0, d_1, \cdots, d_{J-1}, d_J)$, and the matrices Λ and C from system (4.16), where Λ^* is the transposed matrix of Λ and

$$\begin{aligned} d_j &= \frac{1}{\epsilon_2}[u_P - U_j]^+, \qquad 0 \le j \le J, \\ g_j &= \frac{1}{\epsilon_3}[u_P - u_j]^+, \qquad 1 \le j \le J. \end{aligned} \tag{4.37}$$

Moreover,

$$\hat{W}_j = \frac{\hat{w}_j + \hat{w}_{j+1}}{2} + \frac{h}{4E_j} d_j, \qquad 1 \le j \le J-1, \tag{4.38}$$

$$\begin{aligned} &\hat{q}(\xi_j^+) = \frac{E_j}{h}[\hat{w}_j - \hat{w}_{j+1}] + \frac{d_j}{2}, \quad \hat{q}(\xi_j^-) = \hat{q}(\xi_j^+) - d_j, \quad 1 \le j \le J-1, \\ &\hat{q}(\xi_0^+) = \frac{2E_0}{h}[\hat{W}_0 - \hat{w}_1], \qquad \hat{q}(\xi_J^-) = \frac{2E_J}{h}[\hat{w}_J - \hat{W}_J]. \end{aligned} \tag{4.39}$$

Note that

$$\frac{1}{2}[\hat{q}(\xi_j^+) + \hat{q}(\xi_j^-)] = \frac{E_j}{h}[\hat{w}_j - \hat{w}_{j+1}], \qquad 1 \le j \le J-1.$$

Equation (4.36) only contains the vector $\vec{w}$ since the $\hat{q}_j$'s and the other $\hat{W}_j$'s can be explicitly computed from identities (4.38)–(4.39).

The partial derivatives of M_ϵ are

$$\frac{\partial M_\epsilon}{\partial m_i} = 1 + \frac{1}{\epsilon_1}[m_i]^- - \frac{1}{A(x_i)} \sum_{j=0}^{J} \hat{W}_j B\left(\frac{\xi_j - x_i}{h}\right), \qquad 1 \le i \le N, \tag{4.40}$$

$$\frac{\partial M_\epsilon}{\partial x_i} = \frac{m_i}{hA(x_i)} \sum_{j=0}^{J} \hat{W}_j B'\left(\frac{\xi_j - x_i}{h}\right), \qquad 1 \le i \le N, \tag{4.41}$$

where $B'(\zeta)$ is the derivative of the function $B(\zeta)$.

5. Numerical experimentation. We have chosen a segment of the Amoutchou river in Togo, West Africa. The estimated parameters are:

$$\begin{aligned}&V = 0.25\text{m}/\sec, \quad R = 1.1\ 10^{-4}\sec^{-1}, \quad E = 10\text{m}^2/\sec, \quad A = 6.8\text{m}^2,\\ &u_P = 0.5\text{kg} \times \sec/\text{m}^3, \quad L = 12\ 10^3\text{m}, \quad J = 120, \quad h = 100\text{m}.\end{aligned}$$

5.1. Case without diffusion (theoretical results). For $E = 0$ the theoretical optimal masses and site locations are tabulated in Table 5.1 as a function of the number of sites N for $N = 1, 2, 3$, and 20 and $N = \infty$ (the asymptotic solution).

TABLE 5.1
Optimal masses and positions for $E = 0$.

N	m_1	x_1	m_2	x_2	m_3	x_3	$M(\vec{m})$
Unit	kg	m	kg	m	kg	m	kg
1	166.9	0.					166.9
2	11.91	0.	11.06	6000			22.97
3	4.941	0.	4.090	4000	4.090	8000	13.12
20	1.107	0.	0.2568	600	0.2568	1200	5.986
∞							5.34

5.2. Case with diffusion by penalization. For a diffusion $E = 10\,\text{m}^2/\sec$ we have computed optimal masses and site locations for $N = 1$, 2, and 3 sites and the theoretical optimal total mass for the asymptotic solution ($N = \infty$). The starting points were obtained from the cases without diffusion (see Table 5.2).

TABLE 5.2
Initial masses and positions for $E = 10\,\text{m}^2/\sec$.

N	m_1	x_1	m_2	x_2	m_3	x_3	$M(\vec{m})$
Unit	kg	m	kg	m	kg	m	kg
1	166.	10.					166.
2	11.9	10.	11.0	6005			22.9
3	4.93	10.	4.08	4006	4.08	8003	13.1
∞							5.34

With a penalization parameter $\epsilon = 10^{-3}$, $\epsilon_1 = \epsilon_2 = \epsilon$, and $\epsilon_3 = h\epsilon$, the results are as presented in Table 5.3.

TABLE 5.3
Optimal masses and positions for $E = 10\,\text{m}^2/\sec$.

N	Number of	m_1	x_1	m_2	x_2	m_3	x_3	$M(\vec{m})$
Unit	iterations	kg	m	kg	m	kg	m	kg
1	850	122.	399.					122.
2	800	10.3	199.	9.40	6250			19.6
3	500	4.41	199.	3.55	4279	3.55	8275	11.5
∞								5.37

The distribution of the dose for each N is given in Figs. 5.1 and 5.2.

5.3. Case with diffusion by the simplex method. For the sake of comparison we have used the simplex method to determine the minimizing masses for a fixed number of sites and fixed locations of the spraying sites. The optimal site locations

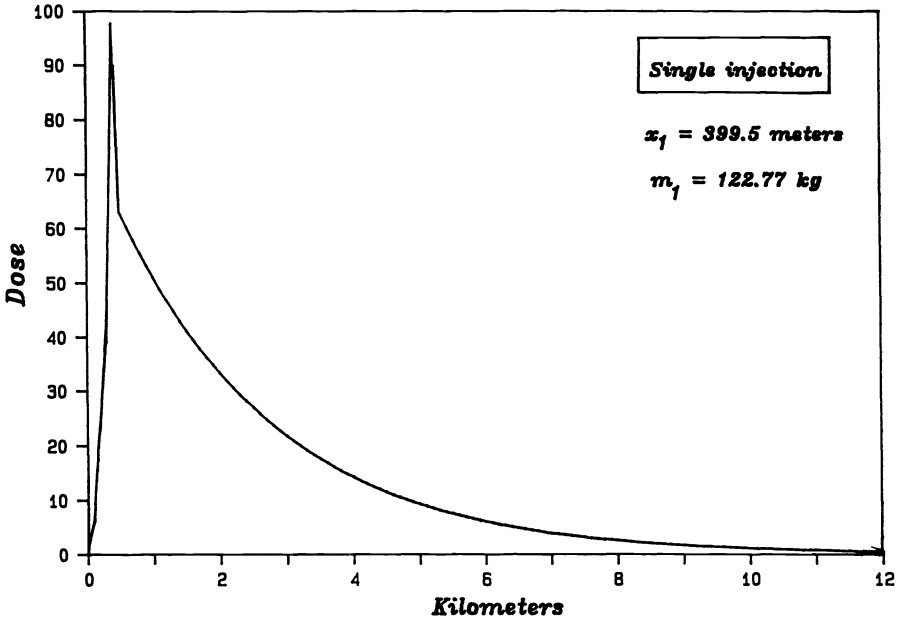


FIG. 5.1. *Dose as a function of x for $N = 1$.*

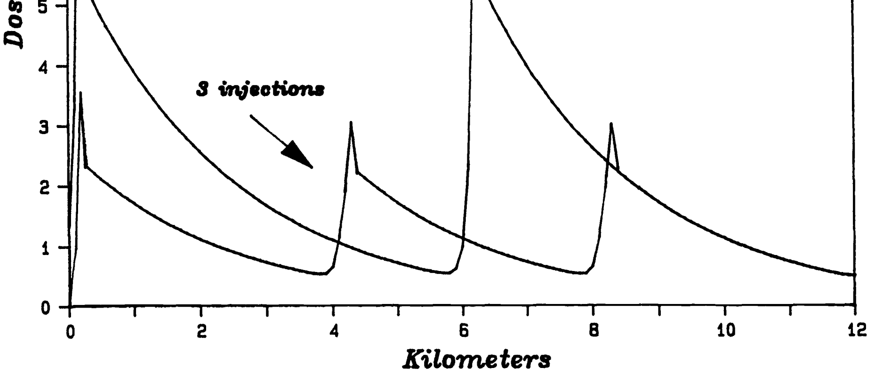


FIG. 5.2. *Dose as a function of x for $N = 2$ and 3.*

were determined by a systematic search over all discretization nodes, that is, every 50m.

The first observation is that the simplex method failed for the 12km river that contains 240 50m elements. So we considered a shorter 6km river with three injection points.

5.3.1. Gradient method with penalization. Two cases were considered:

(a) $B(\sigma) = 1 - |\sigma|$,

(b) $B(\sigma) = \frac{1}{2}[1 + \cos(\pi\sigma)]$.

For the two cases, the results are similar for approximatively 1000 iterations starting from a total initial mass of 4.438kg and the optimal positions of the case without diffusion, $E = 0$.

TABLE 5.4
Gradient method with penalization.

(a)			(b)		
i	x_i	m_i	i	x_i	m_i
Unit	m	kg		m	kg
1	49.5	1.997	1	49.9	1.989
2	2110.0	1.147	2	2092.2	1.139
3	4250.3	1.066	3	4249.9	1.080
	Total	4.210		Total	4.208

5.3.2. Simplex method for the masses. We have limited the search to ten nodes downstream of the optimal injection points of the case with no diffusion ($E = 0$). Moreover, we have introduced an upper bound

$$m_i \leq M_3, \qquad i = 1, 2, 3,$$

where M_3 is the total optimal mass for $E = 0$.

TABLE 5.5
Simplex method for the masses.

i	x_i	m_i
Unit	m	kg
1	50.	2.06103
2	2300.	1.06984
3	4300.	1.06829
	Total	4.19916

The simplex method gave an improvement of 0.2 percent over the gradient method (compare Tables 5.4 and 5.5). However, the gradient method works for a 12km river while the simplex method fails. Moreover, it is faster. For the simplex each minimization of the total mass for fixed locations requires 20 seconds of computing time (122+3 unknowns, 120+2+3 inequality constraints, and 122 equality constraints). For the gradient method one iteration (state equation, adjoint state equation, gradient evaluation, and display of the solution) requires 0.3 second. So for 1000 iterations the total time was 300 seconds, to be compared with 20,000 seconds for the simplex method.

By increasing the number of gradient iterations to 2000 with a tighter control over the descent parameters we achieved the results in Table 5.6, which improved the simplex solution by 1.6 percent.

TABLE 5.6
Gradient method with 2000 *iterations.*

i	x_i	m_i
Unit	m	kg
1	49.95	1.944
2	2199.6	1.094
3	4219.0	1.094
	Total	4.132

The closest simplex solutions were at (50m, 2200m, 4200m) with 4.226kg and at (50m, 2250m, 4200m) with 4.2044kg.

A. Appendix.

A.1. Proofs.

Proof of Theorem 2.1. Clearly, $V \in H^1(0,L)$ and $V^{-1} \in L^\infty(0,L)$ imply that $V^{-1} \in H^1(0,L)$. Under the assumption $V^{-1} \in H^1(0,L)$ and $R \in L^2(0,L)$, the system

$$\text{(A.1)}\qquad \begin{aligned} &-\frac{dv}{dx} + \frac{1}{V(x)}\left[R(x) - \frac{dV}{dx}(x)\right] v = \frac{g}{V(x)} \in L^2(0,L),\\ &v(L) = \frac{1}{V(L)} g_L \in \mathbb{R} \end{aligned}$$

has a unique solution in $H^1(0,L)$. System (A.1) is equivalent to

$$\begin{aligned} &-\frac{d}{dx}(V(x)v) + R(x)v = g,\\ &V(L)v(L) = g_L \end{aligned}$$

and the map

$$v \mapsto \left(-\frac{d}{dx}(V(x)v) + R(x)v, V(L)v(L)\right) : H^1(0,L) \longrightarrow L^2(0,L) \times \mathbb{R}$$

is an isomorphism. As a result, its adjoint is also an isomorphism and for any ℓ in $H^1(0,L)'$,

$$\begin{aligned} &\exists (u, u_L) \in L^2(0,L) \times \mathbb{R}, \quad \text{such that } \forall v \in H^1(0,L)\\ &\int_0^L u\left[-\frac{d}{dx}(V(x)v) + Rv\right] dx + u_L V(L)v(L) = \langle \ell, v\rangle_{H^1} \end{aligned}$$

has a unique solution. The last part follows by a standard argument. □

Proof of Theorem 2.2. It is sufficient to prove that the continuous bilinear form (2.18) is coercive. For all v in $H^1(\mathbb{R})$

$$V\frac{dv}{dx}v = V\frac{1}{2}\frac{dv^2}{dx} = \frac{1}{2}\frac{d}{dx}[Vv^2] - \frac{1}{2}\frac{dV}{dx}v^2.$$

Hence

$$\text{(A.2)}\qquad a(v,v) = \int_{\mathbb{R}} E\left|\frac{dv}{dx}\right|^2 + \frac{1}{2}\frac{d}{dx}[Vv^2] + \left[R - \frac{1}{2}\frac{dV}{dx}\right] v^2\, dx.$$

By hypotheses

$$V \in W^{1,\infty}(\mathbb{R}) \quad \text{and} \quad v \in H^1(\mathbb{R}) \Longrightarrow Vv^2 \in W^{1,1}(\mathbb{R}) \cap H^1(\mathbb{R})$$

and recall that for any function φ in $H^1(\mathbb{R})$,

$$\lim_{|x|\to\infty} \varphi(x) = 0.$$

Therefore, the integral of the middle term in (A.2) is zero and by hypothesis,

$$a(v,v) \geq \alpha \int_{\mathbb{R}} \left|\frac{dv}{dx}\right|^2 + |v|^2\,dx = \alpha\|v\|_{H^1}^2. \qquad \square$$

Proof of Theorem 2.3. (i) System (2.24) is clearly equivalent to the variational equation (2.25) with the bilinear form defined in (2.26). To establish the existence of a solution to (2.25) it is sufficient to show that the continuous bilinear form b is coercive. So for u in $H^1(0,L)$,

$$\begin{aligned} b(u,u) &= \int_0^L E(x)\left|\frac{du}{dx}\right|^2 + \frac{1}{2}V(x)\frac{du^2}{dx} + R(x)u^2\,dx \\ &\quad + \tfrac{1}{2}\left[\sqrt{V^2(L)+4E(L)R(L)} - V(L)\right]u(L)^2 \\ &\quad + \tfrac{1}{2}\left[\sqrt{V^2(0)+4E(0)R(0)} + V(0)\right]u(0)^2 \\ &= \int_0^L E(x)\left|\frac{du}{dx}\right|^2 + \left(R(x) - \frac{1}{2}\frac{dV}{dx}(x)\right)u^2\,dx \\ &\quad + \tfrac{1}{2}\sqrt{V^2(L)+4E(L)R(L)}\;u(L)^2 \\ &\quad + \tfrac{1}{2}\sqrt{V^2(0)+4E(0)R(0)}\;u(0)^2 \\ &\geq \int_0^L E(x)\left|\frac{du}{dx}\right|^2 + \left(R(x) - \frac{1}{2}\frac{dV}{dx}(x)\right)u^2\,dx \\ &\geq \alpha\left\|\frac{du}{dx}\right\|_{L^2}^2 + \alpha\|u\|_{L^2}^2 = \alpha\|u\|_{H^1}^2. \end{aligned}$$

Noting that the right-hand side of (2.25) is a continuous linear form on $H^1(0,L)$, the existence and uniqueness of the solution follows by the Lax–Milgram theorem.

(ii) Now we show that the solution of (2.25) on $[0,L]$ is the restriction of the solution of (2.19) on $\mathbb{R}$. To do that we use test functions with support in $]-\infty,0[$, $]0,L[$, and $]L,\infty[$ in (2.19), solve in $]-\infty,0[$ and $]L,\infty[$, and use the continuity of u at 0 and L to obtain a variational equation in $]0,L[$ that coincides with (2.25). Then the result follows by uniqueness of the solution.

Substitute $v \in H_0^1(-\infty,0)$ in (2.19) to obtain

$$\text{(A.3)} \qquad \begin{aligned} &-E(0)\frac{d^2u}{dx^2} + V(0)\frac{du}{dx} + R(0)u = c_0(x) \quad \text{in} \quad]-\infty,0[, \\ &u(-\infty) = 0, \qquad u(0^-) = u(0^+), \end{aligned}$$

where the boundary conditions arise from the fact that $u \in H^1(\mathbb{R})$ implies that $u(-\infty) = 0$ and u is continuous at $x = 0$. But system (A.3) has a unique solution given by

$$u(x) = u(0)e^{\beta_0^+ x} + \int_x^0 dz \int_{-\infty}^z dy \, \frac{c_0(y)}{E(0)} \, e^{\beta_0^+ (x-z)} e^{\beta_0^- (y-z)},$$

where

$$E(0)\beta_0^+ = \beta(0), \qquad E(0)\beta_0^- = \tfrac{1}{2}\left(\sqrt{V(0)^2 + 4E(0)R(0)} - V(0)\right).$$

In particular,

$$\text{(A.4)} \qquad E(0)\frac{du}{dx}(0^-) = \beta(0)u(0^+) - \int_{-\infty}^0 e^{\beta_0^- y} c_0(y)\, dy = \beta(0)u(0) - C_0.$$

Similarly, by choosing $v \in H_0^1(L, \infty)$ in (2.19) we obtain

$$\text{(A.5)} \qquad \begin{aligned} &-E(L)\frac{d^2u}{dx^2} + V(L)\frac{du}{dx} + R(L)u = c_0(x) \quad \text{in} \quad]L, \infty[, \\ &u(\infty) = 0, \quad u(L^+) = u(L^-). \end{aligned}$$

Its solution is given by

$$u(x) = u(L)e^{-\beta_L^- (x-L)} - \int_0^x dz \int_z^\infty dy \; e^{-\beta_L^- (x-z)} e^{-\beta_L^+ (y-z)} \frac{c_0(y)}{E(L)} \, dy,$$

where

$$E(L)\beta_L^- = \beta(L), \qquad E(L)\beta_L^+ = \tfrac{1}{2}\left(\sqrt{V(L)^2 + 4E(L)R(L)} + V(L)\right).$$

In particular,

$$\text{(A.6)} \quad E(L)\frac{du}{dx}(L^+) = -\beta(L)u(L^-) + \int_L^\infty e^{-\beta_L^+ (y-L)} c_0(y)\, dy = -\beta(L)u(L) + C_L.$$

The variational equation (2.19) is the sum of three terms

$$a + b + c = 0,$$

where

$$\begin{aligned} a &= \int_{-\infty}^0 E(0)\frac{du}{dx}\frac{dv}{dx} + V(0)\frac{du}{dx}v + R(0)uv - c^0(x)v \, dx, \\ b &= \int_0^L E(x)\frac{du}{dx}\frac{dv}{dx} + V(x)\frac{du}{dx}v + R(x)uv - c^0(x)v - \sum_{i=1}^N \frac{m_i}{A(x_i)} v(x_i), \\ c &= \int_L^\infty E(L)\frac{du}{dx}\frac{dv}{dx} + V(L)\frac{du}{dx}v + R(L)uv - c^0(x)v \, dx. \end{aligned}$$

In view of (A.3) and (A.5),

$$a = E(0)\frac{du}{dx}(0^-)v(0) = (\beta(0)u(0) - C_0)v(0),$$
$$c = -E(L)\frac{du}{dx}(L^+)v(L) = (\beta(L)u(L) - C_L)v(L),$$

and finally

$$0 = a + b + c = b(u,v) - \int_0^L c^0(x)v(x)\,dx - \sum_{i=1}^N \frac{m_i}{A(x_i)}v(x_i) - C_0 v(0) - C_L v(L),$$

which is precisely the variational equation (2.25). Hence the restriction to $[0,L]$ of the solution u of system (2.9) coincides with the solution of (2.25) on $[0,L]$. □

Proof of Theorem 3.1. Denote by u_i the solution of

$$\text{(A.7)} \qquad \begin{aligned} &V(x)\frac{du}{dx} + R(x)u = \delta(x - x_i), \qquad x > 0,\\ &u(0) = 0, \end{aligned}$$

where $\delta(x - x_i)$ is the Dirac delta function in $x = x_i$. Then

$$\text{(A.8)} \qquad u_i(x) = \begin{cases} 0, & 0 \le x < x_i,\\ \dfrac{1}{V(x_i)} e^{-\int_{x_i}^{x} (R(y)/V(y))\,dy}, & x_i \le x, \end{cases}$$

and the solution of (2.11)–(2.12) with $c^0 = 0$ and $u_0 = 0$ is given by

$$\text{(A.9)} \qquad u(x) = \sum_{i=1}^N \frac{m_i}{A(x_i)} u_i(x), \qquad x \ge 0.$$

We shall use the notation $\overrightarrow{m} \ge 0$ to say that for all i, $1 \le i \le N$, $m_i \ge 0$. Since $A(x) > 0$, $R(x) \ge 0$, and $V(x) > 0$, then for all $\overrightarrow{m} \ge 0$, $u(x) \ge 0$, $0 \le x \le L$. Now since u is decreasing on each interval $]x_{i-1}, x_i[$,

$$\forall x,\ 0 \le x \le L, \quad u(x) \ge u_P \Longleftrightarrow x_1 = 0, \quad u(x_i^-) \ge u_P > 0, \quad 2 \le i \le N+1.$$

Therefore,

$$0 = x_0 = x_1 \le x_2 \le \cdots \le x_N \le x_{N+1} = L.$$

Denote by $u \ge u_P$ the condition

$$\text{(A.10)} \qquad u(x_{i+1}^-) \ge u_P, \qquad 1 \le i \le N$$

and introduce the notation

$$Q_i = A(x_i)V(x_i), \quad e_i = e^{\int_{x_i}^{x_{i+1}} (R(y)/V(y))\,dy}, \quad 1 \le i \le N.$$

Given $\overrightarrow{m}$ the corresponding solution u of (2.11)–(2.12) must verify $u \ge u_P$, that is,

$$u(x_{i+1}^-) = \sum_{j=1}^i \frac{\widetilde{m}_j}{Q_j} \prod_{\ell=j}^i e_\ell^{-1} \ \ge\ u_P, \qquad 1 \le i \le N,$$

or in matrix form,

$$\Lambda \vec{m} \geq \vec{u}_P,$$

where $\vec{u}_P$ is the N-vector with components equal to u_P. The matrix Λ is positive, lower triangular, and invertible, and

$$\Lambda^{-1} = \begin{bmatrix} Q_1e_1 & 0 & 0 & \cdots & \cdots & 0 \\ -Q_2 & Q_2e_2 & 0 & \cdots & \cdots & 0 \\ 0 & -Q_3 & Q_3e_3 & \cdots & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & \ddots & 0 \\ \cdots & \cdots & \cdots & 0 & -Q_N & Q_Ne_N \end{bmatrix}.$$

Let $\vec{\tilde{m}}$ be the solution of the equation

$$\Lambda \vec{\tilde{m}} = \vec{u}_P$$

and let $\tilde{u}$ be the corresponding solution of (2.11)–(2.12). Then

$$\tilde{m}_1 = u_P Q_1 e_1,$$
$$\tilde{m}_i = u_P Q_i (e_i - 1), \qquad 2 \leq i \leq N$$

and

$$\tilde{u}(x_{i+1}^-) = u_P, \qquad 1 \leq i \leq N.$$

Moreover, since Λ is a positive matrix,

$$\vec{m} \geq \vec{\tilde{m}} \Rightarrow \Lambda(\vec{m} - \vec{\tilde{m}}) \geq 0 \Rightarrow u \geq u_P.$$

Now we claim that if condition (3.4) is verified, that is,

$$\text{(A.11)} \qquad Q_{i+1} - Q_i e_i \leq 0, \quad \forall i, \quad 1 \leq i \leq N,$$

then for any vector $\vec{m} \geq 0$ such that

$$\text{(A.12)} \qquad \exists i, \quad 0 \leq m_i < \tilde{m}_i,$$

we can construct another vector $\vec{m'}$ from $\vec{m}$ such that

$$\text{(A.13)} \qquad u'(x_j^-) \geq u_P, \quad 2 \leq j \leq N+1, \quad m'_j \geq \tilde{m}_j, \quad 1 \leq j \leq N,$$

and

$$\text{(A.14)} \qquad \sum_{j=1}^{N} m'_j \leq \sum_{j=1}^{N} m_j,$$

where u' is the solution of (2.11)–(2.12) corresponding to $\vec{m}'$. This means that to find the infimum, it is sufficient to consider vectors $\vec{m}$ such that $\vec{m} \geq \vec{\tilde{m}}$. Therefore

$$\operatorname*{Inf}_{\substack{\vec{m} \geq 0 \\ u \geq u_P}} \sum_{i=1}^{N} m_i = \operatorname*{Inf}_{\vec{m} \geq \vec{\tilde{m}}} \sum_{i=1}^{N} m_i$$

and the second Inf is achieved for $\vec{m} = \vec{\widetilde{m}}$.

To prove the previous technical result, rewrite the condition

$$\forall i, \quad 1 \le i \le N, \quad u(x_{i+1}^-) \ge u_P \tag{A.15}$$

in the form

$$u(x_{i+1}^-) - u_P = \sum_{j=1}^{i} \frac{\Delta m_j}{Q_j} \prod_{\ell=j}^{i} e_\ell^{-1} \ge 0, \qquad 1 \le i \le N, \tag{A.16}$$

where

$$\Delta m_j = m_j - \widetilde{m}_j, \qquad 1 \le j \le N. \tag{A.17}$$

For $i = 1$,

$$\frac{\Delta m_1}{Q_1} e_1^{-1} \ge 0 \implies \Delta m_1 \ge 0$$

and there is nothing to prove. Let k, $2 \le k \le N$, be the first index such that

$$\Delta m_k = m_k - \widetilde{m}_k < 0 \quad \Rightarrow \quad \forall j, \quad 1 \le j \le k-1, \quad m_j \ge \widetilde{m}_j.$$

Since $\vec{m}$ satisfies all the constraints for $i = k-1$ and k, we have

$$u(x_k^-) - u_P = \frac{\Delta m_{k-1}}{Q_{k-1}} e_{k-1}^{-1} \sum_{j=1}^{k-2} \frac{\Delta m_j}{Q_j} \prod_{\ell=j}^{k-1} e_\ell^{-1} \ge 0, \tag{A.18}$$

$$u(x_{k+1}^-) - u_P = e_k^{-1} \left\{ \frac{\Delta m_k}{Q_k} + \frac{\Delta m_{k-1}}{Q_{k-1}} e_{k-1}^{-1} + \sum_{j=1}^{k-2} \frac{\Delta m_j}{Q_j} \prod_{\ell=j}^{k-1} e_\ell^{-1} \right\} \ge 0. \tag{A.19}$$

Choose

$$\begin{aligned} m'_j &= \widetilde{m}_j, \qquad 1 \le j \le k-2 \quad \text{and} \quad j = k, \\ m'_j &= m_j, \qquad k+1 \le j \le N, \end{aligned} \tag{A.20}$$

and m'_{k-1} such that

$$e_{k-1}^{-1} \frac{\Delta m'_{k-1}}{Q_{k-1}} = \frac{\Delta m_k}{Q_k} + \frac{\Delta m_{k-1}}{Q_{k-1}} e_{k-1}^{-1} + \sum_{j=1}^{k-2} \frac{\Delta m_j}{Q_j} \prod_{\ell=j}^{k-1} e_\ell^{-1} \ge 0, \tag{A.21}$$

where

$$\Delta m'_j = m'_j - \widetilde{m}_j, \qquad 1 \le j \le N.$$

By construction for $\vec{m}'$, $m'_{k-1} \ge \widetilde{m}_{k-1}$ and

$$u'(x_{i+1}^-) - u_P = \sum_{j=1}^{i} \frac{\Delta m'_j}{Q_j} \prod_{\ell=j}^{i} e_\ell^{-1} = 0, \qquad 1 \le i \le k-2,$$

since $\Delta m'_j = 0$, $1 \leq j \leq k-2$. For $i = k-1$ and k, the terms reduce to

$$u'(x_k^-) - u_P = e_{k-1}^{-1}\frac{\Delta m'_{k-1}}{Q_{k-1}} = \left\{\frac{\Delta m_k}{Q_k} + \frac{\Delta m_{k-1}}{Q_{k-1}}e_{k-1}^{-1}\sum_{j=1}^{k-2}\frac{\Delta m_j}{Q_j}\prod_{\ell=j}^{k-1}e_\ell^{-1}\right\} \geq 0,$$

$$u'(x_{k+1}^-) - u_P = e_k^{-1}\left\{\frac{\Delta m'_k}{Q_k} + \frac{\Delta m'_{k-1}}{Q_{k-1}}e_{k-1}^{-1}\right\} = e_k^{-1}\frac{\Delta m'_{k-1}}{Q_{k-1}}e_{k-1}^{-1} = u(x_{k+1}^-) - u_P \geq 0$$

from (A.21) and (A.19). Finally, for $i > k$,

$$u'(x_{i+1}^-) - u_P = \sum_{j=1}^{i}\frac{\Delta m'_j}{Q_j}\prod_{\ell=j}^{i}e_\ell^{-1}$$

and since $\Delta m'_j = 0$ for $1 \leq j \leq k-2$, and $j = k$,

$$u'(x_{i+1}^-) - u_P = e_{k-1}^{-1}\,\frac{\Delta m'_{k-1}}{Q_{k-1}}\prod_{\ell=k}^{i}e_\ell^{-1} + \sum_{j=k+1}^{i}\frac{\Delta m'_j}{Q_j}\prod_{\ell=j}^{i}e_\ell^{-1}.$$

But from (A.20), $\Delta m'_j = \Delta m_j$, $k+1 \leq j \leq N$, and by construction of $\Delta m'_{k-1}$ in (A.21), we obtain

$$u'(x_{i+1}^-) - u_P = \sum_{j=1}^{i}\frac{\Delta m_j}{Q_j}\prod_{\ell=j}^{i}e_\ell^{-1} = u(x_{i+1}^-) - u_P \geq 0.$$

So the new vector $\overrightarrow{m}'$ verifies the constraint $u' \geq u_P$. Now consider the difference in costs:

$$\delta = \sum_{j=1}^{N}m'_j - \sum_{j=1}^{N}m_j = \sum_{j=1}^{k-2} -\Delta m_j + m'_{k-1} - m_{k-1} - \Delta m_k.$$

Again, by construction,

$$m'_{k-1} - m_{k-1} = Q_{k-1}e_{k-1}\left[\frac{\Delta m_k}{Q_k} + \sum_{j=1}^{k-2}\frac{\Delta m_j}{Q_j}\prod_{\ell=j}^{k-1}e_\ell^{-1}\right]$$

and

$$\begin{aligned}\delta &= [Q_{k-1}e_{k-1} - Q_k]\frac{\Delta m_k}{Q_k} + \sum_{j=1}^{k-2}\left[Q_{k-1}\prod_{\ell=j}^{k-2}e_\ell^{-1} - Q_j\right]\frac{\Delta m_j}{Q_j}\\ &= [Q_{k-1}e_{k-1} - Q_k]\frac{\Delta m_k}{Q_k} - \sum_{j=1}^{k-2}\sum_{n=j}^{k-2}[Q_n e_n - Q_{n+1}]\prod_{\ell=j}^{n}e_\ell^{-1}\frac{\Delta m_j}{Q_j}.\end{aligned}$$

Always by construction

$$\Delta m_k < 0 \quad \text{and} \quad \Delta m_j \geq 0, \quad 1 \leq j \leq k-1,$$

and by hypothesis (A.11)

$$Q_i e_i - Q_{i+1} \geq 0, \qquad 1 \leq i \leq k-1.$$

Therefore, $\delta \leq 0$ and we have constructed a new vector $\overrightarrow{m'}$ from $\overrightarrow{m}$ that verifies the constraints

$$u'(x_j^-) \geq u_P, \quad 2 \leq j \leq N+1, \quad m'_j \geq \tilde{m}_j, \quad 1 \leq j \leq k$$

and does not increase the cost. We can now repeat the construction for the next index $k' > k$ such that $\Delta m_{k'} < 0$ up to N. This proves the technical result.

We have established that under assumption (A.11), $\overrightarrow{\tilde{m}}$ is a minimizing solution. Now it is easy to see that if assumption (A.11) is verified with a strict inequality, the points x_i's verify the conditions

$$0 = x_1 < x_2 < \cdots < x_{N-1} < x_N < L$$

and the previous constructions yield a $\overrightarrow{m'}$ for which the cost is strictly less than the one for the minimizing vector $\overrightarrow{m}$. This contradiction shows that the set of solutions of the two problems

$$\operatorname*{Inf}_{\substack{\vec{m} \geq 0 \\ u \geq u_P}} \sum_{i=1}^{N} m_i \quad \text{and} \quad \operatorname*{Inf}_{\vec{m} \geq \vec{\tilde{m}}} \sum_{i=1}^{N} m_i$$

coincide. But for the second problem, $\overrightarrow{\tilde{m}}$ is obviously the unique minimizing element. This completes the proof of Theorem 3.1. □

Proof of Theorem 3.2. We have seen from Theorem 3.1 that for fixed x_i's the optimal cost is given by

$$j(\vec{x}) = u_P \left\{ Q(0) + \sum_{i=1}^{N} Q(x_i)[e_i - 1] \right\}$$

where

$$a(y) = \frac{R(y)}{V(y)}, \quad e_i = e^{\int_{x_i}^{x_{i+1}} a(y)\, dy}, \quad \text{and} \quad x_1 = 0.$$

Moreover, from assumption (3.8), inequalities (3.4) are verified for any fixed set of x_i's. So the next step is to minimize this cost function over the following nonempty compact subset of $\mathbb{R}^N$:

$$K = \{\vec{x} : 0 = x_1 \leq x_2 \leq \cdots \leq x_N \leq x_{N+1} = L\}.$$

For A and V in $C([0, L])$, $j(\vec{x})$ is continuous on K and there exists a minimizing solution $\vec{x}$ in K. This solution can be characterized by using Lagrange multipliers

$$L(\vec{x}, \vec{\lambda}) = \sum_{i=1}^{N} Q(x_i) \left[e^{\int_{x_i}^{x_{i+1}} a(y)\, dy} - 1 \right] + \sum_{i=1}^{N} \lambda_i (x_i - x_{i+1}).$$

This yields

$$\lambda_i(x_i - x_{i+1}) = 0, \quad \lambda_i \geq 0, \quad x_i - x_{i+1} \leq 0, \quad 1 \leq i \leq N,$$
$$Q_i'[e_i - 1] + a_i[Q_{i-1}e_{i-1} - Q_ie_i] + \lambda_i - \lambda_{i-1} = 0, \qquad 2 \leq i \leq N,$$

where

$$Q_i = Q(x_i), \quad Q_i' = Q'(x_i), \quad a_i = a(x_i).$$

It is convenient to rewrite the above system as follows:

$$\text{(A.22)} \qquad \lambda_i(x_i - x_{i+1}) = 0, \quad \lambda_i \geq 0, \quad x_i - x_{i+1} \leq 0, \quad 1 \leq i \leq N,$$

$$\text{(A.23)} \quad [Q_i' - a_iQ_i](e_i - 1) + a_i[Q_{i-1}e_{i-1} - Q_i] + \lambda_i - \lambda_{i-1} = 0, \qquad 2 \leq i \leq N.$$

Assume that $x_N = x_{N+1}$. Then $e_N = 1$ and

$$a_N[Q_{N-1}e_{N-1} - Q_N] + \lambda_N - \lambda_{N-1} = 0.$$

If $x_N > x_{N-1}$, then $\lambda_{N-1} = 0$ and by hypothesis (3.8),

$$0 < a_N[Q_{N-1}e_{N-1} - Q_N] = \lambda_{N-1} - \lambda_N = -\lambda_N \leq 0,$$

which is a contradiction. Therefore, $x_{N-1} - x_N = 0$ and

$$a_{N-1}[Q_{N-2}e_{N-2} - Q_{N-1}] + \lambda_{N-1} - \lambda_{N-2} = 0.$$

If $x_{N-1} > x_{N-2}$, then $\lambda_{N-2} = 0$ and by hypothesis (3.8),

$$0 < a_{N-1}[Q_{N-2}e_{N-2} - Q_{N-1}] = \lambda_{N-2} - \lambda_{N-1} = -\lambda_{N-1} \leq 0,$$

which again is a contradiction. By repeating this argument we finally obtain

$$x_i = x_{N+1}, \qquad 1 \leq i \leq N.$$

But

$$x_1 = 0 < L = x_{N+1}$$

and this is a contradiction. This proves that

$$x_N - x_{N+1} < 0, \quad \lambda_N = 0, \quad \text{and} \quad e_N > 1.$$

Now if $x_{N-1} = x_N$, then $e_{N-1} = 1$ and

$$a_{N-1}[Q_{N-2}e_{N-2} - Q_{N-1}] + \lambda_{N-1} - \lambda_{N-2} = 0,$$

and by the same argument as above we obtain

$$x_i = x_N, \qquad 1 \leq i \leq N \implies x_N = 0.$$

This means that the resulting cost is

$$j_0 = u_P\left\{Q(0) + Q(0)\left[e^{\int_0^L a(y)\,dy} - 1\right]\right\}.$$

For $N \geq 2$ this is not a minimum since we can choose

$$x'_{N-1} = 0 < x'_N < x_{N+1} = L$$

and

$$j_1 = u_P \left\{ Q(0) + Q(0)\left[e^{\int_0^{x'_N} a(y)\,dy} - 1 \right] + Q(x'_N)\left[e^{\int_{x'_N}^{L} a(y)\,dy} - 1 \right] \right\} < j_0.$$

To see this,

$$\begin{aligned} \frac{j_0 - j_1}{u_P} &= Q(0)\left[e^{\int_0^L a(y)\,dy} - e^{\int_0^{x'_N} a(y)\,dy} \right] - Q(x'_N)\left[e^{\int_{x'_N}^{L} a(y)\,dy} - 1 \right] \\ &= Q(0)e^{\int_0^{x'_N} a(y)\,dy}\left[e^{\int_{x'_N}^{L} a(y)\,dy} - 1 \right] - Q(x'_N)\left[e^{\int_{x'_N}^{L} a(y)\,dy} - 1 \right] \\ &= \left[Q(0)e^{\int_0^{x'_N} a(y)\,dy} - Q(x'_N) \right]\left[e^{\int_{x'_N}^{L} a(y)\,dy} - 1 \right] > 0, \end{aligned}$$

since both terms are strictly positive. From this contradiction we conclude that $x_{N-1} - x_N < 0$.

By repeating this argument we find that

$$\text{(A.24)} \qquad \forall i, \quad 1 \leq i \leq N, \quad x_i - x_{i+1} < 0, \quad \Rightarrow \quad \lambda_i = 0.$$

Thus the constraints are not active and

$$Q_{i-1}[e_{i-1} - 1] = \frac{[a_i Q_i - Q'_i]}{a_i Q_i} Q_i(e_i - 1) + Q_i - Q_{i-1}, \qquad 2 \leq i \leq N.$$

The last equation is

$$\text{(A.25)} \qquad \prod_{i=1}^{N} e^{\int_{x_i}^{x_i+1} a(y)\,dy} = e^{\int_0^L a(y)\,dy}.$$

Finally, in view of (A.24) and assumption (3.8), the inequalities (3.4) are verified with strict inequalities. As a result, for a minimizing set of positions x_i's, the optimal distribution of masses is unique and given by $\overrightarrow{m}$ in (3.5). □

Proof of Theorem 3.3. By construction, the solution

$$u(x) = u_P, \quad 0 \leq x \leq L, \quad u_L = u_P$$

generates a positive measure (see (3.24)) for $R(x) \geq 0$, $A(x) \geq 0$, and $V(0) \geq 0$. We only need to show that it is minimal. For all $m \geq 0$ in $C([0, L])'$ such that $u \geq u_P$ and $u_L \geq u_P$ we have

$$\begin{aligned} M(m) = b_0((u, u_L), A) &= b_0((u - u_P, u_L - u_P), A) + b_0((u_P, u_P), A) \\ &\geq b_0((u_P, u_P), A) \end{aligned}$$

by condition (3.25), since $u-u_P \geq 0$ and $u_L-u_P \geq 0$. Hence (u_P, u_P) is a minimizing solution. It is obvious that when the inequality (3.25) is strict for $(u, u_L) \neq (0,0)$, then this solution is unique. Moreover, from (3.25) with $u=0$,

$$V(L)A(L)u_L > 0, \quad \forall u_L > 0 \quad \Rightarrow \quad Q(L) = V(L)A(L) > 0.$$

Similarly, for all $u \geq 0$ in $L^2(0,L)$, $u \neq 0$, with $u_L = 0$,

$$\int_0^L \left[-\frac{d}{d\xi}(VA) + RA\right] u\, d\xi > 0.$$

In particular, fix x and y such that $0 \leq x < y \leq L$ and define

$$u(\xi) = \begin{cases} e^{-\int_x^\xi (R(z)/V(z))\,dz}, & \xi \in [x,y], \\ 0, & \text{otherwise.} \end{cases}$$

Then

$$\int_x^y -\frac{d}{d\xi}\left[Q(\xi)e^{-\int_x^\xi (R(z)/V(z))\,dz}\right] d\xi > 0,$$

which implies that

$$Q(y) < Q(x)e^{-\int_x^y (R(z)/V(z))\,dz}.$$

But this is precisely condition (3.8). □

Proof of Corollary 1. The equivalence of (i), (ii), and (iii) is obvious. When (iii) is verified, multiply the left-hand side by the strictly positive function

$$e(\xi) = \exp^{-\int_0^\xi (R(z)/V(z))\,dz}.$$

Then the right-hand side of (iii) becomes

$$-\frac{d}{d\xi}[Q(\xi)e(\xi)] \geq 0$$

and by integrating from x to y we recover (iv). Conversely, from (iv) for all $x < y$,

$$\frac{Q(y)e(y) - Q(x)e(x)}{y-x} \leq 0 \quad \Rightarrow \quad \frac{d}{dx}[Q(x)e(x)] \leq 0$$

and we obtain (iii). □

Proof of Corollary 2. (i) ⇒ (ii). From Corollary 1 introduce the function

$$f = -\frac{d}{dx}(VA) + RA \geq 0.$$

Denote by Z the set $\{x \in [0,L] : f(x) = 0\}$. If $\text{meas}(Z) > 0$, choose $u = \chi_Z$, the characteristic function of Z, and $u_L = 0$. Then since $u \neq 0$,

$$0 < \int_0^L \chi_Z f\, dx = \int_Z f\, dx = 0,$$

which contradicts the fact that $\text{meas}(Z) > 0$. Hence $f(x) > 0$ almost everywhere in $[0,L]$. Now for $u = 0$, we get from (i)

$$\forall\, u_L > 0, \quad u_L V(L)A(L) > 0 \quad \Rightarrow \quad V(L)A(L) > 0.$$

The converse, (ii) ⇒ (i), and the equivalence of (ii) and (iii), are obvious. Clearly (iii) implies (iv) by the same technique as in Corollary 1 with a strict inequality. □

Proof of Theorem 3.4. The technique is the same as in the proof of Theorem 3.3. First note that the solution u_P generates a positive measure m_P given by (3.35) for $R(x) \geq 0$, $A(x) \geq 0$. Then for all $m \geq 0$ in $C([0,L])'$ such that $u \geq u_P$, we have

$$M(m) = b(u,A) = b(u-u_L, A) + b(u_L, A) \geq b(u_L, A)$$

by condition (3.36) since $u - u_L \geq 0$. Hence u_L is a minimizing solution. It is clearly unique when inequality (3.36) is strict for all $u \neq 0$. □

A.2. Discretization of the state equation (4.15). We compute the algorithm (4.17)–(4.19) from (4.15) under the conditions (4.14). We introduce the notation

$$\begin{aligned}&p_0 = p(\xi_0^+), \quad p_j = p(\xi_j^-), \quad 1 \le j \le J,\\ &u_j = \text{ value of } u(x) \text{ on the interval } I_j =]\xi_{j-1}, \xi_j[, \qquad 1 \le j \le J.\end{aligned}$$

By setting $q = 0$ and $w = 0$ we obtain from (4.15)

$$\begin{aligned}&-p_0 + \beta_0 U_0 = c_0, \qquad p_J + \beta_J U_J = c_J,\\ &p(\xi_j^+) = p_j - c_j, \qquad 1 \le j \le J-1.\end{aligned} \tag{A.26}$$

Then set $w = 0$ and $W = 0$ and use the trapeze formula to evaluate the integrals on each interval I_j:

$$\begin{aligned}&\frac{h}{2}\left[E_j^{-1}p(\xi_j^-)q(\xi_j^-) + E_{j-1}^{-1}p(\xi_{j-1}^+)q(\xi_{j-1}^+)\right]\\ &\qquad = -u_j\left[q(\xi_j^-) - q(\xi_{j-1}^+)\right] + U_j q(\xi_j^-) - U_{j-1}q(\xi_{j-1}^+)\end{aligned}$$

or simply

$$\left.\begin{aligned}&\frac{h}{2}E_j^{-1}p_j + u_j - U_j = 0,\\ &\frac{h}{2}E_{j-1}^{-1}p(\xi_{j-1}^+) - u_j + U_{j-1} = 0,\end{aligned}\right. \qquad 1 \le j \le J. \tag{A.27}$$

Now set $q = 0$ and $W = 0$ to obtain on I_j

$$-[p(\xi_j^-) - p(\xi_{j-1}^+)] + hR_j u_j - u_j[V_j - V_{j-1}] + u_j V_j - u_{j-1}V_{j-1} = hf_j$$

or

$$\begin{aligned}&-[p_1 - p_0] + hR_1 u_1 + V_0[u_1 - u_0] = hf_1,\\ &-[p_j - p_{j-1} + c_{j-1}] + hR_j u_j + V_{j-1}[u_j - u_{j-1}] = hf_j, \qquad 2 \le j \le J.\end{aligned} \tag{A.28}$$

Now use the first equation of (A.27), to eliminate $U_1, \cdots, U_{J-1}$ in the second one:

$$\begin{aligned}&\frac{h}{E_j}p_j + u_j - u_{j+1} = \frac{h}{2E_j}c_j, \qquad 1 \le j \le J-1,\\ &\frac{h}{2E_0}p_0 + U_0 - u_1 = 0.\end{aligned} \tag{A.29}$$

The U_j's can be recovered through the formula

$$U_j = \frac{u_j + u_{j+1}}{2} + \frac{h}{4E_j}c_j, \qquad 1 \le j \le J-1.$$

Now use equations (A.29) and the first equation (A.27) for $j = J$ to eliminate the p_j's in (A.28) :

$$\begin{aligned}&p_j = \frac{E_j}{h}(u_{j+1} - u_j) + \frac{c_j}{2}, \qquad 1 \le j \le J-1,\\ &p_0 = \frac{2E_0}{h}(u_1 - U_0), \qquad p_J = \frac{2E_J}{h}(U_J - u_J)\end{aligned} \tag{A.30}$$

to obtain

(A.31)

$$
\begin{aligned}
&-\left[\frac{2E_0}{h}+V_0\right]U_0+\left[\frac{E_1+2E_0}{h}+hR_1+V_0\right]u_1-\frac{E_1}{h}u_2=\frac{c_1}{2}+hf_1, \qquad j=1,\\
&\begin{cases} -\left[\dfrac{E_{j-1}}{h}+V_{j-1}\right]u_{j-1}+\left[\dfrac{E_j+E_{j-1}}{h}+hR_j+V_{j-1}\right]u_j-\dfrac{E_j}{h}u_{j+1}\\ =\dfrac{c_j+c_{j-1}}{2}+hf_j, \qquad 2\le j\le J-1,\end{cases}\\
&\begin{cases} -\left[\dfrac{E_{J-1}}{h}+V_{J-1}\right]u_{J-1}+\left[\dfrac{2E_J+E_{J-1}}{h}+hR_J+V_{J-1}\right]u_J-\dfrac{2E_J}{h}U_J\\ =\dfrac{c_{J-1}}{2}+hf_J, \qquad j=J.\end{cases}
\end{aligned}
$$

Finally, also eliminate p_0 and p_J in the first two equations of (A.26) to obtain

$$
\text{(A.32)}\qquad \left(\frac{2E_0}{h}+\beta_0\right)U_0-\frac{2E_0}{h}u_1=c_0, \qquad -\frac{2E_J}{h}u_J+\left(\frac{2E_J}{h}+\beta_J\right)U_J=c_J.
$$

System (A.31)–(A.32) is equivalent to a system of $(J+2)$ equations in the $(J+2)$ variables $(U_0, u_1, \cdots, u_J, U_J)$. This is precisely system (4.16), where the $(J+2)\times(J+1)$ matrix C is given by

$$
C=\begin{bmatrix}
1 & 0 & 0 & 0 & \cdots & \cdots & 0\\
0 & \frac12 & 0 & 0 & \cdots & \cdots & 0\\
0 & \frac12 & \frac12 & 0 & \cdots & \cdots & 0\\
0 & 0 & \ddots & \ddots & \ddots & \cdots & 0\\
\vdots & \vdots & \ddots & \ddots & \ddots & 0 & \vdots\\
0 & \cdots & \cdots & \cdots & \frac12 & \frac12 & 0\\
0 & \cdots & \cdots & \cdots & 0 & \frac12 & 0\\
0 & \cdots & \cdots & \cdots & 0 & 0 & 1
\end{bmatrix}.
$$

A.3. Computation of the adjoint system (4.36). With the notation (4.32)–(4.33) of §4.2.2 the Lagrangian $L=L(\overrightarrow{m},\vec{x},Y,P)$ is

$$
\begin{aligned}
L=\sum_{i=1}^{N}m_i+\frac{1}{2\epsilon_1}([m_i]^-)^2+\frac{1}{2\epsilon_2}\sum_{j=0}^{J}([u_P-U_j]^+)^2+\frac{1}{2\epsilon_3}\int_0^L([u_P-u(x)]^+)^2\,dx\\
+\sum_{j=1}^{J}\Bigg\{\int_{I_j}\left[E^{-1}(x)pq+u\frac{dq}{dx}\right]dx-U_jq(\xi_j^-)+U_{j-1}q(\xi_{j-1}^+)\\
+\int_{I_j}\left[-\frac{dp}{dx}+Ru\right]w-u\frac{d}{dx}[V(x)w]\,dx\\
+u(\xi_j^-)V(\xi_j)w(\xi_j^-)-u(\xi_{j-1}^-)V(\xi_{j-1})w(\xi_{j-1}^+)\Bigg\}
\end{aligned}
$$

$$+\sum_{j=1}^{J-1}\left[p(\xi_j^-)-p(\xi_j^+)-\sum_{i=1}^{N}\frac{m_i}{A(x_i)}B\left(\frac{\xi_j-x_i}{h}\right)\right]W_j$$
$$+\left[-p(0^+)+\beta_0U_0-\sum_{i=1}^{N}\frac{m_i}{A(x_i)}B\left(\frac{-x_i}{h}\right)\right]W_0$$
$$+\left[p(\xi_J^-)+\beta_JU_J-\sum_{i=1}^{N}\frac{m_i}{A(x_i)}B\left(\frac{L-x_i}{h}\right)\right]W_J,$$

where it should be recalled that we have defined $u(0^-)=U_0$. We successively compute the partial derivatives with respect to U_j, p, and u at the point $(\{\hat{U}_j\}$, $\hat{u}$, $\hat{p}$, $\hat{W}_j$, $\hat{w}$, $\hat{q})$:

$$\left[-\frac{1}{\epsilon_2}[u_P-\hat{U}_0]^+ +\hat{q}(\xi_0^+)-V(0)\hat{w}(\xi_0^+)+\beta_0\hat{W}_0\right]U_0=0, \tag{A.33}$$

$$\left[-\frac{1}{\epsilon_2}[u_P-\hat{U}_j]^+ -\hat{q}(\xi_j^-)+\hat{q}(\xi_j^+)\right]U_j=0, \qquad 1\le j\le J-1, \tag{A.34}$$

$$\left[-\frac{1}{\epsilon_2}[u_P-\hat{U}_J]^+ -\hat{q}(\xi_J^-)+\beta_J\hat{W}_J\right]U_J=0 \tag{A.35}$$

for all U_j, $0\le j\le J$;

$$\int_{I_1}\left[E^{-1}(x)p\hat{q}-\frac{dp}{dx}\hat{w}\right]dx+p(\xi_1^-)\hat{W}_1-p(0^+)\hat{W}_0=0, \tag{A.36}$$

$$\int_{I_j}\left[E^{-1}(x)p\hat{q}-\frac{dp}{dx}\hat{w}\right]dx+p(\xi_j^-)\hat{W}_j-p(\xi_{j-1}^+)\hat{W}_{j-1}=0, \qquad 2\le j\le J-1, \tag{A.37}$$

$$\int_{I_J}\left[E^{-1}(x)p\hat{q}-\frac{dp}{dx}\hat{w}\right]dx+p(\xi_J^-)\hat{W}_J-p(\xi_{J-1}^+)\hat{W}_{J-1}=0 \tag{A.38}$$

for all p in $P^1(I_j)$; and finally,

$$\begin{aligned}&\int_{I_1}\left\{-\frac{1}{\epsilon_3}[u_P-\hat{u}(x)]^+ +\frac{d\hat{q}}{dx}+R\hat{w}-\frac{d}{dx}[V(x)\hat{w}]\right\}u\,dx\\&+V(\xi_1)\hat{w}(\xi_1^-)u(\xi_1^-)=0,\end{aligned} \tag{A.39}$$

$$\begin{aligned}&\int_{I_j}\left\{-\frac{1}{\epsilon_3}[u_P-\hat{u}(x)]^+ +\frac{d\hat{q}}{dx}+R\hat{w}-\frac{d}{dx}[V(x)\hat{w}]\right\}u\,dx\\&+V(\xi_j)[\hat{w}(\xi_j^-)-\hat{w}(\xi_j^+)]u(\xi_j^-)=0, \qquad 2\le j\le J-1,\end{aligned} \tag{A.40}$$

$$\begin{aligned}&\int_{I_J}\left\{-\frac{1}{\epsilon_3}[u_P-\hat{u}(x)]^+ +\frac{d\hat{q}}{dx}+R\hat{w}-\frac{d}{dx}[V(x)\hat{w}]\right\}u\,dx\\&+V(\xi_J)\hat{w}(\xi_J^-)u(\xi_J^-)=0,\end{aligned} \tag{A.41}$$

for all u in $P^0(I_j)$.

The piecewise constant functions $\hat{u}$ and $\hat{w}$ will be represented by the sequences $\{\hat{u}_j\ :\ 1\le j\le N\}$ and $\{\hat{w}_j\ :\ 1\le j\le N\}$ where $\hat{u}_j$ and $\hat{w}_j$ in $\mathbb{R}$ are the constant values of $\hat{u}$ and $\hat{w}$ on the interval I_j. In addition, we introduce the variables

$$\begin{aligned}
&\tilde{q}_J = \hat{q}(\xi_J^-), \quad \tilde{q}_j = \hat{q}(\xi_j^+), \quad 0 \le j \le J-1,\\
&d_j = \frac{1}{\epsilon_2}[u_P - \hat{U}_j]^+, \qquad 0 \le j \le J,\\
&g_j = \frac{1}{\epsilon_3 h}\int_{I_j}[u_P - \hat{u}(x)]^+ dx = \frac{1}{\epsilon_3}[u_P - \hat{u}_j]^+, \qquad 1 \le j \le J.
\end{aligned} \tag{A.42}$$

From (A.33)–(A.35),

$$\begin{aligned}
&\tilde{q}_0 + \beta_0 \hat{W}_0 - V_0 \hat{w}_1 = d_0\\
&\hat{q}(\xi_j^-) = \tilde{q}_j - d_j, \qquad 1 \le j \le J-1\\
&\tilde{q}_J = \beta_J \hat{W}_J - d_J.
\end{aligned} \tag{A.43}$$

From (A.36)–(A.38) with the trapeze formula for integrals,

$$\begin{aligned}
&\frac{h}{2E_0}\tilde{q}_0 + \hat{w}_1 - \hat{W}_0 = 0,\\
&\frac{h}{2E_1}(\tilde{q}_1 - d_1) - \hat{w}_1 + \hat{W}_1 = 0,
\end{aligned} \tag{A.44}$$

$$\left.\begin{aligned}
&\frac{h}{2E_{j-1}}\tilde{q}_{j-1} + \hat{w}_j - \hat{W}_{j-1} = 0,\\
&\frac{h}{2E_j}(\tilde{q}_j - d_j) - \hat{w}_j + \hat{W}_j = 0,
\end{aligned}\right. \qquad 2 \le j \le J-1, \tag{A.45}$$

$$\begin{aligned}
&\frac{h}{2E_{J-1}}\tilde{q}_{J-1} + \hat{w}_J - \hat{W}_{J-1} = 0,\\
&\frac{h}{2E_J}\tilde{q}_J - \hat{w}_J + \hat{W}_J = 0.
\end{aligned} \tag{A.46}$$

From (A.39)–(A.41), using the trapeze formula and the second identity of (A.43),

$$\begin{aligned}
&\tilde{q}_1 - \tilde{q}_0 + (hR_1 + V_0)\hat{w}_1 = hg_1 + d_1,\\
&\tilde{q}_j - \tilde{q}_{j-1} + (hR_j + V_{j-1})\hat{w}_j - V_j \hat{w}_{j+1} = hg_j + d_j, \qquad 2 \le j \le J-1,\\
&\tilde{q}_J - \tilde{q}_{J-1} + (hR_J + V_{J-1})\hat{w}_J = hg_J.
\end{aligned} \tag{A.47}$$

Now from system (A.44)–(A.46) we can express $\hat{W}_j$, $1 \le j \le J-1$, and the $\tilde{q}_j$'s in terms of the $\hat{w}_j$'s and $\hat{W}_0$ and $\hat{W}_J$:

$$\begin{aligned}
&\tilde{q}_0 = \frac{2E_0}{h}[\hat{W}_0 - \hat{w}_1],\\
&\tilde{q}_j = \frac{E_j}{h}[\hat{w}_j - \hat{w}_{j+1}] + \frac{d_j}{2}, \qquad 1 \le j \le J-1,\\
&\tilde{q}_J = \frac{2E_J}{h}[\hat{w}_J - \hat{W}_J],
\end{aligned} \tag{A.48}$$

$$\hat{W}_j = \frac{\hat{w}_j + \hat{w}_{j+1}}{2} + \frac{h}{4E_j}d_j, \qquad 1 \le j \le J-1. \tag{A.49}$$

Its substitution in (A.47) yields

$$-\frac{2E_0}{h}\hat{W}_0+\left[\frac{E_1+2E_0}{h}+hR_1+V_0\right]\hat{w}_1-\frac{E_1}{h}\hat{w}_2=hg_1+\frac{d_1}{2}, \tag{A.50}$$

$$\begin{aligned}&-\frac{E_{j-1}}{h}\hat{w}_{j-1}+\left[\frac{E_j+E_{j-1}}{h}+hR_j+V_{j-1}\right]\hat{w}_j-\left[\frac{E_j}{h}+V_j\right]\hat{w}_{j+1}\\&=hg_j+\frac{d_j+d_{j-1}}{2},\qquad 2\le j\le J-1,\end{aligned} \tag{A.51}$$

$$-\frac{E_{J-1}}{h}\hat{w}_{J-1}+\left[2\frac{E_J+E_{J-1}}{h}+hR_J+V_{J-1}\right]\hat{w}_J-\frac{2E_J}{h}\hat{W}_J=hg_J+\frac{d_{J-1}}{2}. \tag{A.52}$$

By adding the two boundary equations from (A.43) to the above system,

$$\left[\frac{2E_0}{h}+\beta_0\right]\hat{W}_0-\left[\frac{2E_0}{h}+V_0\right]\hat{w}_1=d_0, \tag{A.53}$$

$$-\frac{2E_J}{h}\hat{w}_J+\left[\frac{2E_J}{h}+\beta_J\right]\hat{W}_J=d_J. \tag{A.54}$$

Equations (A.50)–(A.54) yield (4.36) and equations (A.48)–(A.49) coincide with (4.38)–(4.39). The derivatives (4.40) and (4.41) are now readily obtained by taking partial derivatives of the Lagrangian L. This completes the argument. □

Acknowledgment. The authors are very grateful to Professor Michel Fortin at Université Laval, Québec, Canada for enlightening discussions on the numerical solution of the diffusion-transport equation.

REFERENCES

[1] R. Aris, *On the dispersion of a solute in a fluid flowing through a tube*, Proc. Soc. Ser. A, 235 (1956), pp. 66–77.

[2] A. Chalifour, J. Boisvert, and C. Back, *Optimization of insecticide treatments in rivers: An application of graph theory for planning a black fly larvae control program*, Canad. J. Fish. Aquat. Sci., 47 (1990), pp. 2049–2056.

[3] A. Chalifour and M. C. Delfour, *Optimisation des épandages insecticides en eaux courantes*, in Proc. First Internat. Workshop on Sensors and Actuators in Distributed Parameter Systems, A. El Jai and M. Amouroux, eds., ISGMP, Université de Perpignan, Perpignan, France, 1987, pp. 9–30.

[4] R. Correa and A. Seeger, *Directional derivatives of a minimax function*, Nonlinear Anal., Theory, Methods, Appl., 9 (1985), pp. 13–22.

[5] M. C. Delfour and J. P. Zolésio, *Shape sensitivity analysis via MinMax differentiability*, SIAM J. Control Optim., 26 (1988), pp. 834–862.

[6] ———, *Velocity method and Lagrangian formulation for the computation of the shape Hessian*, SIAM J. Control Optim., 29 (1991), pp. 1414–1442.

[7] G. Dhatt, A. Soulaimani, A. Ouellet, and M. Fortin, *Development of new triangular elements for free surface flows*, Internat. J. Numer. Methods Fluids, 6 (1986), pp. 895–911.

[8] P. Guillet, H. Escaffre, J. M. Prud'hom, and S. Bakayoko, *Etude des facteurs conditionnant l'efficacité des préparations à base de Bacillus thuringiensis $H14$ vis-à-vis des larves du complexe Simulium damnosum (Diptera, Simuliidae)* (1), (2), Cahiers ORSTOM, Ser. Ent. Méd. et Parasitol. XXIII, 4 (1985), pp. 257–264, 265–271.

[9] P. Guillet, J. M. Hougard, J. Doannio, H. Escaffre, and J. Duval, *Evaluation de la sensitivité des larves du complexe Simulium damnosum à la toxine de Bacillus thuringiensis* $H14$ (1), (2), Cahiers ORSTOM, Ser. Ent. Méd. et Parasitol. XXIII, 4 (1985), pp. 241–250, 251–255.

[10] P. Khalig, *One-dimensional transient model for short-term prediction of downstream pollution in rivers*, Water Res., 13 (1978), pp. 1311–1316.

[11] P. Lesaint and P.-A. Raviart, *On a finite element method for solving the neutron transport equation*, in Mathematical Aspects of Finite Elements in Partial Differential Equations, C. de Boor, ed., Academic Press, New York, 1974, pp. 89–123.

[12] G. I. Marchuk, *Mathematical Models in Environmental Problems*, North–Holland, Amsterdam, New York, Oxford, Tokyo, 1986.

[13] OCP, *Ten years of onchocerciasis control in West Africa, onchocerciasis control programme*, Report OCP/GVA 85.1B, World Health Organization, Geneva, 1985.

[14] P.-A. Raviart, *Les méthodes d'éléments finis en mécanique des fluides*, in Collection de la direction des études et recherches d'électricité de France, Editions Eyrolles, Paris, 1981.

[15] G. I. Taylor, *Dispersion matter in turbulent flow slowly through a tube*, Proceedings, Royal Society of London Ser. A, 223 (1954), pp. 446–468.

SIAM J. OPTIMIZATION
Vol. 2, No. 2, pp. 304–324, May 1992

006

ON THE SUPERLINEAR AND QUADRATIC CONVERGENCE OF PRIMAL-DUAL INTERIOR POINT LINEAR PROGRAMMING ALGORITHMS*

YIN ZHANG[†], RICHARD A. TAPIA[‡], AND JOHN E. DENNIS, JR.[‡]

Abstract. This paper presents a convergence rate analysis for interior point primal-dual linear programming algorithms. Conditions that guarantee Q-superlinear convergence are identified in two distinct theories. Both state that, under appropriate assumptions, Q-superlinear convergence is achieved by asymptotically taking the step to the boundary of the positive orthant and letting the barrier parameter approach zero at a rate that is superlinearly faster than the convergence of the duality gap to zero. The first theory makes no nondegeneracy assumption and explains why in recent numerical experimentation Q-superlinear convergence was always observed. The second theory requires the restrictive assumption of primal nondegeneracy. However, it gives the surprising result that Q-superlinear convergence can still be attained even if centering is not phased out, provided the iterates asymptotically approach the central path. The latter theory is extended to produce a satisfactory Q-quadratic convergence theory. It requires that the step approach the boundary as fast as the duality gap approaches zero and the barrier parameter approach zero as fast as the square of the duality gap approaches zero.

Key words. linear programming, primal-dual interior point algorithms, Newton's method, Q-superlinear and Q-quadratic convergence

AMS(MOS) subject classifications. 65K05, 90C05

1. Introduction.

This paper considers linear programs in the standard form:

$$\text{(1.1)}\qquad \begin{array}{ll} \text{minimize} & c^T x \\ \text{subject to} & Ax = b, \qquad x \geq 0, \end{array}$$

where $c, x \in \mathbf{R}^n$, $b \in \mathbf{R}^m$, $A \in \mathbf{R}^{m\times n} (m < n)$, and A has full rank m. The dual linear program of (1.1) can be expressed in the following symmetric form:

$$\text{(1.2)}\qquad \begin{array}{ll} \text{minimize} & d^T y \\ \text{subject to} & By = Bc, \qquad y \geq 0, \end{array}$$

where $y \in \mathbf{R}^n$ is the vector of dual slack variables, $d = A^T(AA^T)^{-1}b$, $B \in \mathbf{R}^{(n-m)\times n}$ has full row rank, and $AB^T = 0$ (i.e., the columns of B^T form a basis for the null space of A). This form of the dual was introduced by Todd and Ye in [21]. A pair (x, y) is called strictly feasible if x and y are feasible for (1.1) and (1.2), respectively, and are positive as well.

The weak duality theorem says that the duality gap $x^T y$ is nonnegative for any feasible pair (x, y). We will assume that the primal feasibility set contains strictly feasible points and that the set of optimal solutions for the primal linear program is nonempty and bounded. For any optimal feasible pair (x_*, y_*), the duality gap is closed, i.e., $x_*^T y_* = 0$.

* Received by the editors August 6, 1990; accepted for publication (in revised form) March 12, 1991. This research was supported in part by National Science Foundation Cooperative Agreement CCR-8809615, Air Force Office of Scientific Research grant 89-0363, Department of Energy grant DEFG05-86ER25017, and Army Research Office grant 9DAALO3-90-G-0093. A preliminary version of this paper was presented at the Second Asilomar Workshop on Progress in Mathematical Programming, Monterey, CA, February 4–7, 1990.

[†] Department of Mathematics and Statistics, University of Maryland, Baltimore, Maryland 21228.

[‡] Department of Mathematical Sciences, Rice University, Houston, Texas 77251-1892.

Primal-dual interior point algorithms attempt to solve the primal and dual linear programs simultaneously by generating a sequence of strictly feasible pairs $\{(x_k, y_k)\}$ (and often another dual variable vector—the Lagrange multipliers associated with the primal constraints $Ax = b$) that converges to an optimal feasible pair (x_*, y_*). The objective of such algorithms is to drive the duality gap $x_k^T y_k$ to zero. Primal-dual approaches of this form were first introduced by Megiddo [14] using a logarithmic barrier function method. Megiddo's idea was developed by Kojima, Mizuno, and Yoshise [9] into a full algorithm with a polynomial complexity bound. A conceptually different approach was proposed by Todd and Ye [21] based on reducing a primal-dual potential function that is analogous to the Karmarkar primal potential function [8]. Other works on primal-dual interior point algorithms include Monteiro and Adler [17]; Lustig [11], [10]; Gonzaga and Todd [6]; Huang and Kortanek [7]; Choi, Monma, and Shanno [3]; McShane, Monma, and Shanno [13]; and Lustig, Marsten, and Shanno [12].

The above works can be classified roughly into two groups. Papers in the first group ([9], [21], for example) focused on designing algorithms with polynomial complexity bounds. Papers in the second group ([3], [12], [13], for example) were more concerned with computational and implementational issues. Unfortunately, there is a discrepancy between the two groups. That is, the algorithms that were described in the second group and were shown to have good practical performance are not those that were studied in the first group and were shown to possess polynomial complexity bounds. This discrepancy is understandably due to the limitation of the worst case analysis used in deriving polynomial complexity bounds. Recently, there have been works aimed at narrowing this discrepancy from a probabilistic point of view; see Mizuno, Todd, and Ye [15], [16]. In the current work, we try to shed light on another fundamental aspect of continuous optimization algorithms; namely, the blending of two often conflicting objectives: global convergence and fast local convergence. A convergence rate analysis for algorithms that belong to a very general class of primal-dual interior point methods is presented. This theory shows how superlinear and quadratic convergence can be attained by primal-dual interior point algorithms.

It is well understood, in the continuous optimization community, that fast local convergence is an important factor in evaluating the efficiency of an iterative method. Moreover, while interior point algorithms for linear programming are certainly iterative methods, local convergence properties have not received much attention. A plausible explanation for this lack of attention is the common belief that interior point algorithms essentially possess finite termination. That is, once one gets close enough to the optimal solution set, the interior point method can be terminated and available information (mainly the zero-nonzero structure of an optimal solution) can be used to obtain an optimal solution through some finite procedure. In the context of this guessing strategy, it is natural to question the value of fast local convergence in linear programming applications. However, our computational experience has taught us that although a correct early guess, on occasion, is certainly possible, especially in the case of a nondegenerate optimal vertex, in general one needs to be very close to the solution set in order to *guarantee* a correct guess. In addition, fast convergence usually occurs much earlier than the standard Newton's method theory predicts—a property often referred to in nonlinear applications as the semilocal behavior of Newton's method. Therefore, the construction of algorithms with fast local convergence can be an important and beneficial activity even in linear programming applications. However, in the interest of conciseness we have decided to present only theory in the present study. A comprehensive numerical investigation is the subject of a current study.

The concept of the central path (trajectory) plays an important role in designing and analyzing interior point algorithms. It was first studied in linear programming by Sonnevend [18] and by Bayer and Lagarias [1], [2]; see also Megiddo [14]. The central path can be expressed in several ways. Perhaps the simplest is that a strictly feasible pair (x, y) is on the central path if and only if it satisfies

$$[x]_1[y]_1 = [x]_2[y]_2 = \cdots = [x]_n[y]_n,$$

where $[x]_i$ ($[y]_i$) is the ith element of x (y), or equivalently,

$$[x]_i[y]_i = x^T y/n, \qquad i = 1, 2, \cdots, n. \tag{1.3}$$

This paper is organized as follows. In §2, we describe a general primal-dual interior point algorithmic framework. Then in §3, we present our superlinear convergence rate analysis and in §4, we present our quadratic convergence rate analysis. Concluding remarks are given in §5.

2. A primal-dual algorithmic framework.

In this section, we describe a general primal-dual interior point algorithmic framework. This general framework can also be derived from the point of view of barrier function methods or potential function reduction methods, as was done, for example, in [9] and [21]. We hope that our somewhat different approach offers new insight into these algorithms.

If the primal variables and the dual slack variables are updated at a given strictly feasible pair (x, y) by the formulas

$$x_+ = X(e + \alpha p) \quad \text{and} \quad y_+ = Y(e + \alpha q), \tag{2.1}$$

where $X = \text{diag}(x)$, $Y = \text{diag}(y)$, $e \in \mathbf{R}^n$ has all components equal to one, $p, q \in \mathbf{R}^n$, and $\alpha > 0$ is the step-length, then in order for x_+ and y_+ to be strictly feasible, p, q, and α must satisfy

$$AXp = 0 \quad \text{and} \quad e + \alpha p > 0, \tag{2.2}$$

$$BYq = 0 \quad \text{and} \quad e + \alpha q > 0. \tag{2.3}$$

We will consider projected gradient–type methods. Namely, the feasible directions p and q are obtained by projecting the negative gradients of relevant functions into the null spaces of AX and BY, respectively. Therefore, we first need to construct two $n \times n$ projection matrices H_p and H_q such that $AXH_p = 0$ and $BYH_q = 0$. If A and B were not scaled by X and Y, respectively, then it would be sufficient to define $H_q = P_A$ and $H_p = I - P_A$, where $P_A = A^T(AA^T)^{-1}A$. This definition would give $AH_p = 0$ and $BH_q = 0$ because $A^T \perp B^T$. Obviously, in this case both H_p and H_q would be orthogonal projections and therefore would be symmetric and positive semidefinite. The symmetry and positive semidefiniteness of H_p (H_q) is important because for any function $\phi : \mathbf{R}^n \to \mathbf{R}$, the projected negative gradient $-H_p\nabla\phi$ ($-H_q\nabla\phi$) will be not only a primal (dual) feasible direction but also a descent direction for ϕ as long as $H_p\nabla\phi \neq 0$ ($H_q\nabla\phi \neq 0$). Furthermore, it is worth noting that one would only need to compute either H_p or H_q because $H_p + H_q = I$.

Even though the matrices A and B are scaled by X and Y, respectively, it is still possible to construct two projection matrices H_p and H_q based on just one orthogonal projection matrix (though H_p and H_q themselves will not be orthogonal projections) and obtain the desirable property that both H_p and H_q are symmetric positive semidefinite. Consider the following matrices that we will call *scaled projections*:

$$H_p = \hat{D}(I - \hat{P})\hat{D} \quad \text{and} \quad H_q = \hat{D}\hat{P}\hat{D}.$$

Here $\hat{D}$ is a positive-definite diagonal matrix and $\hat{P}$ is an orthogonal projection matrix, both contained in $\mathbf{R}^{n \times n}$. The equations $A(XH_p) = 0$ and $B(YH_q) = 0$ and the fact $A^T \perp B^T$ imply that $H_p XY H_q = 0$, which in turn requires that $\hat{P}(\hat{D}XY\hat{D})(I - \hat{P}) = 0$. The last equation will hold for any orthogonal projection matrix $\hat{P}$ if $\hat{D}XY\hat{D} = I$. This leads to the following choice for $\hat{D}$:

$$\hat{D} = (XY)^{-1/2}.$$

It now follows from $AXH_p = 0$ that $(AX^{1/2}Y^{-1/2})(I - \hat{P}) = 0$. Hence we need to define the orthogonal projection matrix $\hat{P}$ as the orthogonal projection into the range space of $X^{1/2}Y^{-1/2}A^T$, namely,

$$\hat{P} = X^{1/2}Y^{-1/2}A^T(AXY^{-1}A^T)^{-1}AX^{1/2}Y^{-1/2}. \tag{2.4}$$

This definition of $\hat{P}$ gives not only $AXH_p = 0$, but also $BYH_q = 0$. Therefore, we finally conclude that the choices for the two scaled projection matrices H_p and H_q should be

$$H_p = (XY)^{-1/2}(I - \hat{P})(XY)^{-1/2}, \tag{2.5}$$

$$H_q = (XY)^{-1/2}\hat{P}(XY)^{-1/2}, \tag{2.6}$$

where $\hat{P}$ is defined by (2.4). The proof of the following proposition is now straightforward.

PROPOSITION 2.1. *If H_p and H_q are defined by* (2.5) *and* (2.6), *respectively, then*

(1) *Both H_p and H_q are symmetric positive semidefinite;*

(2) $AXH_p = 0$ *and* $BYH_q = 0$;

(3) $H_p XY H_q = 0$;

(4) $H_p + H_q = (XY)^{-1}$.

Obviously, the scaled projection H_p (H_q) will project the negative gradient into a primal (dual) feasible direction, which is also a descent direction (provided that the projection is nonzero). It is worth noting that in order to construct the two scaled projections we only need to calculate one orthogonal projection matrix $\hat{P}$.

To derive the directions p and q in (2.1), we first define a function

$$\phi(u, v) = (e + u)^T XY(e + v). \tag{2.7}$$

Obviously, if $x_+ = X(e + u)$ and $y_+ = Y(e + v)$ are primal and dual feasible, respectively, then $\phi(u, v) = x_+^T y_+ \geq 0$ represents the duality gap at the updated pair (x_+, y_+). It is easy to see that

$$\nabla_u \phi(0, 0) = \nabla_v \phi(0, 0) = XYe. \tag{2.8}$$

Now define

$$p_\phi = -H_p \nabla_u \phi(0, 0) = -[(XY)^{-1/2}(I - \hat{P})(XY)^{-1/2}]XYe, \tag{2.9}$$

$$q_\phi = -H_q \nabla_v \phi(0, 0) = -[(XY)^{-1/2}\hat{P}(XY)^{-1/2}]XYe. \tag{2.10}$$

From Proposition 2.1 (1), (2), the above-defined (p_ϕ, q_ϕ) is clearly a feasible descent direction for $\phi(u,v)$ at the current point $(0,0)$. We call (p_ϕ, q_ϕ) the duality gap–reducing direction.

Using the formulas (2.1), we define the barrier function at the given strictly feasible pair (x,y) as

$$\psi(u,v) = -\sum_{i=1}^{n} \ln([X(e+u)]_i[Y(e+v)]_i), \tag{2.11}$$

where $[a]_i$ denotes the ith element of the vector a. The gradient of $\psi(u,v)$ at the current point $(u,v) = (0,0)$ satisfies

$$\nabla_u \psi(0,0) = \nabla_v \psi(0,0) = -e. \tag{2.12}$$

The scaled projections of the components of the negative gradient direction of ψ into the primal and dual feasible spaces are, respectively,

$$p_\psi = -H_p \nabla_u \psi(0,0) = [(XY)^{-1/2}(I - \hat{P})(XY)^{-1/2}]e, \tag{2.13}$$

$$q_\psi = -H_q \nabla_v \psi(0,0) = [(XY)^{-1/2}\hat{P}(XY)^{-1/2}]e. \tag{2.14}$$

The direction (p_ψ, q_ψ) defined above is a descent direction for the barrier function $\psi(u,v)$ at the current point $(0,0)$; thus it pulls the next iterate towards the interior of the primal and dual feasible sets. We will call (p_ψ, q_ψ) the centering direction.

In almost every primal-dual interior point algorithm, the step direction in the primal or dual space is a linear combination of the duality gap–reducing direction and the centering direction. More specifically, for some $\sigma \in [0,1)$,

$$p = p_\phi + \sigma \frac{x^T y}{n} p_\psi = -H_p \left(XYe - \sigma \frac{x^T y}{n} e \right), \tag{2.15}$$

$$q = q_\phi + \sigma \frac{x^T y}{n} q_\psi = -H_q \left(XYe - \sigma \frac{x^T y}{n} e \right). \tag{2.16}$$

Hereafter, we will use the notation:

$$\min(u) = \min_{1 \le i \le n} [u]_i, \qquad \min(u,v) = \min_{1 \le i \le n} ([u]_i, [v]_i)$$

for $u, v \in \mathbf{R}^n$; the corresponding quantities for the maximums are similarly defined.

The following proposition can be easily verified using Proposition 2.1 and direct substitution.

PROPOSITION 2.2. *If p and q are defined by* (2.15) *and* (2.16), *respectively, then*

(1) *$AXp = 0$ and $BYq = 0$;*

(2) *$p^T XYq = 0$;*

(3) *$p + q = -e + \sigma \frac{x^T y}{n} (XY)^{-1} e$;*

(4) *$(e + \alpha p)^T XY(e + \alpha q) = x^T y[1 - \alpha(1 - \sigma)]$.*

We define the step-length α in (2.1) by the formula

$$\alpha = \frac{-\tau}{\min(p,q)}, \qquad \tau \in (0,1). \tag{2.17}$$

These choices of p, q, and α guarantee that the new primal and dual variables x_+ and y_+ obtained from formulas (2.1) will remain strictly feasible.

We now state an algorithmic framework for interior point primal-dual algorithms.

ALGORITHM 1. *Given a strictly feasible pair* (x_0, y_0), *for* $k = 0, 1, 2, \cdots$, *let*

$$x_{k+1} = X_k(e + \alpha_k p_k) \quad \textit{and} \quad y_{k+1} = Y_k(e + \alpha_k q_k), \tag{2.18}$$

where p_k, q_k, *and* α_k *are defined by* (2.15), (2.16), *and* (2.17), *respectively, and all the quantities involved* (*including* σ *and* τ) *are indexed by* k.

This algorithm generates strictly feasible sequences $\{x_k\}$ and $\{y_k\}$. It is a descent algorithm for the duality gap, which is reduced at iteration k by a factor $1 - \alpha_k(1 - \sigma_k) < 1$. Almost all the existing primal-dual algorithms that use only one projection per iteration fit into the above algorithmic framework with different choices for the parameters σ_k and τ_k.

For example, in the primal-dual algorithm of Kojima, Mizuno, and Yoshise [9], at each iteration a constant σ_k is chosen from $(0, 1)$ and, depending on this value of σ_k, restrictions are put on the parameter τ_k to ensure a polynomial complexity bound. In similar primal-dual algorithms implemented by Choi, Monma, and Shanno [3]; McShane, Monma, and Shanno [13]; and Lustig, Marsten, and Shanno [12], very small values of σ_k were used and long steps were taken. Impressive numerical results have been obtained for these implementations, though a polynomial complexity bound is no longer known.

Other examples include Todd and Ye's primal-dual potential reduction algorithm [21] and Monteiro and Adler's path-following primal-dual algorithms [17]. Todd and Ye's primal-dual potential function is

$$\Phi_\rho(x, y) = (n + \rho)\ln(\text{trace}[XY]) - \ln(\det[XY]).$$

This choice was motivated by the Karmarkar primal potential function [8]. At a given strictly feasible pair (x, y), if we define $\hat{\Phi}_\rho(u, v) = \Phi_\rho(X(e + u), Y(e + v))$, then we can see, though this was not the way the authors derived their algorithm, that the scaled projected negative gradient direction of $\hat{\Phi}_\rho(u, v)$ at $(0, 0)$ gives the updating directions for (x, y) proposed by Todd and Ye and they are of the form of (2.15) and (2.16). Todd and Ye used $\rho = \nu\sqrt{n}$ in their algorithm where ν is a positive constant. This choice of ρ leads, at each iteration, to the choice

$$\sigma_k = \frac{\sqrt{n}}{\sqrt{n} + \nu}$$

in (2.15) and (2.16). In Monteiro and Adler's path-following primal-dual algorithms [17], one can show that

$$\sigma_k = 1 - \frac{\delta}{\sqrt{n}},$$

where δ is chosen to be a number in $(0, \sqrt{n})$ subject to a certain restriction. The restriction is such that δ is bounded above as $n \to \infty$ (Monteiro and Adler actually chose $\delta = 0.35$ in their analysis).

3. Superlinear convergence. We first introduce two quantities defined at each iteration of Algorithm 1. At the kth iteration, let

$$\theta_k = \frac{x_k^T y_k / n}{\max(X_k Y_k e)} \quad \text{and} \quad \eta_k = \frac{x_k^T y_k / n}{\min(X_k Y_k e)}.$$

Since $x_k^T y_k/n$ is the average value of the elements of $X_k Y_k e$, it is clear that $\theta_k \leq 1$ and $\eta_k \geq 1$. Moreover, it follows from (1.3) that the pair (x_k, y_k) is on the central path if and only if $\theta_k = 1$ or, equivalently, $\eta_k = 1$.

In this section, we present two distinct Q-superlinear convergence theories, namely, Theorems 3.1 and 3.6. Our first Q-superlinear convergence theory is quite general and makes no nondegeneracy assumption. Some relevant comments will follow its proof.

THEOREM 3.1. *Let $\{x_k\}$ and $\{y_k\}$ be generated by Algorithm* 1, *$x_k \to x_*$, and $y_k \to y_*$. Assume* (i) *strict complementarity,* (ii) *the sequence $\{\eta_k\}$ is bounded, and* (iii) *$\tau_k \to 1$ and $\sigma_k \to 0$. Then the duality gap sequence $\{x_k^T y_k\}$ converges to zero Q-superlinearly. That is, the Q_1-factor*

$$Q_1 = \lim_{k\to\infty} \sup \frac{x_{k+1}^T y_{k+1}}{x_k^T y_k} = 0. \tag{3.1}$$

Proof. From Proposition 2.2 (4), we have

$$Q_1 = 1 - \lim_{k\to\infty} \inf \alpha_k (1 - \sigma_k).$$

Since $\sigma_k \to 0$, $Q_1 = 0$ if and only if $\liminf_{k\to\infty} \alpha_k = 1$. We will prove that $\alpha_k \to 1$.

Multiply both sides of the equation in Proposition 2.2 (3) by $(X_k Y_k)^{1/2}$ and consider the square of the ℓ_2-norm of both sides. From Proposition 2.2 (2) we have

$$\|(X_k Y_k)^{1/2} p_k\|_2^2 + \|(X_k Y_k)^{1/2} q_k\|_2^2 = x_k^T y_k \left(1 - 2\sigma_k + \sigma_k^2 \frac{x_k^T y_k}{n} \frac{e^T (X_k Y_k)^{-1} e}{n}\right);$$

or equivalently,

$$\|T_k^{-1/2} p_k\|_2^2 + \|T_k^{-1/2} q_k\|_2^2 = n \left(1 - 2\sigma_k + \sigma_k^2 \frac{e^T T_k e}{n}\right), \tag{3.2}$$

where $T_k = (x_k^T y_k/n)(X_k Y_k)^{-1}$. Assumption (ii) implies that $\{T_k\}$ is bounded above and $\{T_k^{-1/2}\}$ is bounded away from zero. Therefore, from (3.2), both $\{p_k\}$ and $\{q_k\}$ are bounded. It follows from (2.17) that $\{\alpha_k\}$ is bounded away from zero.

Now assume $[x_*]_i > 0$. Obviously,

$$1 = \lim_{k\to\infty} \frac{[x_{k+1}]_i}{[x_k]_i} = \lim_{k\to\infty} (1 + \alpha_k [p_k]_i).$$

This implies $[p_k]_i \to 0$, because $\{\alpha_k\}$ is bounded away from zero. Since $\sigma_k \to 0$, from Proposition 2.2 (3) we have $(p_k + q_k) \to -e$. Hence $[q_k]_i \to -1$. On the other hand, if $[x_*]_i = 0$, then $[y_*]_i > 0$ by strict complementarity. The same argument, interchanging the roles of p_k and q_k, gives $[q_k]_i \to 0$ and $[p_k]_i \to -1$. Therefore, the components of p_k and q_k converge to either 0 or -1. Consequently, from (2.17), $\alpha_k \to 1$ since $\tau_k \to 1$. This completes the proof. □

In Theorem 3.1, a source of concern has been the compatibility of assumptions (ii) and (iii). On the surface, it seems as if letting $\tau_k \to 1$ and $\sigma_k \to 0$ might force $\eta_k \to \infty$. However, our numerical experience has shown this not to be the case. Indeed, Theorem 3.1 was the direct consequence of a rather extensive numerical experimentation. The superlinear convergence theory presented in the first draft of this paper consisted only of Theorem 3.6 and required the assumption that $\{\eta_k\}$ be bounded. This assumption has been removed in the present version. In subsequent numerical studies with highly degenerate Netlib problems, we let $\tau_k \to 1$ and $\sigma_k \to 0$ and we always observed strict complementarity, $\{\eta_k\}$ bounded, $\alpha_k \to 1$, and Q-superlinear convergence. This phenomenon motivated us to search for a theory that could explain this occurrence and consequently led to the discovery of Theorem 3.1. We feel that Theorem 3.1 offers a satisfactory explanation of what we observed in practice. (In a more recent study, Zhang and Tapia [22] have proved that it is possible to choose $\sigma_k \to 0$ and $\tau_k \to 1$ while maintaining global convergence and the boundedness of $\{\eta_k\}$. Thus, the compatibility of the assumptions in Theorem 3.1 has been demonstrated.)

In numerical computation, the boundedness of $\{\eta_k\}$ requires some qualification because an algorithm is always stopped in a finite number of iterations. In our numerical experiments, we did not observe the trend of continued growth in the values of η_k as our algorithm was about to stop, while the observed convergence was clearly Q-superlinear and $\alpha_k \to 1$. Of course, the behavior of $\{\eta_k\}$ varies with several factors, including how fast $\{\tau_k\}$ converges to one and $\{\sigma_k\}$ to zero. We do not imply that unbounded $\{\eta_k\}$ can never occur. Instead, we feel that it appears to be more an exception than the rule. This topic undoubtedly merits further study.

In the following development, we show that if we assume nondegeneracy, then we can obtain Q-superlinear convergence without assuming the boundedness of $\{\eta_k\}$. The following theorem concerns the Q_1 factor of the duality gap sequence.

THEOREM 3.2. *Let $\{x_k\}$ and $\{y_k\}$ be generated by Algorithm* 1, *$x_k \to x_*$, and $y_k \to y_*$. Assume* (i) *strict complementarity and* (ii) *x_* is a nondegenerate vertex. Then the duality gap sequence $\{x_k^T y_k\}$ converges to zero and the Q_1-factor is*

$$Q_1 = \limsup_{k\to\infty} \frac{x_{k+1}^T y_{k+1}}{x_k^T y_k} = 1 - \liminf_{k\to\infty} \frac{\tau_k(1-\sigma_k)}{1-\theta_k\sigma_k}. \tag{3.3}$$

To prove Theorem 3.2, we need the following two lemmas. The first lemma has been proved in [20] under slightly different assumptions. For the sake of completeness, we include its proof here.

LEMMA 3.3. *Let $\hat{P}_k$ be defined by* (2.4) *with X and Y indexed by k. Without loss of generality, assume that the first m elements of x_* are positive. Then under the assumptions of Theorem* 3.2,

$$\lim_{k\to\infty} \hat{P}_k = \begin{bmatrix} I_m & 0 \\ 0 & 0 \end{bmatrix}. \tag{3.4}$$

Proof. Let

$$d_k = X_k^{1/2} Y_k^{-1/2} e \quad \text{and} \quad D_k = \operatorname{diag}(d_k).$$

Then

$$\hat{P}_k = D_k A^T (A D_k^2 A^T)^{-1} A D_k.$$

By our assumptions, we have

$$[y_k]_i \to 0, \qquad i = 1, 2, \cdots, m$$

and

$$[x_k]_i \to 0, \qquad i = m+1, m+2, \cdots, n.$$

It then follows from the definition of d_k that

$$[d_k]_i \to +\infty, \qquad i = 1, 2, \cdots, m$$

and

$$[d_k]_i \to 0, \qquad i = m+1, m+2, \cdots, n.$$

Now let A_1 be the $m \times m$ submatrix of A consisting of its first m columns and A_0 be the $m \times (n-m)$ submatrix of A consisting of its last $n-m$ columns. Clearly, A_1 is nonsingular. Similarly, let D_k^∞ and D_k^0 be the diagonal matrices of dimensions m and $n-m$, respectively, with the first m and the last $n-m$ elements of d_k on their diagonals, respectively. Evidently, D_k^∞ is nonsingular for all k and $\{D_k^0\}$ converges to zero.

Substituting $AD_k = [A_1 D_k^\infty \;\; A_0 D_k^0]$ for $AX_k^{1/2} Y_k^{-1/2}$ in (2.4), we obtain

$$\hat{P}_k = [A_1 D_k^\infty \;\; A_0 D_k^0]^T [A_1 (D_k^\infty)^2 A_1^T + A_0 (D_k^0)^2 A_0^T]^{-1} [A_1 D_k^\infty \;\; A_0 D_k^0].$$

Note that A_1 is nonsingular and let

$$R_k = (D_k^\infty)^{-1} A_1^{-1} A_0 D_k^0. \tag{3.5}$$

We have

$$\hat{P}_k = \begin{bmatrix} (I_m + R_k R_k^T)^{-1} & (I_m + R_k R_k^T)^{-1} R_k \\ R_k^T (I_m + R_k R_k^T)^{-1} & R_k^T (I_m + R_k R_k^T)^{-1} R_k. \end{bmatrix}. \tag{3.6}$$

Since $D_k^0 \to 0$ and $(D_k^\infty)^{-1} \to 0$, so does R_k. Now it is evident that

$$\lim_{k \to \infty} \hat{P}_k = \begin{bmatrix} I_m & 0 \\ 0 & 0 \end{bmatrix},$$

which completes the proof. □

Our next lemma will be used not only in the proof of Theorem 3.2, but also in our quadratic convergence theory.

LEMMA 3.4. *Let $\hat{P}_k$, p_k, and q_k be defined by* (2.4), (2.15), *and* (2.16), *respectively, with X and Y indexed by k. Under the assumptions of Lemma* 3.3,

$$p_k = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ -1 \\ \vdots \\ -1 \end{pmatrix} + \sigma_k \begin{pmatrix} 0 \\ \vdots \\ 0 \\ \dfrac{x_k^T y_k / n}{[X_k Y_k e]_{m+1}} \\ \vdots \\ \dfrac{x_k^T y_k / n}{[X_k Y_k e]_n} \end{pmatrix} + O(x_k^T y_k)$$

and

$$q_k = \begin{pmatrix} -1 \\ \vdots \\ -1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} + \sigma_k \begin{pmatrix} \dfrac{x_k^T y_k/n}{[X_k Y_k e]_1} \\ \vdots \\ \dfrac{x_k^T y_k/n}{[X_k Y_k e]_m} \\ 0 \\ \vdots \\ 0 \end{pmatrix} + O(x_k^T y_k),$$

where the number of zeros is m in p_k, and $n - m$ in q_k.

Proof. Since $R_k \to 0$,

$$(I_m + R_k R_k^T)^{-1} = I_m - R_k R_k^T + R_k O(\|R_k\|^2) R_k^T = I_m - R_k O(1) R_k^T.$$

Hence, from (3.6),

$$\hat{P}_k = \begin{bmatrix} I_m & 0 \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} 0 & R_k \\ R_k^T & 0 \end{bmatrix} - E_k,$$

where

$$E_k = \begin{bmatrix} R_k & 0 \\ 0 & R_k^T \end{bmatrix} \begin{bmatrix} O(1) & O(1)R_k^T \\ R_k O(1) & O(1) \end{bmatrix} \begin{bmatrix} R_k^T & 0 \\ 0 & R_k \end{bmatrix}.$$

It follows from the definition of p_k and q_k (see (2.15) and (2.16)) that

$$p_k = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ -1 \\ \vdots \\ -1 \end{pmatrix} + \sigma_k \begin{pmatrix} 0 \\ \vdots \\ 0 \\ \dfrac{x_k^T y_k/n}{[X_k Y_k e]_{m+1}} \\ \vdots \\ \dfrac{x_k^T y_k/n}{[X_k Y_k e]_n} \end{pmatrix} + r_k^p,$$

where

$$r_k^p = (X_k Y_k)^{-1/2} \left(\begin{bmatrix} 0 & R_k \\ R_k^T & 0 \end{bmatrix} - E_k \right) (X_k Y_k)^{-1/2} \left(X_k Y_k e - \sigma_k \frac{x_k^T y_k}{n} e \right).$$

By strict complementarity, $[X_k Y_k e]_i = O([x_k]_i)$ when $[x_*]_i = 0$ and $[X_k Y_k e]_i = O([y_k]_i)$ when $[x_*]_i > 0$. Also note that $x_k^T y_k = \|X_k Y_k e\|_1$. From (3.5) it can be verified that

$$\begin{bmatrix} R_k^T & 0 \\ 0 & R_k \end{bmatrix} (X_k Y_k)^{-1/2} = O((x_k^T y_k)^{1/2}).$$

Hence, $(X_k Y_k)^{-1/2} E_k (X_k Y_k)^{-1/2} = O(x_k^T y_k)$ and consequently

$$r_k^p = (X_k Y_k)^{-1/2} \begin{bmatrix} 0 & R_k \\ R_k^T & 0 \end{bmatrix} \left((X_k Y_k)^{1/2} e - \sigma_k \frac{x_k^T y_k}{n} (X_k Y_k)^{-1/2} e \right) + O(x_k^T y_k).$$

A straightforward matrix-vector multiplication shows that

$$[r_k^p]_i = \begin{cases} \frac{1}{[x_k]_i} \sum_{j=m+1}^{n} [A_1^{-1} A_0]_{i,j-m} \left([x_k]_j - \sigma_k \frac{x_k^T y_k}{n} \frac{1}{[y_k]_j} \right) + O(x_k^T y_k), & 1 \le i \le m, \\ \frac{1}{[y_k]_i} \sum_{j=1}^{m} [A_0^T A_1^{-T}]_{i-m,j} \left([y_k]_j - \sigma_k \frac{x_k^T y_k}{n} \frac{1}{[x_k]_j} \right) + O(x_k^T y_k), & m < i \le n. \end{cases}$$

Since

$$\lim_{k\to\infty} [x_k]_i = \begin{cases} [x_*]_i > 0, & 1 \le i \le m, \\ 0, & m < i \le n, \end{cases}$$

and

$$\lim_{k\to\infty} [y_k]_i = \begin{cases} 0, & 1 \le i \le m, \\ [y_*]_i > 0, & m < i \le n, \end{cases}$$

it is evident that $r_k^p = O(x_k^T y_k)$. This proves the first equality for p_k. Similarly, we can prove the second equality for q_k. □

Now we are ready to prove Theorem 3.2.

Proof of Theorem 3.2. Without loss of generality, we assume that the first m components of x_* are positive and consequently the remaining $n-m$ components are zero.

It follows from Lemma 3.3, Lemma 3.4, and (2.17) that

$$\alpha_k = \frac{-\tau_k}{-1 + \sigma_k \frac{x_k^T y_k / n}{\max(X_k Y_k e)} + O(x_k^T y_k)} = \frac{\tau_k}{1 - \sigma_k \theta_k + O(x_k^T y_k)}. \tag{3.7}$$

From Proposition 2.2 (4),

$$\frac{x_{k+1}^T y_{k+1}}{x_k^T y_k} = 1 - \alpha_k (1 - \sigma_k) = 1 - \frac{\tau_k (1 - \sigma_k)}{1 - \sigma_k \theta_k + O(x_k^T y_k)}.$$

Now (3.3) follows immediately. This completes the proof. □

Observe that $\tau_k \in (0,1)$, $\sigma_k \in [0,1)$, and $\theta_k \in (0,1]$. Therefore, for all k,

$$\frac{\tau_k (1 - \sigma_k)}{1 - \theta_k \sigma_k} < 1.$$

Thus from (3.3), $Q_1 = 0$ if and only if

$$\lim_{k\to\infty} \tau_k \frac{1 - \sigma_k}{1 - \theta_k \sigma_k} = 1. \tag{3.8}$$

By examining (3.8), we have the following corollary. Its proof should be straightforward.

COROLLARY 3.5. *Under the assumptions of Theorem* 3.2,

(1) *If* $\lim_{k\to\infty} \tau_k < 1$, *then* $Q_1 > 0$.

(2) *If* $\lim_{k\to\infty} \tau_k = 1$, *then* $Q_1 = 0$ *if and only if*

$$\lim_{k\to\infty} \frac{1 - \sigma_k}{1 - \theta_k \sigma_k} = 1.$$

In particular, the above limit is 1 *if* $\sigma_k \to 0$ *or* $\theta_k \to 1$.

To emphasize the significance of Corollary 3.5, we formally state its interpretation as the following theorem. It is important to remember that $\{\sigma_k\}$ and $\{\tau_k\}$ are directly under our control, but $\{\theta_k\}$ is not.

THEOREM 3.6. *Under the assumptions of Theorem* 3.2, *the duality gap sequence* $\{x_k^T y_k\}$ *generated by Algorithm* 1 *converges to zero Q-superlinearly if the sequence* $\{\tau_k\}$ *converges to* 1 *and either of the following two conditions holds:*

(1) *The centering step is phased out asymptotically, i.e.,* $\lim_{k\to\infty} \sigma_k = 0$.

(2) *The convergence of the primal-dual sequence* $\{(x_k, y_k)\}$ *to* (x_*, y_*) *is along the central path, i.e.,* $\lim_{k\to\infty} \theta_k = 1$.

The convergence of $\{x_k^T y_k\}$ *is no better than Q-linear if* $\lim_{k\to\infty} \tau_k < 1$.

It is interesting to compare Theorem 3.1 to Theorem 3.6. The assumptions for the two theorems are different. In the proofs of the two theorems, we used different approaches and obtained distinct results.

Theorem 3.6 states that it is not necessary to have $\sigma_k \to 0$ in order to attain superlinear convergence. Admittedly, the case where the iterates converge asymptotically along the central path is a very special and perhaps unlikely case.

Observe that $\lim_{k\to\infty} \tau_k = 1$ means that our step asymptotically approaches the boundary of the positive orthant. Another interesting observation from (3.3) is that, assuming $\lim_{k\to\infty} \tau_k = 1$,

$$\limsup_{k\to\infty} \frac{(1-\theta_k)\sigma_k}{1-\sigma_k\theta_k} = Q_1 \le \limsup_{k\to\infty} \sigma_k. \tag{3.9}$$

Therefore, even in the case of linear convergence, in general the smaller σ_k is, the faster the convergence will be. This may in part explain why good numerical performance was obtained from the implementations of primal-dual algorithms by Choi, Monma, and Shanno [3]; McShane, Monma, and Shanno [13]; and Lustig, Marsten, and Shanno [12], where very small values of σ ($\sigma = 1/n$ or $1/\sqrt{n}$) were used.

If the Todd and Ye potential function method [21] is used to generate updating directions with the choice $\rho = n + \nu\sqrt{n}$, then as previously mentioned,

$$\sigma = \frac{\sqrt{n}}{\sqrt{n}+\nu}.$$

Evidently, σ approaches 1 rapidly as n increases. Since the left-hand side of (3.9) tends to 1 as $\sigma \to 1$, unless $\theta_k \to 1$, the Q-linear convergence rate for this choice of σ will generally deteriorate towards 1 with the increase of n. Here we see clearly an inverse relationship between a good polynomial complexity bound (Todd and Ye proved that their algorithm converges in $O(\sqrt{n}L)$ iterations) and a good Q-convergence rate. Such a relationship also exists in Monteiro and Adler's $O(\sqrt{n}L)$-iteration path-following algorithms [17] where

$$\sigma = 1 - \frac{\delta}{\sqrt{n}}$$

and δ is bounded. Clearly, their path-following algorithms also show a deterioration of Q-convergence rate as the problem size increases. However, it is quite possible that the above-mentioned two algorithms can still have reasonable R-behavior.

Now we prove a stronger convergence result for those primal and dual variables that converge to zero.

THEOREM 3.7. *Let* $\{x_k\}$ *and* $\{y_k\}$ *be generated by Algorithm* 1, $x_k \to x_*$, *and* $y_k \to y_*$. *Assume* (i) *strict complementarity, and either* (ii) $\sigma_k\eta_k \to 0$ *and* $x_k^T y_k \to 0$

Q-superlinearly, or (iii) $\tau_k \to 1$, $\theta_k \to 1$, *and* x_* *is a nondegenerate vertex. Then the primal and dual variables that converge to zero do so Q-superlinearly.*

Proof. From (2.18), we have

$$X_k^{-1}x_{k+1} = e + \alpha_k p_k \quad \text{and} \quad Y_k^{-1}y_{k+1} = e + \alpha_k q_k.$$

Hence, by Proposition 2.2 (3), we have

$$X_k^{-1}x_{k+1} + Y_k^{-1}y_{k+1} = (2-\alpha_k)e + \alpha_k\sigma_k\frac{x_k^Ty_k}{n}(X_kY_k)^{-1}e. \tag{3.10}$$

Under assumption (ii), since $\sigma_k\eta_k \to 0$ the second term in the right-hand side of (3.10) vanishes in the limit (notice that $\eta_k = \|(x_k^Ty_k/n)(X_kY_k)^{-1}e\|_\infty$). Also, $\alpha_k \to 1$. Therefore,

$$\lim_{k\to\infty}(X_k^{-1}x_{k+1} + Y_k^{-1}y_{k+1}) = e. \tag{3.11}$$

On the other hand, under assumption (iii) the second term in the right-hand side of (3.10) converges to $\alpha_k\sigma_k e$. Meanwhile, it follows from (3.7) that $\alpha_k(1-\sigma_k) \to 1$. Hence (3.11) also holds.

If $[x_*]_i = 0$, then by strict complementarity, $[y_*]_i > 0$ and $[y_{k+1}]_i/[y_k]_i \to 1$. It follows from (3.11) that $[x_{k+1}]_i/[x_k]_i \to 0$. Therefore, $[x_k]_i \to 0$ Q-superlinearly. By the symmetry of the relation (3.11), we have $[y_k]_j \to 0$ Q-superlinearly if $[y_*]_j = 0$. □

Note that it is easy to enforce $\sigma_k\eta_k \to 0$ since we have direct control over σ_k and we can compute η_k before we set σ_k.

Since $x_k^Ty_k = \|X_kY_ke\|_1$, it is evident that when $\{x_k^Ty_k\}$ converges to zero Q-superlinearly, so does the sequence $\{X_kY_ke\}$. We now demonstrate that this superlinear convergence is actually componentwise.

COROLLARY 3.8. *Under the assumptions of Theorem* 3.7, *the sequence* $\{X_kY_ke\}$ *converges to zero Q-superlinearly componentwise.*

Proof. By strict complementarity, either $[x_k]_i \to 0$ or $[y_k]_i \to 0$ for each index i. From Theorem 3.7, we have either

$$\lim_{k\to\infty}\frac{[x_{k+1}]_i}{[x_k]_i} = 0 \quad \text{and} \quad \lim_{k\to\infty}\frac{[y_{k+1}]_i}{[y_k]_i} = 1$$

or

$$\lim_{k\to\infty}\frac{[x_{k+1}]_i}{[x_k]_i} = 1 \quad \text{and} \quad \lim_{k\to\infty}\frac{[y_{k+1}]_i}{[y_k]_i} = 0.$$

In either case,

$$\lim_{k\to\infty}\frac{[x_{k+1}]_i[y_{k+1}]_i}{[x_k]_i[y_k]_i} = \lim_{k\to\infty}\frac{[X_{k+1}Y_{k+1}e]_i}{[X_kY_ke]_i} = 0.$$

This completes the proof of the componentwise Q-superlinear convergence of $\{X_kY_ke\}$. □

In 1980, Tapia [19, Thm. 3] pointed out that an algorithm which at each iteration satisfies the Taylor linearization of the complementarity equation has the property that the variables that converge to zero do so Q-superlinearly. This result assumed strict complementarity and step-length one. Observe that (3.11) is equivalent to

$$X_kY_ke + Y_k(x_{k+1}-x_k) + X_k(y_{k+1}-y_k) \to 0.$$

We see that the Taylor linearization of complementarity is satisfied asymptotically in our situation.

We close this section by commenting that taking different step-lengths in the primal space and in the dual space may result in a larger reduction in the duality gap locally, i.e., at any given iteration; however, it seems unlikely that superlinear convergence could be achieved without both step-lengths approaching one asymptotically.

4. Quadratic convergence. In this section, we show that under the assumptions of Theorem 3.2 quadratic convergence can be achieved by primal-dual algorithms if we both phase out the centering direction and let the steps approach the boundary at a sufficiently fast rate. In contrast to the analysis of superlinear convergence, which is done in a scaled gradient-projection framework, the study of quadratic convergence will be within the framework of Newton's method.

We first reformulate Algorithm 1 as a perturbed and damped Newton's method.

It is well known that at optimality the primal, dual, and dual slack variables x, λ, and y satisfy

$$\begin{pmatrix} Ax-b \\ A^T\lambda + y - c \\ XYe \end{pmatrix} = 0, \tag{4.1}$$

$x \geq 0$, and $y \geq 0$. To eliminate the dual variables λ from the above system, we premultiply the second equation by the nonsingular matrix $[A^T \; B^T]^T$. Noticing that $BA^T = 0$, we obtain

$$0 = \begin{bmatrix} A \\ B \end{bmatrix} (A^T\lambda + y - c) = \begin{pmatrix} AA^T\lambda + A(y-c) \\ By - Bc \end{pmatrix}.$$

Since AA^T is nonsingular, λ is uniquely determined once y is known. Removing the equation for λ, we arrive at the following $2n \times 2n$ system consisting of primal feasibility (see (1.1)), dual feasibility (see (1.2)), and complementarity:

$$F(x,y) = \begin{pmatrix} Ax-b \\ By - Bc \\ XYe \end{pmatrix} = 0, \tag{4.2}$$

as well as the nonnegativity constraints for (x,y).

Similarly, we can show that a strictly feasible pair (x,y) on the central path satisfies

$$\hat{F}(x,y,\mu) = \begin{pmatrix} Ax-b \\ By - Bc \\ XYe - \mu e \end{pmatrix} = 0 \tag{4.3}$$

for some $\mu > 0$. Evidently, $\hat{F}(x,y,\mu) = 0$ is a perturbation of the system $F(x,y) = 0$ with the perturbation term $-\mu e$ added to the nonlinear portion of $F(x,y)$. It is also obvious that $\hat{F}(x,y,0) = F(x,y)$.

The following proposition relates the search direction (p,q) in Algorithm 1 to a perturbed Newton's direction $(\Delta x, \Delta y)$.

PROPOSITION 4.1. *Let (x,y) be a strictly feasible pair and let p and q be defined by* (2.15) *and* (2.16). *Then p and q satisfy*

$$\begin{pmatrix} Xp \\ Yq \end{pmatrix} = \begin{pmatrix} \Delta x \\ \Delta y \end{pmatrix} \equiv -[F'(x,y)]^{-1}\hat{F}(x,y,\mu) \tag{4.4}$$

for the choice

$$\mu = \sigma \frac{x^T y}{n}.$$

Proof. Notice that

$$\hat{F}'_{(x,y)}(x, y, \mu) = F'(x, y) = \begin{bmatrix} A & 0 \\ 0 & B \\ Y & X \end{bmatrix};$$

consequently, we have

$$\begin{bmatrix} A & 0 \\ 0 & B \\ Y & X \end{bmatrix} \begin{pmatrix} \Delta x \\ \Delta y \end{pmatrix} = - \begin{pmatrix} 0 \\ 0 \\ w \end{pmatrix},$$

where

$$w = XYe - \sigma \frac{x^T y}{n} e.$$

Thus $A\Delta x = 0$, so $\Delta x = B^T u$, where $u \in \mathbf{R}^{n-m}$. Similarly, $\Delta y = A^T v$, where $v \in \mathbf{R}^m$. Substituting Δx and Δy into the third equation block of the system and multiplying both sides by AY^{-1}, we obtain

$$AXY^{-1}A^T v = -AY^{-1} w.$$

Thus

$$\Delta y = A^T v = -A^T (AXY^{-1}A^T)^{-1} AY^{-1} w.$$

It is now straightforward to verify from (2.16) that

$$Y^{-1}\Delta y = -H_q w = q.$$

Consequently, by Proposition 2.2 (4),

$$X^{-1}\Delta x = -(XY)^{-1} w - Y^{-1}\Delta y = -H_p w = p.$$

This completes the proof. □

We can therefore view a primal-dual algorithm as a perturbed and damped Newton's method. At the kth iteration the iterate is obtained from the perturbed system $\hat{F}(x, y, \mu_k) = 0$. The sequence of the perturbation parameters $\{\mu_k\}$ converges to zero as $x_k^T y_k \to 0$. We use the qualifier damped because at each iteration the step-length is determined by formula (2.17) in order to keep the iterates in the interior of the feasibility set. The positivity requirements for x and y generally prevent a full Newton step from being taken. It is well known that taking full steps asymptotically is a critical ingredient for the Q-quadratic convergence of Newton's method (see Dennis and Moré [4, Cor. 2.3]).

We now rewrite Algorithm 1 in the following equivalent form of a perturbed and damped Newton's method.

ALGORITHM 2. *Given a strictly feasible pair (x_0, y_0), for $k = 0, 1, 2, \cdots$, let*

$$x_{k+1} = x_k + \alpha_k \Delta x_k \quad \text{and} \quad y_{k+1} = y_k + \alpha_k \Delta y_k, \tag{4.5}$$

where Δx_k and Δy_k are defined by (4.4), *α_k is defined by* (2.17), *and all the quantities involved* (*including σ and τ*) *are indexed by k.*

To establish Q-quadratic convergence for Algorithm 2, we need to address the following three issues:

1. Is the Jacobian matrix $F'(x, y)$ nonsingular at optimality?
2. How quickly must the centering direction—the perturbation controlled by μ—be phased out?
3. Can full Newton steps be taken asymptotically and at a rate that ensures quadratic convergence?

The following lemma answers the first question.

LEMMA 4.2. *Let (x_*, y_*) be an optimal pair for the linear programs* (1.1) *and* (1.2). *Under the assumptions of Theorem* 3.2, *the $2n \times 2n$ matrix*

$$F'(x_*, y_*) = \begin{bmatrix} A & 0 \\ 0 & B \\ Y_* & X_* \end{bmatrix}$$

is nonsingular.

Proof. It can be shown that assumptions (i) and (ii) of Theorem 3.2 imply that y_* is a nondegenerate vertex of the dual (1.2). Without loss of generality, we assume that the first m components of x_* are positive and consequently, the remaining $n - m$ components are zero. Let

$$x_*^+ = \begin{pmatrix} [x_*]_1 \\ \vdots \\ [x_*]_m \end{pmatrix} \quad \text{and} \quad y_*^+ = \begin{pmatrix} [y_*]_{m+1} \\ \vdots \\ [y_*]_n \end{pmatrix}.$$

By our assumptions, $x_*^+ > 0$ and $y_*^+ > 0$.

Let A_1 be the $m \times m$ submatrix of A consisting of its first m columns and A_0 be the $m \times (n - m)$ submatrix of A consisting of its last $n - m$ columns. Clearly, A_1 is nonsingular because its columns form the optimal basis for the primal linear program (1.1). The same ordering also leads to $B = [B_0 \; B_1]$, where $B_1 \in \mathbf{R}^{(n-m)\times(n-m)}$ is nonsingular and its columns form the optimal basis for the dual linear program (1.2). Using the above-introduced notation, we have

$$F'(x_*, y_*) = \begin{bmatrix} A_1 & A_0 & 0 & 0 \\ 0 & 0 & B_0 & B_1 \\ 0 & 0 & X_*^+ & 0 \\ 0 & Y_*^+ & 0 & 0 \end{bmatrix}. \tag{4.6}$$

By examining blocks of this matrix, we can easily see that it is indeed nonsingular. □

Further examination of (4.6) reveals that Lemma 4.2 is sharp in the sense that $F'(x_*, y_*)$ will be singular if the number of nonzeros in x_* (y_*) is not m ($n - m$).

An answer to the second question is not hard to find. From standard analysis for Newton-like methods, an $O(\|F(x, y)\|^2)$ perturbation term does not destroy quadratic convergence. In our context, this is equivalent to the requirement $\mu_k = O((x_k^T y_k)^2)$, since for any feasible pair (x, y),

$$x^T y = \|XYe\|_1 = \|F(x,y)\|_1.$$

An answer to the third question requires further analysis. For the ease of notation, let us denote the pair (x, y) by $z \in \mathbf{R}^{2n}$. In a damped Newton method for $F(z) = 0$, if Δz_k is the full Newton step at the kth iteration and α_k is the step-length, then

$$z_{k+1} = z_k + \alpha_k \Delta z_k = z_k + \Delta z_k - (1 - \alpha_k)\Delta z_k.$$

From standard analysis,

$$\|z_{k+1} - z_*\| \le O(\|z_k - z_*\|^2) + |1 - \alpha_k| \|\Delta z_k\|.$$

Since

$$\|\Delta z_k\| = O(\|F(z_k)\|) = O(\|z_k - z_*\|),$$

it is clear that if

$$|1 - \alpha_k| = O(\|z_k - z_*\|),$$

then quadratic convergence will be achieved.

From (3.7), we see that the step-length α_k depends on τ_k, σ_k, and an $O(x_k^T y_k)$ term. Since $x_k^T y_k = O(\|z_k - z_*\|)$, in order to ensure $|1 - \alpha_k| = O(\|z_k - z_*\|)$, we see that it is sufficient to have $1 - \tau_k$ and σ_k be $O(\|z_k - z_*\|)$. If we take $\sigma_k = O(x_k^T y_k)$, then we have

$$\sigma_k = O(\|z_k - z_*\|) \quad \text{and} \quad \mu_k = O(\|z_k - z_*\|^2).$$

Moreover, we can easily enforce the requirement $1 - \tau_k = O(x_k^T y_k) = O(\|z_k - z_*\|)$.

Now we are in a position to prove the following quadratic convergence theorem. Its proof is basically a rigorous and detailed treatment of the above discussion. As a by-product, we also obtain a local convergence result.

THEOREM 4.3. *Let* $\{(x_k, y_k)\}$ *be generated by Algorithm* 2. *Assume* (i) *strict complementarity,* (ii) x_* *is a nondegenerate vertex, and* (iii) *the choices of* σ_k *and* τ_k *satisfy at each iteration*

$$0 \le \sigma_k \le \min(\sigma, c_1 x_k^T y_k) \quad \textit{and} \quad \max(\tau, 1 - c_2 x_k^T y_k) \le \tau_k < 1, \tag{4.7}$$

where $\sigma \in [0, 1)$, $\tau \in (0, 1)$, *and* $c_1, c_2 > 0$. *Then*

(1) *whenever* $\{(x_k, y_k)\}$ *converges to* (x_*, y_*), *it does so Q-quadratically, i.e., there exists a constant* $\gamma > 0$ *such that for* k *sufficiently large,*

$$\|(x_{k+1}, y_{k+1}) - (x_*, y_*)\| \le \gamma \|(x_k, y_k) - (x_*, y_*)\|^2; \tag{4.8}$$

(2) *there exists a number* $\delta > 0$ *such that whenever* $\|(x_0, y_0) - (x_*, y_*)\| \le \delta$, *then* $\{(x_k, y_k)\}$ *converges to* (x_*, y_*).

Proof. Again we use the notation $z = (x, y)$. Also let $\hat{e} = (0 \cdots 0 \ 1 \cdots 1)^T \in \mathbf{R}^{2n}$ where the numbers of zeros and ones are both n. As mentioned in §2 after (2.17), the sequence $\{z_k\}$ is always well defined and remains strictly feasible.

Following the standard analysis for Newton-like methods (see Dennis and Schnabel [5], for example), we have

$$\begin{aligned} z_{k+1} - z_* &= z_k - z_* - \alpha_k [F'(z_k)]^{-1} \hat{F}(z_k, \mu_k) \\ &= [F'(z_k)]^{-1}\{[F(z_*) - F(z_k) - F'(z_k)(z_* - z_k)] + (1-\alpha_k)F(z_k) + \alpha_k \mu_k \hat{e}\}. \end{aligned}$$

Therefore,

$$\begin{aligned} \|z_{k+1} - z_*\| \le \|[F'(z_k)]^{-1}\| (&\|F(z_*) - F(z_k) - F'(z_k)(z_* - z_k)\| \\ &+ |1-\alpha_k|\, \|F(z_k)\| + \alpha_k \mu_k \|e\|). \end{aligned} \tag{4.9}$$

Note that $z_k = (x_k, y_k)$ is strictly feasible and that $x_k^T y_k = \|F(z_k)\|_1 = \|F(z_k) - F(z_*)\|_1$. There exist $\delta_1 > 0$ and $c_3 > 0$ such that if $\|z_k - z_*\| \le \delta_1$, then

$$x_k^T y_k \le c_3 \|z_k - z_*\| \quad \text{and} \quad \|F(z_k)\| \le c_3 \|z_k - z_*\|. \tag{4.10}$$

This follows from the fact that $F(z)$ is continuously differentiable. Also note that $F(z)$ is a quadratic, hence there exists $c_4 > 0$ such that for any k,

$$\|F(z_*) - F(z_k) - F'(z_k)(z_* - z_k)\| \le c_4 \|z_k - z_*\|^2. \tag{4.11}$$

In view of the continuity and nonsingularity of $F'(z)$ at z_*, there exist $\delta_2 > 0$ and $c_5 > 0$ such that if $\|z - z_*\| \le \delta_2$, then

$$\|[F'(z)]^{-1}\| \le c_5. \tag{4.12}$$

In addition, there exist $\delta_3 > 0$ and $c_6 > 0$ such that if z_k satisfies $\|z_k - z_*\| \le \delta_3$, then from (3.7),

$$\begin{aligned} |1 - \alpha_k| &= \left| \frac{(1-\tau_k) - \theta_k \sigma_k + O(x_k^T y_k)}{1 - \theta_k \sigma_k + O(x_k^T y_k)} \right| \\ &\le 2|(1-\tau_k) - \theta_k \sigma_k + c_6 x_k^T y_k| \\ &\le 2(c_2 + \theta_k c_1 + c_6) x_k^T y_k. \end{aligned}$$

Here we assumed that δ_3 is sufficiently small, so that $1 - \theta_k \sigma_k - c_6 x_k^T y_k \ge \frac{1}{2}$ and (4.10) holds. We also used the assumptions $\sigma_k \le c_1 x_k^T y_k$ and $1 - \tau_k \le c_2 x_k^T y_k$. Using (4.10) and noting $\theta_k \le 1$, we have

$$|1 - \alpha_k| \le c_7 \|z_k - z_*\|, \tag{4.13}$$

where $c_7 = 2c_3(c_2 + c_1 + c_6)$.

It follows from (4.7) and (4.9)–(4.13) that if z_k satisfies

$$\|z_k - z_*\| \le \min(\delta_1, \delta_2, \delta_3),$$

then (4.8) holds with

$$\gamma = c_5(c_4 + c_7 c_3^2 + c_1 c_3^2 \|e\|/n).$$

Inequality (4.8) implies that if $\{z_k\}$ converges to z_*, then it does so Q-quadratically. This proves the first statement.

Now we only need to prove the second statement—the convergence of $\{z_k\}$. Let

$$\delta = \min(\delta_1, \delta_2, \delta_3, r/\gamma)$$

for some $r \in (0,1)$. If $\|z_0 - z_*\| \leq \delta$, then

$$\|z_1 - z_*\| \leq \gamma \|z_0 - z_*\|^2 \leq r \|z_0 - z_*\|.$$

So $\|z_1 - z_*\| \leq r\delta \leq \delta$. Now we proceed by induction. This establishes the convergence of $\{z_k\}$ to z_*. □

In our numerical experimentation, we found that even for highly degenerate problems the observed convergence was effectively Q-quadratic until the iterates got too close to a solution and the singularity of the Jacobian matrix was encountered. This curious but pleasing phenomenon is the subject of further investigation.

5. Concluding remarks.

The rich structure present in the primal-dual formulation has led us to establish some rather strong convergence rate results.

No superlinear convergence results have been established so far for either primal or dual interior point algorithms. In fact, Gonzaga and Todd [6] showed that an algorithm that takes either primal or dual steps and reduces the Todd–Ye primal-dual potential function cannot have an R convergence rate greater than 1 (independent of n). Thus, from the viewpoint of convergence rate, our results suggest that primal-dual algorithms should be preferred to either primal or dual algorithms. Combined with the favorable numerical results obtained by a number of authors (Choi, Monma, and Shanno [3]; McShane, Monma, and Shanno [13]; and Lustig, Marsten, and Shanno [12]), this preference for primal-dual algorithms seems to be well founded.

We have shown that for the class of primal-dual algorithms studied, approximate centering should be viewed as a globalization strategy for Newton's method. Like other globalization strategies, it may improve the global behavior of the algorithm, but if not properly implemented, it will destroy fast local convergence. This fact lends credibility to the belief that polynomiality alone does not guarantee that local convergence rate properties have not been compromised or that the algorithm is necessarily fast. The algorithms of Kojima, Mizuno, and Yoshise [9]; Monteiro and Adler [17]; and Todd and Ye [21] possess polynomiality but cannot have fast Q-convergence.

Our preliminary numerical experimentation has shown that even without centering, the damped Newton algorithms that take steps close to the boundary of the positive orthant still have reasonable global behavior, although centering usually helps. This should not be totally unexpected since we are applying the damped Newton's method to a mildly nonlinear problem (see (4.2)).

One of the key components of this research is (3.7), which shows that in the damped Newton's method one can asymptotically make the step-length approach 1 at a rate that guarantees the fast convergence of Newton's method.

It seems to be difficult and costly, if at all possible, to ensure that the sequence $\{(x_k, y_k)\}$ converges to (x_*, y_*) along the central path. Therefore, it is our belief that at this stage the only viable strategy for designing a Q-superlinearly or Q-quadratically convergent primal-dual interior point algorithm is to phase out the centering step at the specified speed. The effect of degeneracy on the quadratic rate of convergence and the development of a quadratically convergent practical algorithm are the subjects of current research.

Acknowledgment. We gratefully acknowledge discussions with Florian Potra concerning the assumptions of Theorem 3.2. These discussions motivated us to establish a stronger version of the theorem than the one given in our preliminary version of this paper. We thank Alan Weiser for providing us with an experimental code for doing some of our numerical experiments. We also acknowledge Michael Todd and an anonymous referee for their helpful comments on this paper.

REFERENCES

[1] D. A. Bayer and J. C. Lagarias, *The nonlinear geometry of linear programming, Part* I: *Affine and projective scaling trajectories*, Trans. Amer. Math. Soc., 314 (1989), pp. 499–526.

[2] ———, *The nonlinear geometry of linear programming, Part* II: *Legendre transform coordinates*, Trans. Amer. Math. Soc., 314 (1989), pp. 527–581.

[3] I. C. Choi, C. L. Monma, and D. F. Shanno, *Further development of a primal dual interior point method*, ORSA J. Comput., 2 (1990), pp. 304–311.

[4] J. E. Dennis, Jr. and J. J. Moré, *A characterization of superlinear convergence and its application to quasi-Newton methods*, Math. Comp., 28 (1974), pp. 549–560.

[5] J. E. Dennis, Jr. and R. B. Schnabel, *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*, Prentice–Hall, Englewood Cliffs, NJ, 1983.

[6] C. C. Gonzaga and M. J. Todd, *An $O(\sqrt{n}L)$-iteration large-step primal-dual affine algorithm for linear programming*, Tech. Report 862, School of Operations Research and Industrial Engineering, Cornell University, Ithaca, NY, 1989; SIAM J. Optimization, 2 (1992), to appear.

[7] S. Huang and K. O. Kortanek, *A simultaneous primal- and dual-potential reduction algorithm for linear programming*, Working Paper Series 89–2, College of Business Administration, University of Iowa, Iowa City, IA, 1989.

[8] N. Karmarkar, *A new polynomial time algorithm for linear programming*, Combinatorica, 4 (1984), pp. 373–395.

[9] M. Kojima, S. Mizuno, and A. Yoshise, *A primal-dual interior point method for linear programming*, in Progress in Mathematical Programming, Interior-Point and Related Methods, N. Megiddo, ed., Springer-Verlag, New York, 1989, pp. 29–47.

[10] I. J. Lustig, *Feasibility issues in an interior point method for linear programming*, Math. Programming, 49 (1990/91), pp. 145–162.

[11] ———, *A generic primal-dual interior point algorithm*, Tech. Report SOR 88–3, School of Engineering and Applied Science, Department of Civil Engineering and Operations Research, Princeton University, Princeton, NJ, 1988.

[12] I. J. Lustig, R. E. Marsten, and D. F. Shanno, *Computational experience with a primal-dual interior point method for linear programming*, Linear Algebra Appl., 152 (1991), pp. 191–222.

[13] K. A. McShane, C. L. Monma, and D. F. Shanno, *An implementation of a primal-dual interior point method for linear programming*, ORSA J. Comput., 1 (1989), pp. 70–83.

[14] N. Megiddo, *Pathways to the optimal set in linear programming*, in Progress in Mathematical Programming, Interior-Point and Related Methods, N. Megiddo, ed., Springer-Verlag, New York, 1989, pp. 131–158.

[15] S. Mizuno, M. J. Todd, and Y. Ye, *Anticipated behavior of path-following algorithms for linear programming*, Tech. Report 878, School of Operations Research and Industrial Engineering, Cornell University, Ithaca, NY, 1989.

[16] ———, *Anticipated behavior of long-step algorithms for linear programming*, Tech. Report 882, School of Operations Research and Industrial Engineering, Cornell University, Ithaca, NY, 1990.

[17] R. C. Monteiro and I. Adler, *Interior path-following primal-dual algorithms. Part* I: *Linear programming*, Math. Programming, 44 (1989), pp. 27–41.

[18] G. Sonnevend, *An analytic center for polyhedrons and new classes of global algorithms for linear* (*smooth, convex*) *programming*, in Lecture Notes in Control and Information Science Vol. 84, A. Prekopa, ed., Springer-Verlag, Berlin, 1985, pp. 866–876.

[19] R. A. Tapia, *On the role of slack variables in quasi-Newton methods for constrained optimization*, in Numerical Optimization of Dynamic Systems, L. C. W. Dixon and G. P. Szegö, eds., North–Holland, Amsterdam, pp. 235–246, 1980.

[20] R. A. TAPIA AND Y. ZHANG, *An optimal-basis identification technique for interior-point linear programming algorithms*, Linear Algebra Appl., 152 (1991), pp. 343–363.

[21] M. J. TODD AND Y. YE, *A centered projective algorithm for linear programming*, Math. Oper. Res., 15 (1990), pp. 508–529.

[22] Y. ZHANG AND R. A. TAPIA, *A superlinearly convergent polynomial primal-dual interior-point algorithm for linear programming*, Tech. Report No. 90-40, Department of Mathematical Sciences, Rice University, Houston, TX, 1990; SIAM J. Optimization, 3 (1993), to appear.

SIAM J. OPTIMIZATION
Vol. 2, No. 2, pp. 325–347, May 1992

A ROBUST TRUST REGION METHOD FOR CONSTRAINED NONLINEAR PROGRAMMING PROBLEMS*

JAMES V. BURKE†

Abstract. Most of the published work on trust region algorithms for constrained optimization is derived from the original work of Fletcher on trust region algorithms for nondifferentiable exact penalty functions. These methods are restricted to applications where a reasonable estimate of the magnitude of an optimal Kuhn–Tucker multiplier vector can be given. More recently an effort has been made to extend the trust region methodology to the sequential quadratic programming (SQP) algorithm of Wilson, Han, and Powell. All of these extensions to the Wilson–Han–Powell SQP algorithm consider only the equality-constrained case and require strong global regularity hypotheses. This paper presents a general framework for trust region algorithms for constrained problems that does not require such regularity hypotheses and allows very general constraints. The approach is modeled on the one given by Powell for convex composite optimization problems and is driven by linear subproblems that yield viable estimates for the value of an exact penalty parameter. These results are applied to the Wilson–Han–Powell SQP algorithm and Fletcher's $S\ell_1QP$ algorithm. Local convergence results are also given.

Key words. trust regions, constrained optimization, exact penalty functions

AMS(MOS) subject classifications. 90C30, 65K05

1. Introduction.

Consider the constrained nonlinear programming problem

$$\begin{aligned}\mathcal{P}: &\ \text{minimize } f(x)\\ &\ \text{subject to } x \in \Omega,\end{aligned}$$

where $\Omega := \{x \in X : g(x) \in C\}$, $X \in \mathbb{R}^n$ and $C \subset \mathbb{R}^m$ are nonempty closed convex sets, and $f : \mathbb{R}^n \to \mathbb{R}$ and $g : \mathbb{R}^n \to \mathbb{R}^m$ are Frechet differentiable on an open set U containing X where the Frechet derivatives $f' : \mathbb{R}^n \to \mathbb{R}^n$ and $g' : \mathbb{R}^n \to \mathbb{R}^{m\times n}$ are bounded and continuous on X.

If $C = \mathbb{R}^s_- \times \{0\}_{\mathbb{R}^{m-s}}$ and

$$X := \{x \in \mathbb{R}^n : \underline{z}_i \le x_i \le \overline{z}_i, \quad i = 1, \cdots, n\}$$

where $\overline{z}_i$, $\underline{z}_i \in \mathbb{R} \cup \{\pm\infty\}$ for each $i = 1, \cdots, n$ with $\underline{z}_i \le \overline{z}_i$, $\underline{z}_i \ne +\infty$, and $\overline{z}_i \ne -\infty$ for $i = 1, \cdots, n$, then $\mathcal{P}$ is said to be in standard form. In general, the set X is considered to be some "simple" set of constraints so that the inclusion $x \in X$ is easily maintained.

In this paper we describe a framework for the development of robust trust region methods for solving $\mathcal{P}$. By "robust" we mean that the global convergence theory for these methods does not require assumptions concerning the regularity or the feasibility of $\mathcal{P}$. This is accomplished by designing the algorithm to locate stationary points for the problem

$$\begin{aligned}\widehat{\mathcal{P}}: &\ \text{minimize } f(x)\\ &\ \text{subject to } x \in \arg\min\{\overline{\varphi}(x) : x \in \mathbb{R}^n\},\end{aligned}$$

*Received by the editors December 5, 1989; accepted for publication October 31, 1991.

†Department of Mathematics, GN-50, University of Washington, Seattle, Washington 98195.
This work was supported in part by National Science Foundation grant DMS-9102059.

where

$$\arg\min\{\overline{\varphi}(x) : x \in \mathbb{R}^n\} := \{\overline{x} \in \mathbb{R}^n : \overline{\varphi}(\overline{x}) = \min\{\overline{\varphi}(x) : x \in \mathbb{R}^n\}\},$$

and

$$\overline{\varphi}(x) := \operatorname{dist}(g(x)|C) + \psi(x|X) \tag{1.1}$$

with

$$\operatorname{dist}(y|C) := \inf\{\|y - z\| : z \in C\} \tag{1.2}$$

and

$$\psi(x|X) := \begin{cases} 0 & \text{if } x \in X, \\ +\infty & \text{if } x \notin X \end{cases} \tag{1.3}$$

(here and throughout, the symbols $\|\cdot\|$ denote a given norm on $\mathbb{R}^n$ or $\mathbb{R}^m$). Clearly, if $\mathcal{P}$ is feasible, then $\mathcal{P}$ and $\widehat{\mathcal{P}}$ are equivalent. On the other hand, if $\mathcal{P}$ is not feasible, then further information about $\mathcal{P}$ can be obtained by studying $\widehat{\mathcal{P}}$. In [1], Burke introduces a notion of stationarity for $\widehat{\mathcal{P}}$ which will be reviewed in the next section. Burke [1] also discusses an algorithm for locating points that are stationary for $\widehat{\mathcal{P}}$. This algorithm extends the well-known SQP method of Wilson [28], Han [13], and Powell [17]. The plan of this paper is to extend the techniques of [1] to the trust region framework and then to apply these results to both the $S\ell_1QP$ algorithm of Fletcher [11], [12] and to a trust region implementation of the Wilson–Han–Powell SQP method.

Many other authors have considered trust region algorithms for constrained optimization. One can broadly classify this work into three categories: (1) methods for linear constraints, (2) methods for nonlinear equality constraints, and (3) exact penalization methods. The first class of methods is studied in Conn, Gould, and Toint [9]; Toint [26]; Moré [14]; and Burke, Moré, and Toraldo [5]. This class of methods corresponds to the case of $\mathcal{P}$ with the functional constraint $g(x) \in C$ absent, and is based on projected gradient techniques. The second class of methods concentrates on the instance of $\mathcal{P}$ where $C = \{0\}_{\mathbb{R}^m}$ and $X = \mathbb{R}^n$ and these methods can be viewed as extensions to the Wilson–Han–Powell SQP method. These methods are studied in Celis, Dennis, and Tapia [7]; Vardi [27]; Byrd, Schnabel, and Shultz [6]; and Powell and Yuan [18]. All of these papers require $g'(x)$ to be of full rank on $\mathbb{R}^n$. Under this hypothesis, the method of Celis, Dennis, and Tapia [7] has recently been provided with a convergence theory by El-Alem [10]. The methods of Vardi [27] and Byrd, Schnabel, and Shultz [6] obtain the feasibility of the modified constraint region by including an additional parameter $\eta \in [0, 1]$ in the constraint

$$\eta g(x) + g'(x)s = 0. \tag{1.4}$$

Unfortunately, there are many examples that defeat this trick. For instance, if one takes

$$g(x) := \begin{bmatrix} 1 - e^x \\ x \end{bmatrix} \tag{1.5}$$

with $g : \mathbb{R} \to \mathbb{R}^2$, then $\eta = 0, s = 0$ is the unique solution to (1.4) for all $x \in \mathbb{R}$. The difficulty here is that $g'(x)$ never has full rank. The method introduced in §5 has no difficulty with this example. The method proposed by Powell and Yuan [18] has a flavor that is similar to the approach suggested here for the case $C = \{0\}_{\mathbb{R}^m}$ and $X = \mathbb{R}^n$, but there remain fundamental differences.

There is a large body of work directly associated with the third class of algorithms, exact penalization methods [11], [12], [16], [29], [30], etc. Most of this literature is couched in the language of trust region algorithms for convex composite optimization and is based on the original work of Fletcher. In the context of problem $\mathcal{P}$ all of these methods implicitly require knowledge of an upper bound on the norm of some Kuhn–Tucker multiplier at a Kuhn–Tucker solution to $\mathcal{P}$. They also require that the procedure be initiated close enough to this Kuhn–Tucker solution. One of the fruits of this investigation is a modification of these methods that eliminates the need for hypotheses of this type in the global convergence theory.

We now describe the plan of the paper. In §2, we present the basic algorithm. In §3, the stationarity conditions for $\widehat{\mathcal{P}}$ given in [1] are recalled. In §4 the basic properties of the objects employed in the description of the algorithm are given and the convergence analysis is presented in §5. The application of these results to SQP and $S\ell_1QP$ are given in §§6 and 7, respectively.

The notation that we employ is standard. Nonetheless, a partial listing is given for the readers convenience. Given $x, y \in \mathbb{R}^k$ the inner product is denoted by

$$\langle x, y\rangle := x^T y := \sum_{i=1}^{k} x^i y^i,$$

where $x := (x^1, x^2, \cdots, x^k)^T$ and $y := (y^1, y^2, \cdots, y^k)^T$. If X and Y are subsets of $\mathbb{R}^k$, then

$$\alpha X + \beta Y := \{\alpha x + \beta y : x \in X, y \in Y\}.$$

The polar of X is defined as

$$X^0 := \{w \in \mathbb{R}^k : \langle w, x\rangle \leq 1 \text{ for all } x \in X\}.$$

If X is convex, that is, $\lambda x + (1 - \lambda)y \in X$ for all $x, y \in X$ and $\lambda \in [0, 1]$, then the recession cone of X is defined as

$$\operatorname{rec}(X) := \{y \in \mathbb{R}^k : X + y \subset \operatorname{cl}(X)\}$$

where $\operatorname{cl}(X)$ is the closure of X. The normal cone to X at any point $\overline{x} \in X$ is defined by

$$N(\overline{x}|X) := \{w \in \mathbb{R}^k : \langle w, x - \overline{x}\rangle \leq 0 \text{ for all } x \in X\}.$$

The tangent cone to X at $\overline{x}$ is the polar of the normal cone,

$$T(\overline{x}|X) := N(\overline{x}|X)^0.$$

The support and convex indicator functions for X are given, respectively, by

$$\psi^*(w|X) := \sup\{\langle w, x\rangle : x \in X\}$$

and

$$\psi(x|X) := \begin{cases} +\infty, & \text{if } x \notin X, \\ 0, & \text{if } x \in X. \end{cases}$$

A norm on $\mathbb{R}^k$ is denoted by $\|x\|$ and its unit ball is designated by

$$\mathbb{B} := \{x : \|x\| \le 1\}.$$

The dual norm to $\|x\|$ is given by

$$\|x\|_{\mathrm{o}} := \psi^*(x|\mathbb{B})$$

and consequently the dual unit ball is $\mathbb{B}^0$. The two-norm plays a special role and it is denoted by $\|x\|_2 := (\langle x, x\rangle)^{1/2}$. The distance function for the set X associated with the norms $\|\cdot\|$ and $\|\cdot\|_{\mathrm{o}}$ are given by

$$\mathrm{dist}(y|X) := \inf\{\|y - x\| : x \in X\}$$

and

$$\mathrm{dist}_{\mathrm{o}}(y|X) := \inf\{\|y - x\|_{\mathrm{o}} : x \in X\},$$

respectively. Given $g : \mathbb{R}^n \longrightarrow \mathbb{R}^m$ the Frechet derivative of g at a point $x \in \mathbb{R}^n$, if it exists, is the linear mapping $g'(x) : \mathbb{R}^n \longrightarrow \mathbb{R}^m$ (if it exists, it is unique) for which

$$g(y) = g(x) + g'(x)(y - x) + o(\|y - x\|), \quad \text{where } \lim_{y \to x} \frac{o(\|y - x\|)}{\|y - x\|} = 0.$$

Since $g'(x)$ is a linear mapping from $\mathbb{R}^n$ to $\mathbb{R}^m$, it has a matrix representation in $\mathbb{R}^{m\times n}$, with respect to the standard basis. This representation is called the Jacobian of g at x. In this presentation, we identify $g'(x)$ with its Jacobian. Also, for a set $X \subset \mathbb{R}^n$ and a mapping $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ we define

$$\arg\min\{f(x) : x \in X\} := \{\overline{x} \in X : f(\overline{x}) := \min\{f(x) : x \in X\}\}.$$

The set $\arg\max\{f(x) : x \in X\}$ is defined similarly.

2. The model algorithm. As in [1] our approach is based on a type of "linearization" of the constraint region Ω. Given $x \in X$, $0 \le \rho_1 \le \rho_2$, and $\theta \in [0, 1]$ we define

$$L\Omega(x, \rho_1, \rho_2, \theta) := \{s \in [X - x] \cap \rho_2\mathbb{B}^n |\ g(x) + g'(x)s \in C + \nu(x, \rho_1, \theta)\mathbb{B}^m\}, \tag{2.1}$$

where $\mathbb{B}^n$ and $\mathbb{B}^m$ are the closed unit balls of the norms that are given for $\mathbb{R}^n$ and $\mathbb{R}^m$, respectively, and for $\tau \ge 0$

$$\nu(x, \tau, \theta) := \varphi(x, 0) + \theta[\varphi(x, \tau) - \varphi(x, 0)], \tag{2.2}$$

and

$$\varphi(x, \tau) := \inf\{\mathrm{dist}(g(x) + g'(x)s|C)|s \in [X - x] \cap \tau\mathbb{B}^n\} \tag{2.3}$$

(henceforth the symbol $\mathbb{B}$ is used to denote the unit ball of either $\mathbb{R}^n$ or $\mathbb{R}^m$ unless some ambiguity is possible). We refer to the multifunction $L\Omega : X \times T \times [0, 1] \rightrightarrows \mathbb{R}^n$, where

$$T := \{(\rho_1, \rho_2) : 0 \le \rho_1 \le \rho_2\},$$

as a "linearization" of Ω. Such linearizations are well studied in the literature [2], [19], [21], [22]. Given $(x,(\rho_1,\rho_2),\theta)\in X\times T\times[0,1]$, the set $L\Omega(x,\rho_1,\rho_2,\theta)$ is a nonempty compact convex subset of $\mathbb{R}^n$. Moreover, if

$$(\rho_1,\rho_2,\theta)\in \operatorname{int}(T\times[0,1]) \quad\text{and}\quad \varphi(x,\rho_1)\neq\varphi(x,0),$$

then $\operatorname{int}[L\Omega(x,\rho_1,\rho_2,\theta)]\neq\emptyset$ (the notation $\operatorname{int}(S)$ means the interior of the set S). This observation is significant since we use $L\Omega(x,\rho_1,\rho_2,\theta)$ as the constraint region for our convex programming subproblems. The condition $\operatorname{int}[L\Omega(x,\rho_1,\rho_2,\theta)]\neq\emptyset$ implies that the Slater constraint qualification [25] is satisfied and so these convex programming subproblems have Kuhn–Tucker multipliers [23] at their solution.

The condition $\varphi(x,\rho_1)\neq\varphi(x,0)$ is of particular significance in the construction of the multifunction $L\Omega$. In [3] it is shown that if $\tau>0$, then $\varphi(x,\tau)=\varphi(x,0)$ if and only if x is a stationary point for the function $\overline{\varphi}$ defined in (1.1) (see §3). Moreover, given $(x,\rho_1,\rho_2,\theta)\in X\times T\times[0,1]$, it is shown that the inequality

$$\overline{\varphi}'(x;s)\leq\theta[\varphi(x,\rho_1)-\varphi(x,0)]\leq 0$$

holds for every $s\in L\Omega(x,\rho_1,\rho_2,\theta)$ where

$$\overline{\varphi}'(x;s):=\lim_{t\downarrow 0}\frac{\overline{\varphi}(x+ts)-\overline{\varphi}(x)}{t}$$

is the usual directional derivative of $\overline{\varphi}$ at x in the direction s. Consequently, if $\theta\neq 0$ and x is not a stationary point for $\overline{\varphi}$, then $L\Omega(x,\rho_1,\rho_2,\theta)$ is contained in the set of directions of strict descent for $\overline{\varphi}$ at x. This relationship supports the goal of locating stationary points for $\widehat{\mathcal{P}}$.

If $\mathcal{P}$ is in standard form and the norms chosen for $\mathbb{R}^n$ and $\mathbb{R}^m$ are polyhedral, then $L\Omega(x,\rho_1,\rho_2,\theta)$ is always a polyhedral convex set and the computation of the value $\varphi(x,\rho_1)$ reduces to solving a linear program. Thus, in this case, the set $L\Omega(x,\rho_1,\rho_2,\theta)$ can be specified in finite time.

In order to develop a local convergence theory, it is important that the set $L\Omega(x,\rho_1,\rho_2,\theta)$ closely resemble the constraint region in the standard SQP algorithm whenever possible. For example, if $x\in X$ is such that $\varphi(x,\rho_1)=0$, we would like to set $\theta=1$, since then

$$L\Omega(x,\rho_1,\rho_2,1):=\{s\in[X-x]\cap\rho_2\mathbb{B}\,|\;g(x)+g'(x)s\in C\}.$$

If $\mathcal{P}$ is in standard form and the norms on $\mathbb{R}^n$ and $\mathbb{R}^m$ are polyhedral, then this is indeed possible. However, in general, such a choice of θ is not theoretically sound. Nonetheless, we can choose θ as a function of x so that $\theta(x)\to 1$ as $[\varphi(x,\rho_1)-\varphi(x,0)]\to 0$. Specifically, given $\theta_0>0$ we consider functions $\theta:X\to[\theta_0,1]$ such that if any one of the sets $C,X,\mathbb{B}^n$, or $\mathbb{B}^m$ is not polyhedral, then

$$\theta(x)=1\quad\text{only if } \varphi(x,\rho_1)=\varphi(x,0). \tag{2.4}$$

Two examples of such functions are

$$\theta_1(x):=\theta_0\qquad\text{for all } x\in X \tag{2.5}$$

and

$$\theta_2(x):=\max\{\theta_0,1+[\varphi(x,\rho_1)-\varphi(x,0)]\}. \tag{2.6}$$

The structure of the trust region algorithms that we discuss is standard, and is modeled on the one given by Powell [16] for convex composite functions. There are also similarities to Fletcher's $S\ell_1QP$ method [11], [12]. In particular, the acceptance of the trial step s_k at the kth iteration depends on the quadratic approximation

$$(2.7)\quad P_\alpha(s;x,H) := f(x)+\nabla f(x)^T s+\tfrac{1}{2}s^T Hs+\alpha\ \mathrm{dist}(g(x)+g'(x)s|C)+\psi(x+s|X)$$

to the exact penalty function

$$(2.8)\qquad P_\alpha(x) := f(x)+\alpha\overline{\varphi}(x)$$

for $\mathcal{P}$. As usual, the matrix $H \in \mathbb{R}^{n\times n}$ is intended to approximate the Hessian of the Lagrangian. The trial step s_k is chosen so that the reduction in $P_\alpha(s;x,H)$ is comparable to that which could be obtained by choosing the step that optimizes a linear model of $\mathcal{P}$ at x_k. In the case of constrained optimization, a typical linear model considered by Powell [16] is $P_\alpha(s;x,0)$ for some prespecified $\alpha > 0$. The linear model that we use is given by

$$\begin{aligned} LP(x) :\ & \text{minimize } f(x)+f'(x)s \\ & \text{subject to } s \in L\Omega(x,\rho_1,\rho_2,\theta(x)) \end{aligned}$$

for a fixed choice of $(\rho_1,\rho_2) \in \operatorname{int}(T)$.

The subproblems $LP(x)$ are also used to obtain updates for the penalty parameter α_k. The update rule is similar to the one proposed by Han in [13];

$$(2.9)\qquad \alpha_k := \max\{\|y_k\|_0+\varepsilon,\alpha_{k-1}+4\varepsilon\},$$

where $y_k \in \mathbb{R}^m$ is any Kuhn–Tucker multiplier vector for the constraint

$$(2.10)\qquad g(x_k)+g'(x_k)s \in C+\nu(x_k,\rho_1,\theta(x_k))\mathbb{B}$$

in $LP(x_k)$, where $\|\cdot\|_0$ is the norm dual to the norm $\|\cdot\|$ (i.e., $\|y\|_0 := \sup\{z^T y : z \in \mathbb{B}\}$) and $\varepsilon > 0$. From Burke [1], [3], this set of Kuhn–Tucker multipliers is given by

$$KTM(x) := \{y|(s,y,w,z) \in KT(x) \text{ for some } s,w,z \in \mathbb{R}^m\},$$

where

$$KT(x) := \left\{ (s,y,w,z) \;\middle|\; \begin{array}{l} s \in L\Omega(x,\rho_1,\rho_2,\theta(x)), w \in N(x+s|X), \\ z \in N(s|\rho_2\mathbb{B}), y \in N(g(x)+g'(x)s|C+\nu(x,\rho_1,\theta(x))\mathbb{B}), \\ 0 = f'(x)^T+g'(x)^T y+w+z \end{array} \right\}$$

is the multifunction of Kuhn–Tucker solutions to $LP(x)$. In general, the α_k's can be updated by any rule such that $\|y_k\|_0 \le \alpha_k$ for all $k = 1,2,\cdots$, and α_k is updated infinitely many times if and only if $\sup\{\|y_k\|_0 : k = 1,2,\cdots\} = +\infty$. Having α_k, the trial step s_k is accepted if

$$(2.11)\qquad P_{\alpha_k}(x_k+s_k)-P_{\alpha_k}(x_k) \le \beta_1\Delta P_{\alpha_k}(s_k;x_k,H_k),$$

where $0 < \beta_1 < 1$ and

$$(2.12)\qquad \Delta P_{\alpha_k}(s_k;x_k,H_k) := P_{\alpha_k}(s_k;x_k,H_k)-P_{\alpha_k}(x_k).$$

A detailed description of the algorithm follows.

Initialization: Choose $x_0 \in X, H_0 \in \mathbb{R}^{n\times n}, \alpha_{-1} > 0, \varepsilon > 0$, $t_0 \in (0,1)$, $0 < \gamma_1 \le \gamma_2 < 1 \le \gamma_3, 0 < \beta_1 \le \beta_2 < \beta_3 \le 1$. Set $k = 0$.

Step 1. If $KTM(x_k) = \emptyset$, set $\alpha_k := \alpha_{k-1}$; otherwise choose $y_k \in KTM(x_k)$ and set

$$\alpha_k := \begin{cases} \alpha_{k-1}, & \text{if } \alpha_{k-1} \ge \|y_k\|_0 + \varepsilon, \\ \max\{\|y_k\|_0 + \varepsilon, \alpha_{k-1} + 4\varepsilon\}, & \text{otherwise.} \end{cases}$$

Step 2. Choose $\widehat{s}_k, s_k \in [X - x_k] \cap t_k\rho_2\mathbb{B}$ with $\Delta P_{\alpha_k}(s_k; X_k, H_k) < 0$. If no such s_k exists, then stop.

Step 3. Set $r_k := [P_{\alpha_k}(x_k + s_k) - P_{\alpha_k}(x_k)][\Delta P_{\alpha_k}(s_k; x_k, H_k)]^{-1}$ and $\widehat{r}_k := [P_{\alpha_k}(x_k + \widehat{s}_k) - P_{\alpha_k}(x_k)][\Delta P_{\alpha_k}(s_k; x_k, H_k)]^{-1}$. If $r_k \le \widehat{r}_k$, reset $r_k := \widehat{r}_k$ and $s_k := \widehat{s}_k$. If $r_k \ge \beta_3$, choose $t_{k+1} \in [t_k, \min\{1, \gamma_3 t_k\}]$; if $\beta_2 \le r_k < \beta_3$, set $t_{k+1} := t_k$; if $r_k < \beta_2$, choose $t_{k+1} \in [\gamma_1 t_k, \gamma_2 t_k]$.

Step 4. If $r_k < \beta_1$, set $x_{k+1} := x_k, \alpha_{k+1} := \alpha_k$, $k := k + 1$, and return to Step 2.

Step 5. Choose $H_{k+1} \in \mathbb{R}^{n\times n}$, set $x_{k+1} := x_k + s_k$, $k := k + 1$, and return to Step 1.

Remarks. (1) The alternate trial step $\widehat{s}_k$ in Step 2 of the algorithm is introduced to facilitate the discussion of second-order corrections in §§6 and 7. It will be shown that one may always take $s_k := t_k\widetilde{s}_k$ where $\widetilde{s}_k$ solves $LP(x_k)$ and then set $\widehat{s}_k := s_k$.

(2) The updating formula for the penalty parameter depends upon the knowledge of a dual solution to $LP(x_k)$. This linear subproblem has a fixed trust region radius that could be adjusted finitely many times without affecting the global convergence behavior of the procedure. Nonetheless, it would seem to be more natural, if not more efficient, to let the trust region radius of this subproblem be the same as in the choice of trial step s_k. Unfortunately, our proof theory does not allow such a variation. In particular, if the trust region radius in $LP(x_k)$ is allowed to vary, then we are unable to provide a satisfactory analysis of the cases where the sequence $\{t_k\}$ is not bounded away from zero.

(3) The function $\theta(x)$ is introduced primarily for considerations associated with local convergence and to simplify adjustments in the trust region radius. The ability to adjust the trust region radius in this way follows from the inclusion

$$L\Omega(x, t\rho_1, t\rho_2, \theta) \subset L\Omega(x, \rho_1, t\rho_2, t\theta),$$

to be established in Proposition 4.1. In the polyhedral case the function $\theta(x)$ can also be used to reduce the effort required to obtain y_k in Step 1 whenever $\varphi(x_k, \rho_1) \neq \varphi(x_k, 0) \neq 0$. This is done by implicitly defining $\theta(x)$ in terms of the algorithm used to evaluate $\varphi(x, \rho_1)$ and any other function $\widehat{\theta}(x)$ satisfying (2.4). The algorithm for evaluating $\varphi(x, \rho_1)$ should produce a sequence $\{(\lambda_i, \widehat{s}_k)\} \in \mathbb{R} \times ([X - x] \cap \rho_1\mathbb{B})$ such that

$$\operatorname{dist}(g(x) + g'(x)\widehat{s}_k | C) \downarrow \varphi(x, \rho_1)$$

and

$$\lambda_i \uparrow \varphi(x, \rho_1).$$

One then terminates the procedure when

$$\operatorname{dist}[g(x) + g'(x)s_k | C] - \varphi(x, 0) \le \widehat{\theta}(x)[\lambda_i - \varphi(x, 0)]$$

(which must occur after a finite number of iterations i if $\varphi(x,\rho) \neq \varphi(x,0)$) and define

$$\theta(x) := \begin{cases} \frac{\operatorname{dist}[g(x)+g'(x)\widehat{s}_i|C]-\varphi(x,0)}{\varphi(x,\rho_1)-\varphi(x,0)}, & \text{if } \varphi(x,\rho_1) \neq \varphi(x,0), \\ 1, & \text{otherwise.} \end{cases}$$

In this case

$$\widehat{\theta}(x) \leq \theta(x) \leq 1$$

and

$$\theta(x)[\varphi(x,\rho_1) - \varphi(x)] = \operatorname{dist}[g(x) + g'(x)\widehat{s}_k|C] - \varphi(x,0)$$

so that $\varphi(x,\rho_1)$ need not be computed except when $\varphi(x,\rho_1) = \varphi(x,0) \neq 0$.

(4) There are many ways to update the penalty parameter α_k in order to guarantee the existence of a trial step s_k so that $\Delta P_{\alpha_k}(s_k; x_k, H_k) < 0$, however, not all of these methods guarantee the inequality

$$\alpha_k \geq \operatorname{dist}_0(0|KTM(x_k)).$$

Our proof of convergence requires this inequality since we need to invoke Proposition 4.2(2) when $\{\alpha_k\}$ is bounded.

(5) As described above the sequence of penalty parameters $\{\alpha_i\}$ is necessarily nondecreasing. However, one can employ a clever device proposed by Sahba [24] for reducing the penalty parameter on certain iterations. Specifically, at the end of the kth iteration one evaluates

$$\widetilde{\varphi}_k := \min\{\widetilde{\varphi}_{k-1}, \varphi(x_k, 0)\}.$$

If $\widetilde{\varphi}_k \leq \widetilde{\varphi}_{k-1} - \widetilde{\varepsilon}$ for some prespecified $\widetilde{\varepsilon} > 0$, then one resets α_{i+1} to any positive real number, say,

$$\alpha_{i+1} := \|y_i\|_0 + \varepsilon.$$

Clearly this reinitialization of α_i can only occur a finite number of times. Hence the convergence analysis remains unaltered.

(6) In the case where C and X are polyhedral and the norms on $\mathbb{R}^n$ and $\mathbb{R}^n$ are polyhedral, then $LP(x)$ is a linear program and the evaluation of $\varphi(x,\rho)$ reduces to solving a linear program.

We now proceed to the analysis of the algorithm. The first step in this process is to describe the first-order necessary conditions for optimality in $\widehat{\mathcal{P}}$.

3. Stationarity conditions for $\mathcal{P}$. We say that a point $x \in X$ is a stationary point for $\mathcal{P}$ if it is a stationary point for $\widehat{\mathcal{P}}$. By this we mean that $\overline{x}$ satisfies first-order necessary conditions for optimality in both of the problems

$$\underset{x\in\mathbb{R}^n}{\text{minimize}}\ \overline{\varphi}(x) \tag{3.1}$$

and

$$\begin{aligned} &\text{minimize } f(x) \\ &\text{subject to } x \in X \quad \text{and} \quad g(x) \in C + \overline{\varphi}(\overline{x})\mathbb{B}. \end{aligned} \tag{3.2}$$

It is shown in [1, §2] that these conditions can be expressed in terms of the multifunctions

$$M_1(x) := \left\{ \begin{pmatrix} y \\ w \end{pmatrix} \middle| \begin{array}{l} y \in N(g(x)|C + \overline{\varphi}(x)\mathbb{B}), w \in N(x|X), \\ 0 = f'(x)^T + g'(x)^T y + w \end{array} \right\}$$

and

$$M_0(x) := \left\{ \begin{pmatrix} y \\ w \end{pmatrix} \middle| \begin{array}{l} y \in N(g(x)|C + \overline{\varphi}(x)\mathbb{B}), w \in N(x|X), \\ 0 = g'(x)^T y + w, \end{array} \right\}$$

where $x \in X$.

THEOREM 3.1. (Burke [1, §2].) *Let $\overline{x} \in X$.*

(1) *If $\overline{x}$ is a stationary point for $\overline{\varphi}$, then either $M_0(\overline{x}) \neq \{0\}$ or $g(\overline{x}) \in C$, or both. Moreover, if $\overline{\varphi}(\overline{x}) \neq 0$, then $M_0(\overline{x}) \neq \{0\}$ if and only if $\varphi(\overline{x}, \rho) = \varphi(x, 0)$ for every $\rho > 0$.*

(2) *If $\overline{x}$ is a stationary point for* (3.2), *then either $M_1(x) \neq \emptyset$ or $M_0(x) \neq \{0\}$, or both.*

In Clarke's [8] terminology the sets $M_1(x)$ and $M_0(x)$ are called the normal and abnormal multipliers for (3.2) at $x \in X$. We will call $M_1(x)$ the set of Kuhn–Tucker multipliers for (3.2) at $x \in X$ and $M_0(x)$ the set of Fritz John multipliers for (3.2) at $\overline{x} \in X$. If $\overline{x}$ is such that $\overline{\varphi}(\overline{x}) = 0$ and $\mathcal{P}$ is in standard form, then $M_1(\overline{x})$ is precisely the set of Kuhn–Tucker multipliers for $\mathcal{P}$ that one normally encounters in mathematical programming. A point $\overline{x} \in X$ is called a Kuhn–Tucker point for $\mathcal{P}$ if $\overline{\varphi}(\overline{x}) = 0$ and $M_1(\overline{x}) \neq 0$; it is called a Fritz John point for $\mathcal{P}$ if $\overline{\varphi}(\overline{x}) = 0$ and $M_0(\overline{x}) \neq \{0\}$; and it is called a nonfeasible stationary point for $\mathcal{P}$ if $\overline{\varphi}(\overline{x}) \neq 0$ and $M_0(\overline{x}) \neq \{0\}$. Any point that is either a Kuhn–Tucker point, a Fritz John point, or a nonfeasible stationary point for $\mathcal{P}$ is simply called a stationary point for $\mathcal{P}$.

We conclude this section by recalling certain elementary facts concerning the distance function $\mathrm{dist}(y|C)$, the support function $\psi^*(y|C)$, and normal cones that are used in our study. For the proofs of these facts we refer the reader to [3] and [23].

LEMMA 3.2. *Let K be a nonempty closed convex subset of $\mathbb{R}^q$.*

(1) *The distance function*

$$\mathrm{dist}(y|K) := \inf\{\|y - z\| : z \in K\}$$

is convex on $\mathbb{R}^q$ with convex subdifferential

$$\partial\, \mathrm{dist}(x|K) := \begin{cases} \mathbb{B}^{\circ} \cap N(x|K), & \text{if } y \in K, \\ (bdry\mathbb{B}^{\circ}) \cap N(x|K + \mathrm{dist}(y|K)\mathbb{B}), & \text{if } y \notin K. \end{cases}$$

Consequently, $\mathrm{dist}(\cdot|K)$ *is globally Lipschitz continuous on $\mathbb{R}^q$ with Lipschitz constant of* 1.

(2) *If $x \in K$, then $w \in N(x|K)$ if and only if*

$$\langle w, x\rangle = \psi^*(w|K).$$

(3) *For any $x \in \mathbb{R}^n$ and $w \in \mathbb{R}^q$, it is always the case that*

$$\langle w, x\rangle - \psi^*(w|K) \leq \|w\|_{\circ}\, \mathrm{dist}(x|K).$$

4. The linear subproblem $LP(x)$.

We begin this section with a description of the properties of the linearization $L\Omega$.

PROPOSITION 4.1. *Let $x_1, x_2 \in X$, $0 \leq \rho_1 \leq \rho_2$, $0 \leq \overline{\rho}_1 \leq \overline{\rho}_2$, and θ_1, θ_2, $t, \sigma \in [0, 1]$, and suppose that $M > 0$ is a bound for f' and g' on X.*

(1) *If $s \in [X - x_1] \cap \rho_1\mathbb{B}$, then*

$$\mathrm{dist}[s|[X - x_2] \cap \rho_1\mathbb{B}] \leq 2\|x_1 - x_2\|.$$

(2) $|\varphi(x_1,\rho_1)-\varphi(x_1,0)|\le M\rho_1$.
(3) $\varphi(x_1,\cdot):\mathbb{R}_+\to\mathbb{R}$ *is a convex function.*
(4) $|\varphi(x_1,\rho_1)-\varphi(x_2,\rho_1)|\le 3M\|x_1-x_2\|+\rho_1\|g'(x_1)-g'(x_2)\|$.
(5) $\nu(x_1,\cdot,\theta_1):\mathbb{R}_+\to\mathbb{R}$ *is a convex function and so*

$$\nu(x_1,t\rho_1,\theta_1)\le\nu(x_1,\rho_1,t\theta_1).$$

(6) $|\nu(x_1,\rho_1,\theta_1)-\nu(x_2,\rho_1,\theta_2)|\le 5M\|x_1-x_2\|+M\rho_1|\theta_1-\theta_2|+\rho_1\|g'(x_1)-g'(x_2)\|$.

(7) $L\Omega(x_1,t\rho_1,t\rho_2,\theta_1)\subset L\Omega(x_1,\rho_1,t\rho_2,t\theta_1)$.

(8) $tL\Omega(x,\rho_1,\rho_2,\theta_1)+(1-t)L\Omega(x,\rho_1,\overline{\rho}_2,\theta_2)\subset L\Omega(x,\rho_1,t\rho_2+(1-t)\overline{\rho}_2,t\theta_1+(1-t)\theta_2)$.

(9) *The multifunction* $L\Omega$ *is upper semicontinuous on* $X\times T\times[0,1]$.

(10) *If* $x\in X$ *is such that* $M_0(x)=\{0\}$ *and* $(\rho_1,\rho_2,\theta)\in\operatorname{int}[T\times[0,1]]$, *then the multifunction* $L\Omega$ *is continuous near* (x,ρ_1,ρ_2,θ) *relative to* $X\times T\times[0,1]$.

Proof. (1) If $x_2=x_1+s$ we are done since $0\in(X-x_2)\cap\rho_1\mathbb{B}$ and $\|s-0\|=\|x_1-x_2\|$. If $x_2\ne x_1+s$ choose $\lambda>0$ so that $\lambda\|x_1+s-x_2\|=\rho_1$. If $\lambda\ge1$, then $\widehat{s}:=(x_1+s)-x_2\in[X-x_2]\cap\rho_1\mathbb{B}$ and $\|s-\widehat{s}\|=\|x_1-x_2\|$, from which the result follows. If $\lambda<1$, then again $\widehat{s}:=\lambda[x_1+s-x_2]\in[X-x_2]\cap\rho_1\mathbb{B}$, since $x_2+\lambda[x_1+s-x_2]=\lambda(x_1+s)+(1-\lambda)x_2\in X$. Moreover,

$$\begin{aligned}\|s-\widehat{s}\|&\le\|x_1-x_2\|+(1-\lambda)\|x_1+s-x_2\|\\&=\|x_1-x_2\|+\|x_1+s-x_2\|-\rho_1\\&\le 2\|x_1-x_2\|+\|s\|-\rho_1\\&\le 2\|x_1-x_2\|.\end{aligned}$$

(2) Let $s\in[X-x_1]\cap\rho_1\mathbb{B}$ be such that

$$\varphi(x_1,\rho_1)=\operatorname{dist}(g(x_1)+g'(x_1)s|C).$$

Then, by Lemma 3.2,

$$|\varphi(x_1,\rho_1)-\varphi(x_1,0)|=\operatorname{dist}(g(x_1)+g'(x_1)s|C)-\operatorname{dist}(g(x_1)|C)\le\|g'(x_1)\|\rho_1.$$

(3) This follows immediately from the fact that

$$\lambda[X-x_1]\cap\rho_1\mathbb{B}+(1-\lambda)[X-x_1]\cap\rho_2\mathbb{B}\subset[X-x_1]\cap(\lambda\rho_1+(1-\lambda)\rho_2)\mathbb{B}.$$

(4) Let $s_2\in[X-x_2]\cap\rho_1\mathbb{B}$ be such that

$$\varphi(x_2,\rho_1)=\operatorname{dist}[g(x_2)+g'(x_2)s_2|C]$$

and let $\widehat{s}_2\in[X-x_1]\cap\rho_1\mathbb{B}$ be such that

$$\|s_2-\widehat{s}_2\|=\operatorname{dist}[s_2|[X-x_1]\cap\rho_1\mathbb{B}].$$

Then, by (1) of the proof and Lemma 3.2(1),

$$\begin{aligned}\varphi(x_1,\rho_1)&\le\operatorname{dist}[g(x_1)+g'(x_1)\widehat{s}_2|C]\\&\le\|g(x_1)-g(x_2)\|+\|g'(x_1)-g'(x_2)\|\|\widehat{s}_2\|\\&\qquad+\|g'(x_2)\|\|s_2-\widehat{s}_2\|+\varphi(x_2,\rho_1)\\&\le 3M\|x_1-x_2\|+\rho_1\|g'(x_1)-g'(x_2)\|+\varphi(x_2,\rho_1).\end{aligned}$$

The result now follows by symmetry.

(5) This follows immediately from (3).

(6) This follows immediately from Lemma 3.2(1) and parts (2) and (4) above.

(7) This follows directly from the inequality $\nu(x_1, t_1\rho_1, \theta_1) \leq \nu(x_1, \rho_1, t_1\theta_1)$ in part (5).

(8) This follows from part (5), the convexity of the sets C and X, and the fact that $\eta_1\mathbb{B} + \eta_2\mathbb{B} = (\eta_1 + \eta_2)\mathbb{B}$ for every $\eta_1, \eta_2 \geq 0$.

(9) This follows directly by continuity.

(10) This is established in Burke [1, Thm. 9.3]. □

Let $0 < \rho_1 < \rho_2$ be fixed throughout the remainder of the paper. Also let $\theta : X \to [\theta_0, 1]$ be given so that (2.4) is satisfied unless all of the sets $X, C, \mathbb{B}^n$, and $\mathbb{B}^m$ are polyhedral. Moreover, we assume that θ is chosen so that there are constants $K_1, K_2 \geq 0$ such that

$$(4.1) \qquad |\theta(x) - \theta(y)| \leq K_1\|x - y\| + K_2\|g'(x) - g'(y)\| \quad \text{for all } x, y \in X.$$

The functions θ_1 and θ_2 given in (2.5) and (2.6), respectively, satisfy (4.1). The fact that (2.6) satisfies (4.1) is an easy consequence of Proposition 4.1(4).

Now, given $x \in X$, recall the structure of the linear subproblems discussed in §2:

$$LP(x) : \text{ minimize}\,\{f(x) + f'(x)s : s \in L\Omega(x, \rho_1, \rho_2, \theta(x))\}.$$

As has been observed, the subproblem $LP(x)$ is always well defined and finite valued since $L\Omega(x, \rho_1, \rho_2, \theta(x))$ is a nonempty convex compact subset of $\mathbb{R}^n$ for all $x \in X$. In conjunction with $LP(x)$, we also need to consider the value function for $LP(x)$,

$$\ell(x) := \min\{f(x) + f'(x)s | s \in L\Omega(x, \rho_1, \rho_2, \theta(x))\},$$

the multifunction of Kuhn–Tucker solutions to $LP(x)$, $KT(x)$, and the multifunction of Kuhn–Tucker multipliers for the functional constraint $g(x) + g'(x)s \in C + \nu(x, \rho_1, \theta(x))\mathbb{B}$, $KTM(x)$. The properties of these objects that are important for our study are given in the following proposition.

PROPOSITION 4.2. (1) *Both $KT(x)$ and $KTM(x)$ are nonempty as long as $x \in X$ and $M_0(x) = \{0\}$. Moreover, both KT and KTM are upper semicontinuous on X.*

(2) *If $\alpha > \text{dist}_0(0|KTM(x))$ for all $x \in S \subset X$, then there exist nonnegative constants K_3, K_4, and K_5 such that*

$$(4.2)$$
$$|\ell(x) - \ell(y)| \leq K_3\|x - y\| + K_4\|f'(x) - f'(y)\| + K_5\|g'(x) - g'(y)\| \quad \textit{for all } y \in S.$$

Proof. (1) The first statement follows from Burke [1, Thm. 4.4] and the second follows from Proposition 4.1(9) and Burke [1, Prop. 6.1].

(2) Consider the exact penalty function

$$\widehat{P}_\alpha(s, x) := f(x) + f'(x)s + \alpha \text{ dist}(g(x) + g'(x)s | C + \nu(x, \rho_1, \theta(x))\mathbb{B}) + \psi(s | [X - x] \cap \rho_2\mathbb{B})$$

for $LP(x)$. From the hypothesis on α we obtain from Burke [3, Thms. 10.3 and 10.7] that the solution sets of the two convex programs $LP(x)$ and

$$LP_\alpha(x) : \text{ minimize}\,\{\widehat{P}_\alpha(s; x) | s \in \mathbb{R}^n\}$$

coincide on S with

$$\ell(x) = \min\{\widehat{P}_\alpha(s; x) | s \in \mathbb{R}^n\}.$$

Let s be a solution to $LP(y)$, $\widehat{s} \in [X - x] \cap \rho_2\mathbb{B}$ satisfy

$$\|s - \widehat{s}\| = \operatorname{dist}[s|[X - x] \cap \rho_2\mathbb{B}],$$

and $z \in \mathbb{B}$ satisfy

$$\operatorname{dist}(g(y) + g'(y)s + \nu(y, \rho_1, \theta(y))z|C) = \operatorname{dist}(g(y) + g'(y)s|C + \nu(y, \rho_1, \theta(y))\mathbb{B}).$$

Then, by Lemma 3.2 and parts (1) and (6) of Proposition 4.1, we have

$$\begin{aligned}
\ell(x) - \ell(y) &= \ell(x) - \widehat{P}_\alpha(s; y) \le \widehat{P}_\alpha(\widehat{s}; x) - \widehat{P}_\alpha(s; y) \\
&\le \rho_2\|f'(x) - f'(y)\| + M\|s - \widehat{s}\| + M\|x - y\| \\
&\qquad + \alpha[\operatorname{dist}[g(x) + g'(x)\widehat{s} + \nu(x, \rho_1, \theta(x))z|C] \\
&\qquad\quad - \operatorname{dist}[g(y) + g'(y)s + \nu(y, \rho_1, \theta(y))z|C]] \\
&\le \rho_2\|f'(x) - f'(y)\| + 3M\|x - y\| \\
&\qquad + \alpha\|g(x) + g'(x)\widehat{s} + \nu(x, \rho_1, \theta(x))z \\
&\qquad\quad - (g(y) + g'(y)s + \nu(y, \rho_1, \theta(y))z)\| \\
&\le \rho_2\|f'(x) - f'(y)\| + 3M\|x - y\| \\
&\qquad + \alpha[M\|x - y\| + \rho_2\|g'(x) - g'(y)\| \\
&\qquad\quad + M\|s - \widehat{s}\| + |\nu(x, \rho_1, \theta(x)) - \nu(y, \rho_1, \theta(y))|] \\
&\le \rho_2\|f'(x) - f'(y)\| + 3M\|x - y\| \\
&\qquad + \alpha[3M\|x - y\| + \rho_2\|g'(x) - g'(x)\| + 5M\|x - y\| \\
&\qquad\quad + \rho_1 M|\theta(y) - \theta(y)| + \rho_1\|g'(x) - g'(y)\|] \\
&\le \rho_2\|f'(x) - f'(y)\| + (3 + 8\alpha + \rho_1 K_1\alpha)M\|x - y\| \\
&\qquad + \alpha(\rho_2 + \rho_1 + \rho_1 K_2 M)\|g'(x) - g'(y)\|.
\end{aligned}$$

The result now follows by symmetry. □

Remark. For each $x \in X$ the function given by $\widehat{\ell}(x, t) := \min\{f(x) + f'(x)s | s \in L\Omega(x, \rho_1, t\rho_2, t\theta(x))\}$ is convex in t on $\mathbb{R}_+$. This follows from Proposition 4.1(8). Although this property has interesting consequences, we do not directly make use of it in our study.

The subproblems $LP(x)$ can also be used to characterize stationarity in $\mathcal{P}$ and to obtain descent directions for P_α for an appropriate choice of α.

PROPOSITION 4.3. *Let $x \in X$.*

(1) *Suppose that $KT(x)$ is nonempty and choose*

$$\alpha \ge \operatorname{dist}_0(0|KTM(x)) + \varepsilon \tag{4.3}$$

for some $\varepsilon \ge 0$. Then

$$\Delta_\alpha(x) := \ell(x) - f(x) + \alpha\theta(x)[\varphi(x, \rho_1) - \varphi(x, 0)] \le \varepsilon\theta(x)[\varphi(x, \rho_1) - \varphi(x, 0)]. \tag{4.4}$$

Moreover, if $\Delta_\alpha(x) = 0$, then x is a stationary point for $\mathcal{P}$. If both $\Delta_\alpha(x) = 0$ and $\varphi(x, 0) = 0$, then x is a Kuhn–Tucker point for $\mathcal{P}$.

(2) *If x is a Kuhn–Tucker point for $\mathcal{P}$, then*

$$\Delta_\alpha(x) = 0 \quad \textit{for all } \alpha \ge \operatorname{dist}_0(0|KTM(x)).$$

(3) *If* $(s, y, w, z) \in KT(x)$, *then*

$$P'_\alpha(s; x) \leq \Delta_\alpha(x). \tag{4.5}$$

Remarks. (1) If $x \in X$ is a stationary point for $\mathcal{P}$ that is not a Kuhn–Tucker point, then it is still possible that $KT(x)$ is nonempty and $\Delta_\alpha(x) < 0$ where α satisfies (4.3). This is illustrated by considering the example

$$\min\{x : x^3 \leq 0, -25 \leq x\}$$

at the point $x = 0$. This is an attractive feature of the subproblem $LP(x)$, since even if one is at such a stationary point for $\mathcal{P}$ it may still be possible to obtain descent directions for $\mathcal{P}$.

(2) Observe in (4.5) that if α is chosen with

$$\alpha \geq \|y\|_0 + \varepsilon, \tag{4.6}$$

then (4.3) is satisfied and so

$$P'_\alpha(x; s) \leq \Delta_\alpha(x) \leq 0 \tag{4.7}$$

with $P'_\alpha(x; s) = 0$ only if x is stationary for $\mathcal{P}$.

Proof. We begin by establishing statements (1) and (3) of the proposition. Let $(s, y, w, z) \in KT(x)$ and in the case of (1) we also assume that (s, y, w, z) is chosen so that

$$\|y\|_0 = \text{dist}_0(0|KTM(x)).$$

By Lemma 3.2, we have

$$\begin{aligned} -\langle y, g'(x)s\rangle &= \langle y, g(x)\rangle - \langle y, g(x) + g'(x)s\rangle \\ &= \langle y, g(x)\rangle - \psi^*(y|C + \nu(x, \rho_1, \theta(x))\mathbb{B}) \\ &\leq \|y\|_0 \ \text{dist}(g(x)|C + \nu(x, \rho_1, \theta(x))\mathbb{B}) \\ &= \|y\|_0\theta(x)[\varphi(x, 0) - \varphi(x, \rho_1)] \\ &\leq (\alpha - \varepsilon)\theta(x)[\varphi(x, 0) - \varphi(x, \rho_1)], \end{aligned} \tag{4.8}$$

$$\begin{aligned} -\langle w, s\rangle &= \langle w, x\rangle - \langle w, x + s\rangle \\ &= \langle w, x\rangle - \psi^*(w|X) \\ &\leq \|w\|_0 \ \text{dist}(x|X) \\ &\leq 0, \end{aligned} \tag{4.9}$$

and

$$-\langle z, s\rangle = -\psi^*(z|\rho_2\mathbb{B}) = -\rho_2\|z\|_0. \tag{4.10}$$

Since

$$f'(x)s = -[\langle y, g'(x)s\rangle + \langle w, s\rangle + \langle z, s\rangle], \tag{4.11}$$

these relations yield the inequality

$$\begin{aligned} f'(x)s &\le (\alpha-\varepsilon)\theta(x)[\varphi(x,0)-\varphi(x,\rho_1)]-\rho_2\|z\| \\ &\le (\alpha-\varepsilon)\theta(x)[\varphi(x,0)-\varphi(x,\rho_1)], \end{aligned} \tag{4.12}$$

from which inequality (4.4) immediately follows. Now if $\Delta_\alpha(x)=0$, then (4.4) implies $\varphi(x,\rho_1)=\varphi(x,0)$. Thus, by Theorem 3.1, we may as well assume that $\varphi(x,0)=0$. In this case $f'(x)s=0$ and so (4.12) implies that $z=0$. But then, by (4.11),

$$0=\langle y,g'(x)s\rangle+\langle w,x\rangle,$$

while (4.8) and (4.9) imply

$$0\le\langle y,g'(x)s\rangle$$

and

$$0\le\langle w,s\rangle,$$

respectively. Hence $0=\langle y,g'(x)s\rangle=\langle w,s\rangle$. Consequently, again by (4.8) and (4.9),

$$\langle y,g(x)\rangle=\psi^*(y|C)$$

and

$$\langle w,x\rangle=\psi^*(w|X),$$

and so, by Lemma 3.2, $y\in N(g(x)|C)$ and $w\in N(x|X)$ with

$$0=f'(x)^T+g'(x)^Ty+w.$$

Therefore, x is a Kuhn–Tucker point for $\mathcal{P}$.

To obtain (4.5) we simply observe that

$$\begin{aligned} P'_\alpha(x;s) &\le f'(x)s+\alpha[\operatorname{dist}(g(x)+g'(x)s|C)-\operatorname{dist}(g(x)|C)] \\ &\le f'(x)s+\alpha[\nu(x,\rho_1,\theta(x))-\varphi(x,0)] \\ &=\Delta_\alpha(x). \end{aligned}$$

(2) If $(s,y,w,z)\in KT(x)$, then, as in the proof of part (1), $(x,y,w)\in M_1(x)$. Hence $\alpha\ge\operatorname{dist}_0(0|KTM(x))\ge\operatorname{dist}_0(0|M_c(x))$ where

$$M_c(x):=\{y:(y,w)\in M_1(x)\text{ for some }w\in\mathbb{R}^n\}.$$

Hence, by Burke [3, Thm. 10.7], $P'_\alpha(x;s)\ge0$ for all $s\in\mathbb{R}^n$. Thus $\Delta_\alpha(x)=0$ by (4.4) and (4.5). □

From Proposition 4.2 we know that $KTM(x_k)\ne\emptyset$ as long as $M_0(x_k)=\{0\}$. Proposition 4.3 shows that if $KT(x_k)\ne\emptyset$ and $\Delta_{\alpha_k}(x_k)=0$, then x_k is a stationary point for $\mathcal{P}$. This proposition also assures us of the existence of an element $s\in L\Omega(x_k,\rho_1,t_k\rho_2,t_k\theta(x_k))$ for which $\Delta P_{\alpha_k}(s;x_k,H_k)<0$ whenever $\Delta_{\alpha_k}(x_k)<0$. Therefore, in Step 2 of the algorithm in §2 one can always locate an $s_k\in[X-x_k]\cap t_k\rho_2\mathbb{B}$ (note that s_k need not be in $L\Omega(x_k,\rho_1,t_k\rho_2,t_k\theta(x_k))$) for which $\Delta P_{\alpha_k}(s_k;x_k,H_k)<0$ as long as x_k is not a stationary point for $\mathcal{P}$. If $M_0(x_k)\ne\{0\}$, it may still be possible to obtain s_k such that $\Delta P_{\alpha_k}(s_k;x_k,H_k)<0$ as noted in remark (1) after Proposition 4.3. This is an attractive feature of the algorithm and it explains why we set $\alpha_{k+1} := \alpha_k$ if $KTM(x_k) = \emptyset$. Particular choices of the trial step s_k are studied in §§6 and 7.

5. Convergence.

The convergence theory presented in this section is modeled on that given in Powell [16, §4]. Consequently, we require the following assumption (see Powell [16, Thm. 2]).

ASSUMPTION 5.1. For every $\delta > 0$ there exist constants $\kappa_1, \kappa_2 > 0$ such that the inequality

$$\Delta P_{\alpha_k}(s_k, x_k, H_k) \le -\kappa_1 \min\{\kappa_2, t_k\} \tag{5.1}$$

holds whenever $\Delta_{\alpha_k}(x_k) \le -\delta$.

Inequality (5.1) is used to guarantee that the reduction in $P_{\alpha_k}(s; x_k, H_k)$ induced by s_k is comparable to the reduction that one would expect to obtain by use of the linear model $LP(x_k)$ alone. The following proposition indicates a way to choose s_k which assures the validity of inequality (5.1) when the sequence $\{H_k\}$ is bounded.

PROPOSITION 5.1. *Let $x \in X, H \in \mathbb{R}^{n\times n}$, $\alpha > 0$, and $\bar{t} \in (0,1)$. If $s \in [X - x] \cap \rho_2\mathbb{B}$ solves $LP(x)$, then there exists $\hat{t} \in [0, \bar{t}]$ such that*

$$\Delta P_\alpha(\hat{t}s; x, H) \le \frac{1}{2}\Delta_\alpha(x) \min\left\{\frac{|\Delta_\alpha(x)|}{(\sigma\rho_2)^2\|H\|_2}, \bar{t}\right\}, \tag{5.2}$$

where $\sigma > 0$ is chosen so that $\|z\|_2 \le \sigma\|z\|$ for $z \in \mathbb{R}^n$.

Proof. For $\lambda \in [0, \bar{t}]$ observe that

$$\begin{aligned}
\Delta P_\alpha(\lambda s; x, H) &\le \lambda f'(x)s + \frac{\lambda^2}{2}(\sigma\rho_2)^2\|H\|_2 \\
&\quad + \alpha\lambda[\operatorname{dist}[g(x) + g'(x)s|C] - \operatorname{dist}(g(x)|C)] \\
&\le \lambda f'(x)s + \frac{\lambda^2}{2}(\sigma\rho_2)^2\|H\|_2 \\
&\quad + \alpha\lambda[\nu(x, \rho_1, \theta(x)) - \varphi(x, 0)] \\
&= \lambda\Delta_\alpha(x) + \frac{\lambda^2}{2}(\sigma\rho_2)^2\|H\|_2.
\end{aligned}$$

If we now let

$$\hat{t} \in \arg\min\left\{\lambda\Delta_\alpha(x) + \frac{\lambda^2}{2}(\sigma\rho_2)^2 : \lambda \in [0, \bar{t}]\right\},$$

then it is straightforward to show that (5.1) is satisfied (see, for example, the proof of [16, Lemma 5, p. 20]). □

The following technical lemma greatly facilitates the discussion of convergence.

LEMMA 5.2. *Let $x \in X$, $H \in \mathbb{R}^{n\times n}$, $0 < \beta_1 < \beta_2 < 1$, and $\alpha, \kappa_1, \kappa_2 > 0$, and choose $\bar{t} > 0$ so that*

$$\kappa_1(1 - \beta_2)\min\{\kappa_2, t\} \ge (1 + \alpha)t\rho_2\omega_x(t\rho_2) + \tfrac{1}{2}(t\rho_2)^2\|H\|_2$$

for all $t \in [0, \bar{t}]$ where

$$\omega_x(t\rho_2) := \max\{\|f'(y) - f'(x)\|, \|g'(y) - g'(x)\| : y \in x + t\rho_2\mathbb{B}\}. \tag{5.3}$$

Then for every $t \in [0, \bar{t}]$ and $s \in [X - x] \cap t\rho_2\mathbb{B}$ for which

$$\Delta P_\alpha(s; x, H) \le -\kappa_1 \min\{\kappa_2, t\}, \tag{5.4}$$

one has

$$[P_\alpha(x+s) - P_\alpha(x)] \leq \beta_1 \Delta P_\alpha(s; x, H). \tag{5.5}$$

Proof. By Lemma 3.2(1), we have

$$\begin{aligned} P_\alpha(x+s) - P_\alpha(x) &\leq f(x) + f'(x)s + \alpha \operatorname{dist}(g(x) + g'(x)s|C) - P_\alpha(x) \\ &\qquad + \|f(x+s) - [f(x) + f'(x)s]\| \\ &\qquad + \alpha\|g(x+s) - [g(x) + g'(x)s]\| \\ &\leq \Delta P_\alpha(s; x, H) + \tfrac{1}{2}(t\rho_2)^2\|H\|_2 \\ &\qquad + (1+\alpha)t\rho_2\omega_x(t\rho_2) \\ &\leq \Delta P_\alpha(s; x, H) + \kappa_1(1-\beta_1)\min\{\kappa_2, t\} \\ &\leq \beta_1 \Delta P_\alpha(s; x, H) \end{aligned}$$

for every $t \in [0, \bar{t}]$ and $s \in [X - x] \cap t\rho_2\mathbb{B}$ satisfying (5.4). □

By combining Proposition 5.1 and Lemma 5.2 we see that unless $\Delta_{\alpha_k}(x_k) = 0$, one can always choose s_k in Step 2 of the algorithm of §2 so that $\Delta P_{\alpha_k}(s_k; x_k, H_k) < 0$ and Assumption 5.1 is satisfied. Furthermore, the procedure cannot jam at x_k with $x_{k+i} = x_k$ for all $i = 1, 2, \cdots$.

The main result is now given. The proof of this result is based on the approach of Powell in [16, Thm. 2].

THEOREM 5.3. *Let $\{x_k\}$ be a sequence generated by the algorithm of* §2 *for which Assumption* 5.1 *is satisfied.*

Furthermore, assume that f' and g' are bounded and uniformly continuous on $S := [\overline{co}\{x_i\} + \rho_2\mathbb{B}] \cap X$ and that the sequence $\{H_i\}$ is also bounded. Then at least one of the following must occur:

(1) $\Delta_{\alpha_k}(x_k) = 0$ *for some k and the procedure terminates,*

(2) $\alpha_k \uparrow +\infty$,

(3) $P_{\alpha_k}(x_k) \downarrow -\infty$,

(4) $\Delta_{\alpha_k}(x_k) \to 0$.

Proof. We will assume that none of (1)–(4) occur and derive a contradiction. First note that by Proposition 5.1 the sequence $\{x_k\}$ is infinite. Also observe that since α_k is bounded the updating strategy of Step 1 assures us that α_k remains constant for all k sufficiently large. Thus we may assume that $\alpha_k = \alpha$ for all $k = 1, 2, \cdots$. Now since $\Delta_\alpha(x_k) \nrightarrow 0$ there is a constant $\delta > 0$ and a subsequence $J \subset \mathbb{N}$ such that $\sup\{\Delta_\alpha(x_k) : k \in J\} < -2\delta < 0$. Consequently, by Assumption 5.1, there are constants $\kappa_1, \kappa_2 > 0$ such that (5.1) holds for all $k \in J$. Via Lemma 5.2, the uniform continuity of f' and g' now yield the existence of a $\bar{t} > 0$ such that

$$r_k \geq \beta_1 \quad \text{and} \quad x_{k+1} = x_k + s_k$$

whenever $t_k \leq \bar{t}$. Suppose there is a $\xi > 0$ and a subsequence $\widehat{J}$ of J such that

$$\inf[t_k | k \in \widehat{J}] > \xi.$$

Then for each $k \in \widehat{J}$ let $\sigma(k)$ be the first integer greater than or equal to k for which $x_{\sigma(k)+1} = x_{\sigma(k)} + s_{\sigma(k)}$ and consider the subsequence $\widehat{J}_\sigma := \{\sigma(k) | k \in \widehat{J}\}$. Observe that for each $k \in \widehat{J}_\sigma$ we have $t_k \geq \min\{\gamma_1\bar{t}, \gamma_1\xi\}$. Consequently, $P_\alpha(x_{k+1}) \leq$

$P_\alpha(x_k) - \kappa_1\beta_1 \min\{\kappa_2, \gamma_1\bar{t}, \gamma_1\xi\}$ for each $k \in \widehat{J}_\alpha$. But then $P_\alpha(x_k) \downarrow -\infty$, which is a contradiction. Therefore, we can assume that $t_k \le \bar{t}$ for all $k \in J$ and $\lim_J t_k = 0$.

By Proposition 4.1(4), Proposition 4.2(2), and (4.1), the uniform continuity of f' and g' imply the uniform continuity of Δ_α on S. Hence there is an $\bar{\varepsilon} > 0$ such that

$$|\Delta_\alpha(x_i) - \Delta_\alpha(x_j)| \le \delta$$

whenever $\|x_i - x_j\| \le \varepsilon$, $i, j \in \mathbb{N}$. Given $k \in J$ let $v(k)$ be the first integer greater than k for which one of

$$\|x_{v(k)} - x_k\| \le \bar{\varepsilon} \tag{5.6}$$

and

$$t_{v(k)} \le \bar{t} \tag{5.7}$$

is violated. If (5.6) is violated, then

$$P_\alpha(x_{s+1}) \le P_\alpha(x_s) - \kappa_1\beta_1 \min\{\kappa_2, t_s\}$$

and

$$t_{s+1} \ge t_s$$

for $s = k, \cdots, v(k) - 1$. Hence

$$P_\alpha(x_{v(k)}) \le P_\alpha(x_k) - \kappa_1\beta_1 \min\{\kappa_2, \bar{\varepsilon}/\rho_2\}$$

since

$$\sum_{k}^{\nu(k)-1} t_k\rho_2 \ge \|x_{v(k)} - x_k\| \ge \bar{\varepsilon}.$$

If (5.7) is violated, then

$$P_\alpha(x_{v(k)}) \le P_\alpha(x_{v(k)-1}) - \kappa_1\beta_1 \min\{\kappa_2, \gamma_3^{-1}\bar{t}\}.$$

In either case we have

$$P_\alpha(x_{v(k)}) \le P_\alpha(x_k) - \kappa_1\beta_1 \min\{\kappa_2, \bar{\varepsilon}/\rho_2, \gamma_2^{-1}\bar{t}\},$$

which implies that $P_\alpha(x_k) \downarrow -\infty$. This is the contradiction that establishes the result. □

COROLLARY 5.4. *Let $\{x_k\}, \{H_k\}$, f', and g' be as in Theorem* 5.3.

(1) *If $\alpha_k \uparrow +\infty$, then every cluster point $\bar{x}$ of the subsequence $J := \{i : \alpha_{i+1} > \alpha_i\}$ satisfies $M_0(\bar{x}) \ne \{0\}$ and so is either a Fritz John point or a nonfeasible stationary point for $\mathcal{P}$.*

(2) *If $\alpha := \sup\{\alpha_k\} < \infty$, then every cluster point $\bar{x}$ of $\{x_i\}$ is a stationary point for $\mathcal{P}$. Moreover, if $\varphi(x, 0) = 0$, then $\bar{x}$ is a Kuhn–Tucker point for $\mathcal{P}$.*

Proof. (1) Suppose to the contrary that $M_0(\bar{x}) = \{0\}$. Since $\alpha_k \uparrow \infty$, the multifunction

$$LM_1(x) := \{(y, w, z) | (s, y, w, z) \in KT(x) \text{ for some } s \in \mathbb{R}^n\}$$

is locally unbounded at $\overline{x}$. By Burke [1, Thm. 6.3] it must be the case that $\theta(\overline{x}) = 1$ and so $\varphi(\overline{x}, \rho_1) = \varphi(\overline{x}, 0)$ by (2.4). But then $\varphi(\overline{x}, 0) = 0$ since $M_0(\overline{x}) = \{0\}$. Furthermore, by Burke [3, Prop. 3.7],

$$LM_0(\overline{x}) := \mathrm{rec}[LM_1(\overline{x})] \neq \{0\}$$

since $LM_1(x)$ is locally unbounded at $\overline{x}$. But then by Burke [1, Thm. 4.3], $M_0(\overline{x}) \neq \{0\}$, a contradiction.

(2) Since the α_k's are bounded they eventually equal α. Moreover, for all k sufficiently large, $P_{\alpha_k}(x_k) \geq P_\alpha(\overline{x})$. Therefore, by Theorem 5.3, $\Delta_\alpha(\overline{x}) = 0$. Consequently, by (4.4), $\varphi(\overline{x}, \rho_1) = \varphi(\overline{x}, 0)$. Thus we can assume that $\varphi(\overline{x}, 0) = 0$. We now show that $\overline{x}$ is a Kuhn–Tucker point for $LP(\overline{x})$.

Let $\{(s_k, y_k, w_k, z_k)\}$ be such that $(s_k, y_k, w_k, z_k) \in KT(x_k)$ and $\alpha_k \geq \|y_k\|_0$ for all $k = 1, 2, \cdots$. Let $J \in \mathbb{N}$ be a subsequence for which $x_k \xrightarrow{J} \overline{x}$. If $\{(w_k, z_k)\}_J$ is bounded, then $\overline{x}$ is a Kuhn–Tucker point for $LP(\overline{x})$ since $KT(x)$ is upper semicontinuous. If $\{(w_k, z_k)\}_J$ is unbounded, we can assume that J is such that $(w_k, z_k)/(\|w_k\|_0 + \|z_k\|_0) \xrightarrow{J} (\widehat{w}_k, \widehat{z}_k) \neq (0, 0)$ and $s_k \xrightarrow{J} \widehat{s}$. Then $\widehat{w} \in N(\overline{x} + \widehat{s} | X)$, $\widehat{z} \in N(\widehat{s} | \rho_2 \mathbb{B})$, and $0 = \widehat{w} + \widehat{z}$ since $(s_k, y_k, w_k, z_k) \in KT(x_k)$ for all $k \in J$. Hence, by Lemma 3.2,

$$\begin{aligned} 0 &= -[\langle \widehat{w}, \widehat{s} \rangle + \langle \widehat{z}, \widehat{s} \rangle] \\ &= \langle \widehat{w}, \overline{x} \rangle - \langle \widehat{w}, \overline{x} + \widehat{s} \rangle - \rho_2 \|\widehat{z}\|_0 \\ &= \langle \widehat{w}, \overline{x} \rangle - \psi^*(\widehat{w} | X) - \rho_2 \|\widehat{z}\|_0 \\ &\leq \|\widehat{w}\|_0 \ \mathrm{dist}(\overline{x} | X) - \rho_2 \|\widehat{z}\|_0 \\ &= -\rho_2 \|\widehat{z}\|_0 \leq 0, \end{aligned}$$

but then $\widehat{z} = \widehat{w} = 0$, which is a contradiction.

Since $\overline{x}$ is a Kuhn–Tucker point for $LP(\overline{x})$ at which $\Delta_\alpha(\overline{x}) = 0$ for all $\alpha \geq \mathrm{dist}_0(0 | KTM(\overline{x})) + \varepsilon$, the result follows from Proposition 4.3. □

6. Application to $S\ell_1 QP$.

In this section we assume that $\mathcal{P}$ is given in standard form, the norms chosen for $\mathbb{R}^n$ and $\mathbb{R}^m$ are polyhedral, and the function $\theta : X \to [\theta_0, 1]$ of §4 is such that $\theta(x) = 1$ whenever $\varphi(x, \rho_1) = 0$. We now consider an instance of the algorithm of §2 wherein the choice of trial step s_k is based on the $S_{l_1}QP$ algorithm of Fletcher. The procedure incorporates the second-order correction technique due to Fletcher [11], [12] in order to avoid the Marotos effect.

Initialization. Choose $x_0 \in X$, $H_0 \in \mathbb{R}^{n \times n}$, $\alpha_{-1} > 0$, $\varepsilon > 0$, and $t_0 \in (0, 1)$. Set $k := 0$ and choose $\sigma > 0$ so that $\|x\|_2 \leq \sigma \|x\|$ for all $x \in \mathbb{R}^n$.

Step 1. Choose $(\widetilde{s}_k, \widetilde{y}_k, \widetilde{w}_k, \widetilde{z}_k) \in KT(x_k)$. If $\widetilde{s}_k = 0$, then stop; otherwise set

$$\alpha_k := \begin{cases} \alpha_{k-1}, & \text{if } \alpha_{k-1} \geq \|\widetilde{y}_k\|_0 + \varepsilon, \\ \max\{\|\widetilde{y}_k\|_0 + \varepsilon, \alpha_{k-1} + 4\varepsilon\}, & \text{otherwise,} \end{cases}$$

and

$$\widetilde{t}_k := \arg\min \left\{ \lambda \Delta_{\alpha_k}(x_k) + \frac{\lambda^2}{2} \widetilde{s}_k^T H_k \widetilde{s}_k : 0 \leq \lambda \leq k_k \right\}.$$

Step 2. Let s_k be a stationary point of the subproblem

$$\begin{aligned} QP_1(x_k, t_k) : &\min P_{\alpha_k}(s; x_k, H_k) \\ &\text{subject to } s \in t_k \rho_2 \mathbb{B} \cap S_k \end{aligned}$$

for which

$$P_{\alpha_k}(s_k; x_k, H_k) \le P_{\alpha_k}(\tilde{t}_k \tilde{s}_k; x_k, H_k), \tag{6.1}$$

where S_k is any subspace of $\mathbb{R}^n$ containing $\tilde{s}_k$.

Step 3. Set $r_k := [P_{\alpha_k}(x_k + s_k) - P_{\alpha_k}(x_k)][\Delta P_{\alpha_k}(s_k; x_k, H_k)]^{-1}$. If $r_k > 0.75$, go to Step 9.

Step 4. Let $\widehat{s}_k$ be a stationary point for the problem

$$\widehat{QP}(x_k, t_k) : \min \widehat{P}_k(s) \quad \text{subject to } s \in [X - x_k] \cap t_k \rho_2 \mathbb{B} \cap \widehat{S}_k,$$

where

$$\widehat{P}_k(s) := f(x_k) + f'(x_k)s + \tfrac{1}{2}s^T H_k s + \alpha_k \operatorname{dist}(g(x_k + s_k) - g'(x_k)s_k + g'(x_k)s | C)$$

and $\widehat{S}_k$ is any subspace of $\mathbb{R}^n$ containing S_k.
Set

$$r_k^e := r_k + \frac{\widehat{P}_k(0) - \widehat{P}_k(\widehat{s}_k)}{\Delta P_{\alpha_k}(s_k; x_k, H_k)}.$$

If $r_k < .25$, go to Step 6.

Step 5. If $r_k^e \in [0.9, 1.1]$, set $t_{k+1} := 2t_k$ and go to Step 11; otherwise go to Step 10.

Step 6. If $r_k^e \notin [0.75, 1.25]$, go to Step 7. Set

$$\widehat{r}_k := \frac{P_{\alpha_k}(x_k + \widehat{s}_k) - P_{\alpha_k}(x_k)}{\Delta P_{\alpha_k}(s_k; x_k, H_k)}.$$

If $\widehat{r}_k > 0.75$, set $s_k := \widehat{s}_k, r_k := \widehat{r}_k$, and go to Step 9.
If $\widehat{r}_k \ge 0.25$, set $s_k := \widehat{s}_k, r_k := \widehat{r}_k$, and go to Step 10.
If $\widehat{r}_k \ge r_k$, set $s_k := \widehat{s}_k$ and $r_k := \widehat{r}_k$.

Step 7. Choose $t_{k+1} \in [0.1t_k, 0.5t_k]$. If $r_k > 0.05$, go to Step 11.

Step 8. Set $x_{k+1} := x_k, k := k + 1$, and go to Step 2.

Step 9. If $\|s_k\| < t_k \rho_2$, go to Step 10. If $r_k > 0.9$, then $t_{k+1} := 4t_k$; otherwise $t_{k+1} := 2t_k$. Go to Step 11.

Step 10. Set $t_{k+1} := t_k$.

Step 11. Set $x_{k+1} := x_k + s_k$.

Step 12. Choose $H_{k+1} \in \mathbb{R}^{n \times n}$, set $t_{k+1} := \min\{t_{k+1}, 1\}$, $k := k + 1$, and go to Step 1.

Remarks. (1) The vector $\tilde{s}_k$ in Step 1 is often called the Cauchy step since it naturally corresponds to the best step obtainable from first-order information. The vector $\tilde{s}_k$ is used in (6.1) in order to assure the validity of inequality (5.2). In this way, Assumption 5.1 is satisfied.

(2) Except for the possibility of increasing t_k when $0.25 \le r_k \le 0.75$, this algorithm is an instance of the algorithm of §2. However, it is easily verified that this slight change in the implementation does not nullify the validity of Theorem 5.3 and Corollary 5.4.

The remarks above demonstrate that the results of §5 provide a global convergence theory for the algorithm in this section. Let us now concentrate on the local convergence. These results are obtained by appealing to the work of Yuan [30]. To this end we assume that $X = \mathbb{R}^n$ and we set

$$H_k := \nabla^2_{xx} L(x_k, y_k) = \nabla^2 f(x_k) + \sum_{i=1}^{m} y^{(i)}_{k+1} \nabla^2 g_i(x_k) \tag{6.2}$$

in Step 12, where $L(x, y) := f(x)+y^T g(x)$ is the Lagrangian for $\mathcal{P}$ and y_k is a multiplier estimate. For example, the multiplier estimate may be chosen as the solution to a least squares problem based on the optimality conditions. If $\{x_k\}$ is the sequence generated by the algorithm of this section, then we also assume the existence of a Kuhn–Tucker point $\overline{x}$ of $\mathcal{P}$ to which the sequence $\{x_k\}$ converges and at which the following hypotheses are satisfied:

(H1) (linear independence of the active constraint gradients). The gradients $\{g'_i(\overline{x}) : i \in A(\overline{x}) \cup \{s + 1, \cdots, m\}\}$ are linearly independent where

$$A(x) := \{i \in \{1, \cdots, s\} | g_i(x) \geq 0\}.$$

(H2) (strict complementary slackness). The unique Kuhn–Tucker multiplier vector $\overline{y} \in \mathbb{R}^n$ is such that $\overline{y}^{(i)} > 0$ for each $i \in A(\overline{x})$.

(H3) (second-order sufficiency condition). For each

$$s \in \{d \in \mathbb{R}^n : f'(\overline{x})d = 0 \text{ and } g'(\overline{x})d \in T(g(\overline{x})|C)\}$$

with $s \neq 0$, one has

$$s^T \left[\nabla^2 f(\overline{x}) + \sum_{i=1}^{m} \overline{y}^{(i)} \nabla^2 g_i(\overline{x}) \right] s > 0.$$

THEOREM 6.1. *Let $\{x_k\}$ be a sequence generated by the algorithm of* §6 *with $S_k = \widehat{S}_k = \mathbb{R}^n$ for all $k = 0, 1, 2$. Assume that $x_k \to \overline{x}$ and that hypotheses* (H1)–(H3) *hold at $\overline{x}$. Furthermore, assume that $H_k := \nabla^2_{\overline{x}} L(x_k, y_k)$ and that s_k and $\widehat{s}_k$ solve $QP_1(x_k, t_k)$ and $\widehat{QP}(x_k, t_k)$, respectively, for each $k = 0, 1, \cdots$ with $\{y_k\}$ chosen so that y_k converges to $\overline{y}$, the unique Kuhn–Tucker multiplier for $\mathcal{P}$ at $\overline{x}$. Then $x_k \to \overline{x}$ superlinearly, and if y_k is chosen to be the value of y that minimizes*

$$\|\nabla f(x_k) + g'(x_k)^T y\|_2,$$

then x_k converges to $\overline{x}$ quadratically.

Proof. The hypothesis (H1) implies that $M_0(\overline{x}) = \{0\}$, consequently, by Theorem 5.3, α_k is constant for all k sufficiently large. Therefore the algorithm is eventually an instance of the algorithm studied by Yuan in [30] and so the result follows from [30, Thm. 2.5, Cor. 2.6]. □

Remark. The assumption about the choice of multipliers $\{y_k\}$ is satisfied if, for example, one chooses the y_k's to be solutions to the least squares problems

$$\min\{\tfrac{1}{2}\|\nabla_x L(x_k, y)\|^2 : y \in \mathbb{R}^m\}.$$

7. Application to SQP.

In this section we again assume that $\mathcal{P}$ is given in standard form with $X = \mathbb{R}^n$, that the norms chosen for $\mathbb{R}^n$ and $\mathbb{R}^m$ are polyhedral, and that the function $\theta : X \to [\theta_0, 1]$ of §4 is such that $\theta(x) = 1$ whenever $\varphi(x, \rho_1) = 0$.

We consider an instance of the algorithm of §2 wherein the choice of trial step s_k is based on a modification to the Wilson–Han–Powell SQP subproblem. The algorithm is identical to the algorithm of §6 except that the subproblem $QP_1(x_k, t_k)$ in Step 2 is replaced by Step 2′.

Step 2′. Let s_k be a stationary point of the subproblem

$$\begin{aligned} QP_2(x_k, t_k) : &\min f(x_k) + f'(x_k)s + \tfrac{1}{2}s^T H_k s \\ &\text{subject to } s_k \in L\Omega(x_k, \rho_1, t_k\rho_2, t_k\theta(x_k)) \cap S_k \end{aligned}$$

for which

$$f(x_k) + f'(x_k)s_k + \frac{1}{2}s_k^T H_k s_k \le f(x_k) + f'(x_k)(\tilde{t}_k\tilde{s}_k) + \frac{\tilde{t}_k^2}{2}\tilde{s}_k^T H_k \tilde{s}_k, \tag{7.1}$$

where S_k is any subspace of $\mathbb{R}^n$ containing $\tilde{s}_k$.

Remarks. (1) Observe that from Proposition 4.1(7), we have

$$L\Omega(x_k, t_k\rho_1, t_k\rho_2, \theta(x_k)) \subset L\Omega(x_k, \rho_1, t_k\rho_2, t_k\theta(x_k)).$$

Hence the subproblems $QP_2(x_k, t_k)$ are always well defined. The subspaces S_k (and $\widehat{S}_k$) are introduced to reduce the dimensionality of the feasible region for the subproblems $QP_2(x_k, t_k)$ $(\widehat{QP}(x_k, t_k))$. For example, when $\varphi(x, \rho_1) = 0$, a typical choice for S_k would be the span of $\tilde{s}_k$ and the solution to the Wilson–Han–Powell SQP subproblem when this subproblem has a solution, e.g., see Celis, Dennis, and Tapia [7].

(2) Inequality (7.1) plays a role similar to that of inequality (6.1) in that it guarantees that Assumption 5.1 holds. Consequently, the results of §5 provide a global convergence theory for this modification to the algorithm of §6.

The local convergence theory for the algorithm when Step 2′ is used instead of Step 2 is not yet well understood. However, we conjecture that if the hypotheses (H1)–(H3) hold, then the subproblems QP_1 and QP_2 should produce identical trial steps s_k when x_k is sufficiently close to $\bar{x}$. If this is indeed true, then Theorem 6.1 remains valid when Step 2′ is used instead of Step 2. The resolution of this conjecture is the topic of ongoing research.

In lieu of establishing this conjecture, one can obtain a preliminary local convergence result by assuming that the trust region radius in the modified algorithm is eventually inactive. In this case, a local convergence result is easily obtained by appealing to results in Robinson [20].

THEOREM 7.1. *Let $\{x_k\}$ be a sequence generated by the algorithm in* §6 *with Step* 2 *replaced by Step* 2′ *and $S_k = \widehat{S}_k = \mathbb{R}^n$ for all $k = 0, 1, 2$. Assume that $x_k \to \bar{x}$ where $\bar{x}$ satisfies the assumptions* (H1)–(H3). *Furthermore, assume that $H_k := \nabla_x^2 L(x_k, y_{k-1})$ and that s_k and $\widehat{s}_k$ solve $QP_2(x_k, t_k)$ and $\widehat{QP}(x_k, t_k)$, respectively, for all $k \ge k_0$ for some $k_0 \in \mathbb{N}$, with each y_k chosen as a Kuhn–Tucker multiplier vector associated with the constraint*

$$g(x_k) + g'(x_k)s_k \in C + \nu(x_k, \rho_1, t_k\theta(x_k))\mathbb{B}$$

in $QP_2(x_k, t_k)$. If the trust region radius in QP_2 is eventually inactive, then $x_i \to \bar{x}$ quadratically.

Proof. Since the trust region constraint in the subproblems QP_2 is eventually inactive, the subproblems QP_2 reduce to the standard subproblems employed in the

Wilson–Han–Powell SQP method. Thus quadratic convergence follows from Robinson [20, Thm. 3.1]. □

Before closing, we wish to emphasize that the assumption that the trust region constraint is locally inactive is very strong. A more complete convergence result would establish conditions under which this hypothesis is valid. Until a clearer picture of the convergence properties of this procedure is established, the usefulness of Step $2'$ remains in doubt. Nonetheless, we introduce this alternative to Step 2 since we conjecture that the resulting algorithm possesses convergence properties similar to those described in Theorem 6.1. The resolution of this conjecture is the subject of ongoing research.

REFERENCES

[1] J. V. Burke, *A sequential quadratic programming method for potentially infeasible mathematical programs*, J. Math. Anal. Appl., 139 (1989), pp. 319–351.

[2] ———, *On the identification of active constraints* II: *The nonconvex case*, SIAM J. Numer. Anal., 27 (1990), pp. 1081–1102.

[3] ———, *An exact penalization view point of constrained optimization*, SIAM J. Control Optim., 29 (1991), pp. 968–998.

[4] J. V. Burke and S.-P Han, *A robust sequential quadratic programming method*, Math. Programming, 43 (1989), pp. 277–303.

[5] J. V. Burke, J. J. Moré, and G. Toraldo, *Convergence properties of trust region methods for linear and convex constraints*, Math. Programming, 47 (1990), pp. 305–336.

[6] R. H. Byrd, R. B. Schnabel, and G. A. Shultz, *A trust region algorithm for nonlinearly constrained optimization*, SIAM J. Numer. Anal., 24 (1987), pp. 1152–1170.

[7] M. R. Celis, J. E. Dennis, and R. A. Tapia, *A trust region strategy for nonlinear equality constrained optimization*, in Numerical Optimization 1984, P. T. Boggs, R. H. Byrd, and R. B. Schnabel, eds., Society for Industrial and Applied Mathematics, Philadelphia, PA, pp. 71–82.

[8] F. H. Clarke, *Optimization and Nonsmooth Analysis*, John Wiley and Sons, New York, 1983.

[9] A. R. Conn, N. I. M. Gould, and Ph. L. Toint, *Global convergence of a class of trust region algorithms for optimization problems with simple bounds*, SIAM J. Numer. Anal., 25 (1988), pp. 433–460.

[10] M. El-Alem, *A global convergence theory for the Celis–Dennis–Tapia algorithm for constrained optimization*, Tech. Report 88-19, Department of Mathematical Sciences, Rice University, Houston, TX, 1988.

[11] R. Fletcher, *Practical Methods of Optimization*, Second Edition, John Wiley and Sons, New York, 1987.

[12] ———, *Second order correction for nondifferentiable optimization*, in Numerical Analysis, G. A. Watson, ed., Springer-Verlag, Berlin, 1982, pp. 85–114.

[13] S.-P. Han, *A globally convergent method for nonlinear programming*, J. Optim. Theory Appl., 22 (1977), pp. 297–309.

[14] O. L. Mangasarian and S. Fromovitz, *The Fritz John necessary optimality conditions in the presence of equality and inequality constraints*, J. Math. Anal. Appl., 17 (1967), pp. 37–47.

[15] J. J. Moré, *Trust regions and projected gradients*, in Systems Modelling and Optimization: Proceedings of the 13th IFIP Conference on Systems Modelling and Optimization, M. Iri and K. Yajima, eds., Lecture Notes in Control and Information Sciences 113, Springer-Verlag, Berlin, 1988, pp. 1–13.

[16] M. J. D. Powell, *General algorithm for discrete nonlinear approximation calculations*, in Approximation Theory IV, C. K. Chui, L. L. Schumaker, and J. D. Ward, eds., Academic Press, New York, 1983, pp. 187–218.

[17] ———, *A fast algorithm for nonlinearly constrained optimization calculations*, in Proceedings of the 1977 Dundee Biennial Conference on Numerical Analysis, Springer-Verlag, Berlin, 1977.

[18] M. J. D. Powell and Y. Yuan, *A trust region algorithm for equality constrained optimization*, Math. Programming, 49 (1991), pp. 189–211.

[19] S. M. Robinson, *Stability theory for systems of inequalities. Part* II. *Differentiable nonlinear systems*, SIAM J. Numer. Anal., 13 (1976), pp. 497–513.

[20] S. M. Robinson, *Perturbed Kuhn–Tucker points, and rates of convergence for a class of nonlinear programming algorithms*, Math. Programming, 7 (1974), pp. 1–16.

[21] S. M. Robinson and R. R. Meyer, *Lower semi-continuity of multivalued linearization mappings*, SIAM J. Control, 11 (1973), pp. 525–533.

[22] R. T. Rockafellar, *Lipschitzian properties of multifunctions*, Nonlinear Anal. TMA, 9 (1985), pp. 867–885.

[23] ———, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.

[24] M. Sahba, *Globally convergent algorithm for nonlinearly constrained optimization problems*, J. Optim. Theory Appl., 52 (1987), pp. 291–309.

[25] M. Slater, *Lagrange multipliers revisited: A contribution to non-linear programming*, Cowles Commission Discussion Paper, Math. 403, 1950.

[26] Ph. L. Toint, *Global convergence of a class of trust region methods for nonconvex minimization in Hilbert space*, IMA J. Numer. Anal., 8 (1988), pp. 231–252.

[27] A. Vardi, *A trust region algorithm for equality constrained minimization: Convergence properties and implementation*, SIAM J. Numer. Anal., 22 (1985), pp. 575–591.

[28] R. B. Wilson, *A simplicial algorithm for concave programming*, Ph.D. thesis, Graduate School of Business Administration, Harvard University, Cambridge, MA, 1963.

[29] Y. Yuan, *Conditions for convergence of trust region algorithms for nonsmooth optimization*, Math. Programming, 31 (1985), pp. 220–228.

[30] ———, *On the superlinear convergence of a trust region algorithm for nonsmooth optimization*, Math. Programming, 31 (1985), pp. 269–285.

SIAM J. OPTIMIZATION
Vol. 2, No. 3, pp. 349-359, August 1992

AN $O(\sqrt{n}\, L)$-ITERATION LARGE-STEP PRIMAL–DUAL AFFINE ALGORITHM FOR LINEAR PROGRAMMING*

C. C. GONZAGA† AND M. J. TODD‡

Abstract. An algorithm based on reducing a suitable potential function for linear programming is described. At each iteration, separate scalings are applied to the primal and dual problems, and a step is taken in either the primal or the dual space. It is shown that a constant reduction can always be achieved, leading to a bound of $O(n^{1/2}L)$ iterations. Moreover, it is also shown that a reduction of $\Omega(n^{1/4})$ can usually be obtained, so that $O(n^{1/4}L)$ iterations are expected to suffice. Finally, it is proved that no general algorithm taking either primal or dual steps and guaranteeing the reduction of such a potential function can achieve R-order of convergence greater than one.

Key words. linear programming, interior-point methods, potential-reduction methods, path-following methods

AMS(MOS) subject classification. 90C05

1. Introduction. This paper presents a primal-dual interior-point algorithm for the linear programming problem

$$\text{(P)} \qquad \min c^T x, \quad Ax = b, \quad x \geqq 0.$$

The algorithm reduces a primal-dual potential function at each iteration by applying separate scalings to the primal and dual problems and taking a step in either the primal or the dual space. We show that a constant reduction can always be achieved, leading to a bound of $O(n^{1/2}L)$ iterations, where n is the number of variables and L the size of the input, assuming the data A, b, and c are all integers. Moreover, we give heuristic arguments showing that a greater reduction can usually be obtained, so that fewer iterations are expected to suffice. Finally, we show that no algorithm that drives this primal-dual potential function to minus infinity by taking either primal or dual steps at each iteration can achieve R-order of convergence greater than one, independent of n, although superlinear convergence is possible.

Our algorithm uses a primal-dual potential function described in Todd and Ye [15] based on the primal function of Karmarkar [6]. Other algorithms based on reducing this potential function include Ye [16]; Kojima, Mizuno, and Yoshise [8]; Huang and Kortanek [5]; and Gonzaga [3].

The conventional dual of (P), written with explicit slacks s, is

$$\text{(CD)} \qquad \max b^T y, \quad A^T y + s = c, \quad s \geqq 0.$$

It is more convenient to work with a dual problem also in standard form, involving only the nonnegative slacks s. Thus, let F be a matrix whose rows span the null space of A, and let $g = Fc$. Then $A^T y + s = c$ for some y if and only if $Fs = g$. Let d be any vector satisfying $Ad = b$. Then $b^T y = d^T A^T y = d^T c - d^T s$. Thus (CD) can be written in terms of s alone as

$$\text{(D)} \qquad \min d^T s, \quad Fs = g, \quad s \geqq 0.$$

* Received by the editors August 13, 1990; accepted for publication (in revised form) July 11, 1991. This research was supported in part by National Science Foundation grant ECS-8602534 and Office of Naval Research contract NOOO14-87-K0212.

† COPPE-Federal University of Rio de Janeiro, Cs. Postal 68511, Rio de Janeiro, RJ 21945, Brazil (gonzaga@brlncc.bitnet).

‡ Center for Applied Mathematics and School of Operations Research and Industrial Engineering, Cornell University, Ithaca, New York 14853 (miketodd@cs.cornell.edu).

In this form, weak duality takes the form that $c^Tx+d^Ts \geqq c^Td$ for all feasible x, s, and strong duality states that equality holds if and only if x and s are optimal. Notice that the duality gap c^Tx-b^Ty equals $c^Tx+d^Ts-c^Td$ and also x^Ts. This form of the dual was first presented by Todd and Ye [15].

Let $F_+(P)$ and $F_+(D)$ denote the set of strictly positive feasible solutions to (P) and (D), respectively, and let U_+ denote their Cartesian product $F_+(P)\times F_+(D)$. For $(x, s)\in U_+$, we define the primal-dual potential function ϕ_ρ by

$$\phi_\rho(x, s) := \rho \ln x^Ts - \sum_j \ln x_js_j \tag{1}$$

for $\rho > n$, the primal penalized function by

$$f_\alpha^P(x) := \alpha c^Tx - \sum_i \ln x_j \tag{2}$$

for $\alpha > 0$, and the dual penalized function by

$$f_\alpha^D(s) := \alpha d^Ts - \sum_j \ln s_j \tag{3}$$

for $\alpha > 0$. These functions combine a monotonic function of the objective function or duality gap with a barrier term in primal, dual, or primal-dual space.

We assume that $F_+(P)$ and $F_+(D)$ are nonempty and that $(x^0, s^0)\in U_+$ is known such that

$$\phi_\rho(x^0, s^0) = O(\sqrt{n}\,L), \tag{4}$$

where $\rho = n+\nu\sqrt{n}$ for some constant ν. Manipulating (P) and (D) to obtain such solutions is described in several papers, for example, Ye [16] and Kojima, Mizuno, and Yoshise [7], [8]. If we can reduce ϕ_ρ by a constant at each iteration, then $O(\sqrt{n}\,L)$ iterations will yield a pair $(x, s)\in U_+$ such that

$$(n+\nu\sqrt{n}) \ln x^Ts - \sum \ln x_js_j \leqq -\nu\sqrt{n}\,L$$

or

$$\ln x^Ts \leqq -L + \frac{1}{\nu\sqrt{n}} \sum \ln \frac{x_js_j}{x^Ts} \leqq -L. \tag{5}$$

Hence, $x^Ts \leqq 2^{-L}$, and exact solutions to (P) and (D) can then be obtained in $O(n^3)$ additional arithmetic operations. We use a large but constant ν, as proposed by Gonzaga [2], [3] because then, although the complexity bound is larger, it appears that larger steps can be taken and the algorithm will perform better in practice.

Our analysis is based on the relationship between ϕ_ρ and the penalized functions $f_\alpha^P(x)$ and $f_\alpha^D(x)$. Let P_A denote the projection onto the null space of A. Then, for any $s\in F_+(D)$, $P_As = P_Ac$ (the rows of P_A span the null space of A), so that

$$\begin{aligned} P_A\nabla_x\phi_\rho(x, s) &= \frac{\rho}{x^Ts} P_As - P_AX^{-1}e \\ &= \frac{\rho}{x^Ts} P_Ac - P_AX^{-1}e \\ &= P_A\nabla f_\alpha^P(x), \end{aligned} \tag{6}$$

where $\alpha = \rho/x^Ts$. Here and below, $X(S)$ denotes the diagonal matrix with the components of $x(s)$ down its diagonal, and e denotes the vector of ones in $\mathbb{R}^n$. Similarly,

$$P_F\nabla_s\phi_\rho(x, s) = P_F\nabla f_\alpha^D(s)$$

for the same α. These equations relate the potential function to the penalized functions.

Remark. The choice of F with rows spanning the null space $N(A)$ of A has no effect on the projection $P_F v$ of a given vector $v \in \mathbb{R}^n$. This is so because by definition the range space $R(F^T)$ equals $N(A)$ and hence $N(F) = R(A^T)$. Since $R(F^T) = N(A)$ and $N(F) = R(A^T)$ are orthogonal complements, $P_F v$ is the projection of v into the row space of A and equals $(I - P_A)v$.

The advantage of working with the penalized functions is that they have been well studied (see Fiacco and McCormick [1]), are strictly convex, and (under our assumptions) achieve their minima for each $\alpha > 0$. The minimizer of f_α^P over $F_+(P)$ is called the *α-center* of $F_+(P)$ and is denoted $x(\alpha)$; similarly, $s(\alpha)$, the α-center of $F_+(D)$, minimizes f_α^D over $F_+(D)$. Note that if x is the α-center of $F_+(P)$, then

$$0 = P_A \nabla f_\alpha^P(x) = \alpha P_A c - P_A X^{-1} e = \alpha(P_A c - P_A(\alpha^{-1} X^{-1} e)).$$

Hence, $s := \alpha^{-1} X^{-1} e$ satisfies $s > 0$, $P_A s = P_A c$, so $s \in F_+(D)$. Moreover,

$$P_F \nabla f_\alpha^D(s) = \alpha P_F d - P_F S^{-1} e = \alpha(P_F d - P_F x) = 0,$$

since $Ax = Ad = b$; thus, s is the α-center of $F_+(D)$. We therefore have the equations

$$X(\alpha)S(\alpha)e = \alpha^{-1} e, \qquad x(\alpha)^T s(\alpha) = n/\alpha, \tag{7}$$

relating the α-centers of $F_+(P)$ and $F_+(D)$. The path $\{(x(\alpha), s(\alpha))\}$ is called the *central path.*

Our algorithm makes no mention of the central path, but the convergence proof does. We must prove that the potential function decreases substantially at each iteration. At points where the projected gradients of the potential function are large, this is immediate; we must preclude the possibility of small projected gradients. For the primal or dual penalized functions, small projected gradients characterize nearly central points. We conclude that any reasonable potential restriction method will make good progress far from the central path, but we must pay special attention to its performance near the path. Corollary 2.1 in the next section shows that for our choice of potential function the primal and dual projected gradients can never both be small; its proof uses the contrast between the values $\alpha = n/x(\alpha)^T s(\alpha)$ and $\alpha = \rho/x^T s$ of the last two paragraphs, when $\rho = n + \nu\sqrt{n}$. Of course, other details such as scaling are necessary to guarantee a suitable step length.

In § 2, we show that, if a projected gradient like that in (6) is small, then $x \in F_+(P)$ is close to the α-center $x(\alpha)$ and $c^T x$ is close to $c^T x(\alpha)$. This extends a result in Gonzaga [2]. Hence, we prove that, given $(x, s) \in U_+$, we cannot have both x close to $x(\alpha)$ and s to $s(\alpha)$ for $\alpha = \rho/x^T s$. This result is the basis of our algorithm, which is described in § 3. We show that constant reduction in ϕ_ρ can be achieved at each iteration by taking a step in x or in s (or in both). One reason for taking steps in either the primal or the dual space is that McShane, Monma, and Shanno [9], in their empirical study of primal-dual methods, found that taking a step along an appropriate direction in primal-dual space led to poor computational results. They found it necessary to split the primal-dual direction into directions in primal and dual spaces and use separate step lengths (a fixed proportion of the distance to the boundary) in each. We give heuristic reasons to expect that a reduction of $\Omega(n^{1/4})$ in the potential function can often be achieved. These are based on plausible (but not rigorous) probabilistic assumptions. Rigorous bounds of $O(n^{1/4}L)$ steps have been obtained by Sonnevend, Stoer, and Zhao [13] for special classes of problems.

In § 4 we show that no algorithm that drives ϕ_ρ to $-\infty$ by taking at each iteration either a primal or a dual step can have R-order of convergence greater than one, independent of n. This result applies to our algorithm as well as to that of Ye [16].

Section 5 contains some concluding remarks and summarizes some preliminary computational results.

2. Preliminaries.

2.1. Scaling. Let Λ be a positive definite diagonal matrix. Then we can make the change of variables $\bar{x}=\Lambda^{-1}x$, so that (P) becomes

$$(\bar{\mathrm{P}})\qquad \min \bar{c}^T\bar{x},\quad \bar{A}\bar{x}=b,\quad \bar{x}\geqq 0$$

with

$$\bar{A}=A\Lambda,\qquad \bar{c}=\Lambda c.$$

The dual of $(\bar{\mathrm{P}})$ can be written as

$$(\bar{\mathrm{D}})\qquad \min \bar{d}^T\bar{s},\quad \bar{F}\bar{s}=g,\quad \bar{s}\geqq 0$$

in terms of $\bar{s}=\Lambda s$, where

$$\bar{F}=F\Lambda^{-1},\quad \bar{d}=\Lambda^{-1}d.$$

Note that the rows of $\bar{F}$ still span the null space of $\bar{A}$. If $\bar{f}_\alpha^P, \bar{f}_\alpha^D$, and $\bar{\phi}_\rho$ are defined from $(\bar{\mathrm{P}})$ and $(\bar{\mathrm{D}})$ as in (1)-(3), using $\bar{x}$, $\bar{s}$, $\bar{c}$, and $\bar{d}$, then

$$\begin{aligned}\bar{\phi}_\rho(\bar{x},\bar{s})&=\phi_\rho(x,s),\\ \bar{f}_\alpha^P(\bar{x})&=f_\alpha^P(x)+\ln\det\Lambda,\\ \bar{f}_\alpha^D(\bar{s})&=f_\alpha^D(s)-\ln\det\Lambda,\end{aligned}\tag{8}$$

when $\bar{x}$ and x, and $\bar{s}$ and s, correspond as above. It follows that

$$\begin{aligned}\nabla_{\bar{x}}\bar{\phi}_\rho(\bar{x},\bar{s})&=\Lambda\nabla_x\phi_\rho(x,s),\\ \nabla_{\bar{s}}\bar{\phi}_\rho(\bar{x},\bar{s})&=\Lambda^{-1}\nabla_s\phi_\rho(x,s),\\ \nabla_{\bar{x}}\bar{f}_\alpha^P(\bar{x})&=\Lambda\nabla_x f_\alpha^P(x),\\ \nabla_{\bar{s}}\bar{f}_\alpha^D(\bar{s})&=\Lambda^{-1}\nabla_s f_\alpha^D(s).\end{aligned}\tag{9}$$

Let $(\hat{x},\hat{s})\in U_+$. Then we can choose $\Lambda=\hat{X}=\operatorname{diag}(\hat{x})$ so that $\hat{x}$ is transformed into e in $\bar{x}$-space, or $\Lambda=\hat{S}^{-1}=(\operatorname{diag}(\hat{s}))^{-1}$ to transform $\hat{s}$ into e. In fact, we can use the first to scale x when considering changes in x and the second to scale s when considering changes in s. This separate scaling contrasts with the symmetric scaling using $\Lambda=(\hat{X}\hat{S}^{-1})^{1/2}$, which transforms both $\hat{x}$ and $\hat{s}$ to $(\hat{X}\hat{S})^{1/2}e$, which was used in the primal-dual algorithms of Kojima, Mizuno, and Yoshise [7], [8] and Monteiro and Adler [11], [12], for example. We only have to be careful when using separate scalings in dealing with ϕ_ρ, since (8) and (9) assume that the same scaling is used on both x and s.

The effect of scaling is that we can usually assume without loss of generality that our current x iterate (or our current s iterate) is e. In this case, any step of length less than one maintains positivity, so that, in the scaled space, steps in the direction of negative (projected) gradients are attractive.

2.2. A measure of centrality. Here we refine the analysis of Gonzaga [3] to show that we can effectively measure the distance from $x\in F_+(P)$ to the α-center of $F_+(P)$ by means of the norm of the (scaled) projected gradient of f_α^P. Recall that $f_\alpha^P(x)=\alpha c^Tx-\sum_j \ln x_j$ and that f_A^P is minimized at the α-center $x(\alpha)$.

DEFINITION 2.1. The measure of centrality of $x\in F_+(P)$ to the α-center $x(\alpha)$ is

$$\delta^P(x,\alpha):=\|P_{AX}X\nabla f_\alpha^P(x)\|_2,$$

where $X = \operatorname{diag}(x)$. Similarly,

$$\delta^D(s, \alpha) := \|P_{FS} S \nabla f_\alpha^D(s)\|_2$$

is the measure of centrality of $s \in F_+(D)$ to the α-center $s(\alpha)$.

We can also view $\delta^P(x, \alpha)$ as the length of the Newton step to minimize f_α^P from x; see [3]. Note that $\delta^P(x, \alpha) = \bar{\delta}^P(e, \alpha)$, where the scaling is chosen to transform x into e, i.e., $\Lambda = X$.

PROPOSITION 2.1. *Let* $x \in F_+(P)$ *and*

$$\psi := \frac{\|X^{-1}(x(\alpha) - x)\|_2}{\|X^{-1}(x(\alpha) - x)\|_\infty}. \tag{10}$$

If $\delta := \delta^P(x, \alpha) \leqq \psi/4$, *then*

$$\varepsilon := \|X^{-1}(x(\alpha) - x)\|_2 \leqq 2\delta \tag{11}$$

and

$$|c^T x(\alpha) - c^T x| \leqq 3\sqrt{n}\, \delta/\alpha. \tag{12}$$

Proof. Without loss of generality, we can assume that $x = e$; otherwise, scale so that this is true. Then let

$$h := \frac{x(\alpha) - e}{\|x(\alpha) - e\|_2}$$

so that $\|h\|_2 = 1$ and $x(\alpha) = e + \varepsilon h$. Since f_α^P is convex and minimized at $x(\alpha)$, $h^T \nabla f_\alpha^P(e + \lambda h) < 0$ for $0 \leqq \lambda < \varepsilon$. Now

$$\begin{aligned} \nabla f_\alpha^P(e + \lambda h) &= \alpha c - ((1 + \lambda h_j)^{-1}) \\ &= \alpha c - e + \lambda \left(\frac{h_j}{1 + \lambda h_j} \right) \\ &= \nabla f_\alpha^P(e) + \lambda h - \lambda^2 \left(\frac{h_j^2}{1 + \lambda h_j} \right). \end{aligned}$$

Also,

$$|h^T \nabla f_\alpha^P(e)| = |h^T P_A \nabla f_\alpha^P(e)| \leqq \|h\|_2 \|P_A \nabla f_\alpha^P(e)\| = \delta, \tag{13}$$

using $h = P_A h$ and Definition 2.1. Hence, for $0 \leqq \lambda < \varepsilon$,

$$\begin{aligned} 0 &> h^T \nabla f_\alpha^P(e + \lambda h) \\ &= h^T \nabla f_\alpha^P(e) + \lambda h^T h - \lambda^2 \sum_j \left(\frac{h_j^3}{1 + \lambda h_j} \right) \\ &\geqq -\delta + \lambda - \lambda^2 \sum_{j: hj > 0} \left(\frac{h_j^3}{1 + \lambda h_j} \right) \qquad (14) \\ &\geqq -\delta + \lambda - \lambda^2 \|h\|_\infty \sum_{j: hj > 0} h_j^2 \\ &\geqq -\delta + \lambda - \lambda^2/\psi, \end{aligned}$$

using $\|h\|_2 = 1$ and $\|h\|_\infty = 1/\psi$. Since $\delta \leqq \psi/4$ by hypothesis, (14) gives

$$\frac{(\lambda - \psi/2)^2}{\psi} = \frac{\lambda^2}{\psi} - \lambda + \frac{\psi}{4} \geqq \frac{\lambda^2}{\psi} - \lambda + \delta > 0$$

for $0 \leqq \lambda < \varepsilon$, so $\varepsilon \leqq \psi/2$. Thus $\lambda^2/\psi \leqq \lambda\varepsilon/\psi \leqq \lambda/2$ for $0 \leqq \lambda < \varepsilon$, so (14) yields

$$-\delta + \lambda - \lambda/2 < 0 \quad \text{for } 0 \leqq \lambda < \varepsilon,$$

from which $\varepsilon \leqq 2\delta$ follows.

To show (12), note that

$$\begin{aligned} |c^T(x(\alpha) - e)| &= |\alpha c^T(\varepsilon h)|/\alpha \\ &= |(\alpha c - e)^T(\varepsilon h) + e^T(\varepsilon h)|/\alpha \\ &\leqq (\varepsilon/\alpha)(|\nabla f_\alpha^P(e)^T h| + |e^T h|) \\ &\leqq (2\delta/\alpha)(\delta + |e^T h|) \\ &\leqq 3\sqrt{n}\, \delta/\alpha, \end{aligned} \tag{15}$$

using (13), (11), $e^T h \leqq \|e\|_2 \|h\|_2 = \sqrt{n}$, and $\delta \leqq \psi/4 \leqq \sqrt{n}/4$ by the definition of ψ. □

Of course, an analogous result holds for the dual problem (D). An easy consequence is that, for a suitable choice of α, we cannot simultaneously have x close to $x(\alpha)$ and s close to $s(\alpha)$.

COROLLARY 2.1. *Let $(x, s) \in U_+$ and let $\alpha = \rho/x^T s$, where $\rho = n + \nu\sqrt{n}$, $\nu > 0$. Let $\Delta > 0$ be such that $\Delta \leqq \frac{1}{4}$ and $\Delta < \nu/6$ (for example, $\Delta = \frac{1}{4}$ with $\nu \geqq 2$). Then we cannot have both $\delta^P(x, \alpha) \leqq \Delta$ and $\delta^D(s, \alpha) \leqq \Delta$.*

Proof. Suppose the contrary. Then Proposition 2.1 applies since $\psi \geqq 1$, and we deduce that

$$|c^T x(\alpha) - c^T x| \leqq 3\sqrt{n}\, \delta^P/\alpha$$

and

$$|d^T s(\alpha) - d^T s| \leqq 3\sqrt{n}\, \delta^D/\alpha.$$

Hence the duality gaps $x(\alpha)^T s(\alpha)$ and $x^T s$ satisfy

$$|x(\alpha)^T s(\alpha) - x^T s| \leqq 6\sqrt{n}\, \Delta/\alpha. \tag{16}$$

However,

$$x^T s = \frac{\rho}{\alpha} = \frac{n}{\alpha} + \nu\frac{\sqrt{n}}{\alpha},$$

and since $x(\alpha)^T s(\alpha) = n/\alpha$,

$$|x(\alpha)^T s(\alpha) - x^T s| = \nu\sqrt{n}/\alpha.$$

Comparing this expression with (16), we obtain $\nu \leqq 6\Delta$, contrary to the hypothesis. □

Now recall (6). The corollary implies that, when $\rho = n + \nu\sqrt{n}$ with $\nu \geqq 2$, we cannot have both $P_{AX} X \nabla_x \phi_\rho(x, s)$ and $P_{FS} S \nabla_s \phi_\rho(x, s)$ with norms at most $\frac{1}{4}$. Thus, it is reasonable that the potential function ϕ_ρ can be reduced substantially by taking either a primal or dual step. This is the basis of our method.

3. The algorithm. We can now state our algorithm. To obtain the polynomial bound, (x^0, s^0) must be chosen with $\phi_\rho(x^0, s^0) = O(\sqrt{n}\, L)$ and η should be 2^{-2L}. For the convergence analysis we choose $\nu \geqq 2$, but any positive value $\nu = O(1)$ can be used instead with suitable changes to the constants.

ALGORITHM

Given $(x^0, s^0) \in U_+$ and termination parameter $\eta > 0$:

set $k \leftarrow 0$, $\rho \leftarrow n + \nu\sqrt{n}$ for $2 \leqq \nu = O(1)$.

```
**Repeat** until $(x^k)^T s^k \leqq \eta$;
    **set** $(x, s) \leftarrow (x^k, s^k)$;
    **choose** either a primal or a dual step
        (a primal step can only be chosen if $\|P_{AX} X \nabla_x \phi_\rho(x, s)\|_2 \geqq \frac{1}{4}$,
        a dual step only if $\|P_{FS} S \nabla_s \phi_\rho(x, s)\|_2 \geqq \frac{1}{4}$);
        **primal step:**
        $x^+ = x - \beta X P_{AX} X \nabla_x \phi_\rho(x, s)$;
        $s^+ = s$; or
        **dual step:**
        $x^+ = x$;
        $s^+ = s - \beta S P_{FS} S \nabla_s \phi_\rho(x, s)$;
    where $\beta > 0$ is such that
        $\phi_\rho(x^+, s^+) \leqq \phi_\rho(x, s) - \frac{1}{40}$;
    **update:**
        $k \leftarrow k + 1$;
        $(x^k, s^k) \leftarrow (x^+, s^+)$;
**end.**
```

Corollary 2.1 implies that the conditions for taking a primal or dual step can always be met. If β can be chosen to assure the decrease of $\frac{1}{40}$ in ϕ_ρ, then the argument in the introduction implies that the algorithm will terminate in at most $O(\phi_\rho(x^0, s^0) + \nu\sqrt{n} \ln(\eta^{-1}))$ iterations.

PROPOSITION 3.1. *If β in the algorithm is chosen as*

$$(6\|P_{AX} X \nabla_x \phi_\rho(x, s)\|_2)^{-1} \quad \textit{or} \quad (6\|P_{FS} S \nabla_s \phi_\rho(x, s)\|_2)^{-1}$$

according to whether a primal or dual step is taken, then

$$\phi_\rho(x^+, s^+) \leqq \phi_\rho(x, s) - \tfrac{1}{40}.$$

Proof. This follows from standard arguments. We know that

$$\phi_\rho(x + \Delta x, x) \leqq \phi_\rho(x, s) + \Delta x^T \nabla_x \phi_\rho(x, s) - \frac{\|X^{-1}\Delta x\|_2^2}{2(1 - \|X^{-1}\Delta x\|_\infty)};$$

see, for instance, Ye [16]. Then, if $\Delta x = -\beta X P_{AX} X \nabla_x \phi_\rho(x, s)$, we find $\|X^{-1}\Delta x\|_2 = \frac{1}{6}$, and so

$$\phi_\rho(x + \Delta x, s) \leqq \phi_\rho(x, s) - \frac{\delta}{6} + \frac{(\frac{1}{6})^2}{2(1 - \frac{1}{6})}, \tag{17}$$

where $\delta = \|P_{AX} X \Delta_x \phi_\rho(x, s)\|_2$. Since δ is at least $\frac{1}{4}$ if a primal step is taken, we find ϕ_ρ is reduced by at least $\frac{1}{40}$. A similar argument applies if a dual step is taken.

In fact, we could take steps in both primal and dual spaces, exactly as above. Since the step Δx in x lies in the null space of A, it is orthogonal to the step Δs in s. Hence, $(x^+)^T s^+ = x^T s + \Delta x^T s + x^T \Delta s$ is linear in $(\Delta x, \Delta s)$, and the standard argument shows that

$$\begin{aligned} \phi_\rho(x + \Delta x, s + \Delta s) \leqq{} & \phi_\rho(x, s) + \Delta x^T \nabla_x \phi_\rho(x, s) + \Delta s^T \nabla_s \phi_\rho(x, s) \\ & - \frac{\|X^{-1}\Delta x\|_2^2}{2(1 - \|S^{-1}\Delta s\|_\infty)} - \frac{\|S^{-1}\Delta s\|_2^2}{2(1 - \|S^{-1}\Delta s\|_\infty)}; \end{aligned}$$

if we choose step sizes so that $\|X^{-1}\Delta x\|_2 = \|S^{-1}\Delta s\|_2 = \frac{1}{6}$, then we deduce that

$$\phi_\rho(x + \Delta x, s + \Delta s) \leqq \phi_\rho(x, s) - \frac{\delta^P}{6} - \frac{\delta^D}{6} + \frac{2(\frac{1}{6})^2}{2(1 - \frac{1}{6})}.$$

Since δ^P and δ^D are both nonnegative and one is at least $\frac{1}{4}$, we find that ϕ_ρ is reduced by at least $\frac{1}{120}$.

Remarks. (a) In most iterations we expect a much greater reduction in potential than guaranteed by the worst-case analysis. Indeed, suppose

$$h = h^P := \frac{X^{-1}(x(\alpha)-x)}{\|X^{-1}(x(\alpha)-x)\|_2}$$

satisfies $\|h\|_2/\|h\|_\infty \geqq 4n^{1/4}$ and $|e^Th| \leqq n^{1/4}$ (in what follows we give heuristic reasons for these hypotheses). Then the proof of Proposition 2.1 [see especially (15)] shows that if $\delta^P(x, \alpha) \leqq n^{1/4}$, $|c^Tx(\alpha) - c^Tx| \leqq 4\sqrt{n}/\alpha$. If similar hypotheses hold for the dual direction h^D, we could conclude, as in Corollary 2.1, that for $\nu > 8$ we cannot have both $\delta^P(x, \alpha)$ and $\delta^D(s, \alpha)$ at most $n^{1/4}$. This would imply, from (17), that a reduction of potential of $\Omega(n^{1/4})$ could be obtained. In the integer model, only $O(n^{1/4}L)$ iterations of this kind are necessary. This heuristic analysis gives a similar bound to the rigorous results of Sonnevend, Stoer, and Zhao [13] for special classes of linear programming problems.

In [4], we show that the hypotheses we have made in the previous paragraph on h^P and h^D hold with probability approaching 1 as $n \to \infty$ if h^P and h^D are uniformly distributed on the unit sphere. Of course, this is not a rigorous analysis of expected behavior, but it gives some justification to the heuristic arguments above.

(b) From the description of the algorithm, it seems that two projections must be calculated at each iteration. However, since at most one of x and s is changed, one of these projections is the same as at the last iteration. Hence only one new factorization is required at each iteration after the first, and thus one can choose a primal or dual step corresponding to the larger projected gradient. In practice, P_{FS} can be computed as $I - P_{AS^{-1}}$ (see the remark following (6)).

4. *R*-order of convergence.

In § 3 we described an algorithm that takes either a primal or a dual step at each iteration and that drives the potential function ϕ_ρ to $-\infty$, for $\rho = n + \nu\sqrt{n}$, ν a constant. Ye's algorithm [16] is also of this form. In this section we provide a limit to the convergence rate of such an algorithm. This can be contrasted with the quadratic convergence of Zhang and Tapia's primal-dual algorithm [17].

Recall that, if $x^k \to x^*$, then x^k converges to x^* *R-linearly* if

$$\|x^k - x^*\|_2 \leqq \alpha\gamma^k \tag{18}$$

for some $\alpha > 0$ and $0 < \gamma < 1$, whereas the convergence has *R-order q* $(q > 1)$ if for all $1 \leqq p < q$,

$$\|x^k - x^*\|_2 \leqq \alpha\gamma^{p^k}. \tag{19}$$

THEOREM 4.1. *Consider an algorithm that takes either a primal or a dual step at each iteration and that drives ϕ_ρ to $-\infty$, for $\rho = n + \nu\sqrt{n}$ and ν a constant. Then this algorithm cannot guarantee that the generated sequences $\{x^k\}$ or $\{s^k\}$ have R-order of convergence greater than one (independent of n).*

Proof. We assume that (P) and (D) have unique nondegenerate optimal solutions $\bar{x}$ and $\bar{s}$, respectively, with

$$\bar{x}_j > 0 \quad \text{iff } j \in B, \quad |B| = m,$$

$$\bar{s}_j > 0 \quad \text{iff } j \in N, \quad |N| = d,$$

and $m + d = n$. We indicate the appropriate subvectors of x and s by a subscript B or N. Any convergent algorithm has $x \to \bar{x}$, $s \to \bar{s}$, and we will assume henceforth that the iterates have converged sufficiently that

$$x_B \leqq 2\bar{x}_B, \qquad s_N \leqq 2\bar{s}_N. \tag{20}$$

Given $(x, s) \in U_+$, the duality gap is $x^T s$, and this is the sum of the primal gap (from optimality), denoted

$$\pi := \bar{s}^T x = \bar{s}_N^T x_N,$$

and the dual gap, denoted

$$\theta := \bar{x}^T s = \bar{x}_B^T s_B.$$

Then, assuming that (20) holds, we have for any $\rho \geqq n$,

$$\begin{aligned}
\phi_\rho(x, s) &= \rho \ln(\pi + \theta) - \sum_{j \in B} \ln x_j s_j - \sum_{j \in N} \ln x_j s_j \\
&\geqq \rho \ln(\pi + \theta) - \sum_{j \in B} \ln \bar{x}_j s_j - \sum_{j \in N} \ln x_j \bar{s}_j - n \ln 2 \\
&\geqq \rho \ln(\pi + \theta) - m \ln \frac{\bar{x}_B^T s_B}{m} - d \ln \frac{\bar{s}_N^T x_N}{d} - n \ln 2 \qquad (21) \\
&= \rho \ln(\pi + \theta) - m \ln \theta - d \ln \pi + m \ln 2m + d \ln 2d - n \ln 4 \\
&\geqq \rho \ln(\pi + \theta) - m \ln \theta - d \ln \pi + m \ln \frac{2m}{n} + d \ln \frac{2d}{n} + n \ln n - n \ln 4 \\
&\geqq \max\{(\rho - m) \ln \theta - d \ln \pi, (\rho - d) \ln \pi - m \ln \theta\} + n \ln(n/4);
\end{aligned}$$

the last inequality follows since

$$\frac{2m}{n} \ln \frac{2m}{n} + \frac{2d}{n} \ln \frac{2d}{n}$$

can be written as $(1+\alpha)\ln(1+\alpha) + (1-\alpha)\ln(1-\alpha)$, which is convex and nonnegative, minimized at $\alpha = 0$.

Suppose that ϕ_ρ has been driven below $n \ln(n/4)$. Then from (21) we can deduce that

$$d \ln \pi \geqq (\rho - m) \ln \theta, \qquad m \ln \theta \geqq (\rho - d) \ln \pi,$$

or

$$\pi \geqq \theta^{(\rho - m)/d}, \quad \theta \geqq \pi^{(\rho - d)/m}. \tag{22}$$

Suppose that $m = d = n/2$, so that $(\rho - m)/d = (\rho - d)/m = 1 + 2\nu/\sqrt{n}$. Assume that an algorithm of the type considered generates a sequence $(x^k, s^k) \in U_+$ with primal and dual gaps (π_k, θ_k). Then

$$\min\{\pi_{k+1}, \theta_{k+1}\} \geqq (\min\{\pi_k, \theta_k\})^{1 + 2\nu/\sqrt{n}} \tag{23}$$

for all sufficiently large k. Indeed, suppose $\pi_{k+1} \leqq \theta_{k+1}$ (the other case is similar). If a dual step was just taken, then $\pi_{k+1} = \pi_k$, which is at least the right-hand side of (23) if $\pi_k < 1$. If a primal step was taken, (22) gives $\pi_{k+1} \geqq \theta_{k+1}^{1+2\nu/\sqrt{n}} = \theta_k^{1-2\nu/\sqrt{n}}$, which is again greater than or equal to the right-hand side of (23), as long as (20) holds for (x^{k+1}, s^{k+1}) and $\phi_\rho(x^{k+1}, s^{k+1}) \leqq n \ln(n/4)$.

Of course, (23) yields

$$\min\{\pi_{k+l}, \theta_{k+l}\} \geqq (\min\{\pi_k, \theta_k\})^{(1+2\nu/\sqrt{n})^l}$$

for any sufficiently large k, which implies by (19) that the R-order of convergence cannot be greater than $1 + 2\nu/\sqrt{n}$. Since ν is a constant, this proves the theorem. □

Remarks. (a) The theorem should not be interpreted as stating that algorithms of this type should not be used. Fast R-linear algorithms can be very attractive if α is small and γ close to zero in (18). Moreover, our result does not preclude R-order of convergence $1+2\nu/\sqrt{n}$, which implies R-superlinear convergence. However, asymptotic quadratic convergence cannot be achieved, although it is possible for algorithms that move in primal and dual spaces simultaneously [17].

(b) From (21) we can establish an interesting property of central pairs (x, s), i.e., $x = x(\alpha)$, $s = s(\alpha)$ for some α. Using (7), we find that $\phi_n(x, s) = n \ln n$, so that (noting that $\rho - m = d$, $\rho - d = m$ if $\rho = n$) we conclude

$$\pi \geqq (\tfrac{1}{4})^{n/d}\theta, \quad \theta \geqq (\tfrac{1}{4})^{n/m}\pi. \tag{24}$$

Hence, the primal and dual gaps are of similar order for central pairs sufficiently close to optimal. This result can be contrasted with (22) and provides further motivation for choosing a large value of ν, which gives more freedom in reducing π for a given value of θ or vice versa.

5. Concluding remarks. We have described an algorithm that at each iteration takes a scaled projected steepest-descent step for the primal-dual potential function in either the primal or the dual variables. We have shown that one of these steps will assure a constant decrease in the potential function, hence providing a bound on the number of iterations required. The algorithm is more symmetric than that of Ye [16], which either takes a step in primal space or updates the dual (these operations are not symmetric), but it is perhaps less so than the primal-dual potential reduction methods of Kojima, Mizuno, and Yoshise [8] and Huang and Kortanek [5].

We have tested a preliminary implementation of the algorithm on random problems of sizes 50×100 up to 500×1000, using PRO-MATLAB [10]. For $\nu = 2$ or $\nu = 10$ the number of iterations seems to grow more slowly than $O(n^{1/2})$; for $\nu = 10$ the growth is close to $O(n^{1/4})$, as in remark (a) of § 3. The minimum value of $\max\{\delta^P(x, \alpha), \delta^D(s, \alpha)\}$ over all iterations was 2-2.5, independent of the size of the problem for $\nu = 2$, and 7.5-11 for all sizes for $\nu = 10$; note that $7.5 > n^{1/4}$ even for $n = 1{,}000$, so this does not contradict the heuristic arguments in that remark.

Our best results were obtained for $\rho = 2n$; then the number of iterations ranged from 22-28 for the smallest problems and 31-35 for the largest. These figures are about twice those obtained by a different algorithm [14] for the same problems. It was observed that the iterations usually alternated between primal and dual, so that from a primal point of view, every other iteration only updates a lower bound. On the other hand, most other algorithms change both primal and dual solutions or update the lower bound and change the primal solution in a single iteration (requiring just one factorization).

REFERENCES

[1] A. V. Fiacco and G. P. McCormick, *Nonlinear Programming: Sequential Unconstrained Minimization Techniques*, John Wiley, New York, 1968, reissued in *Classics in Applied Mathematics*, R. E. O'Malley, Jr., ed., Society for Industrial and Applied Mathematics, Philadelphia, PA, 1990.

[2] C. C. Gonzaga, *Large step path-following methods for linear programming. Part* I: *Barrier function method*, SIAM J. Optimization, 1 (1991), pp. 268-279.

[3] C. C. Gonzaga, *Large step path-following methods for linear programming. Part* II: *Potential reduction method*, SIAM J. Optimization, 1 (1991), pp. 280-292.

[4] C. C. Gonzaga and M. J. Todd, *An* $O(\sqrt{n}\,L)$*-iteration large-step primal-dual affine algorithm for linear programming*, Tech. Report No. 862, School of Operations Research and Industrial Engineering, Cornell University, Ithaca, NY, 1989.

[5] S. Huang and K. O. Kortanek, *A note on a potential reduction algorithm for LP with simultaneous primal-dual updating*, Oper. Res. Lett., 10 (1991), pp. 501-507.

[6] N. K. Karmakar, *A new polynomial time algorithm for linear programming*, Combinatorica, 4 (1984), pp. 373-395.

[7] M. Kojima, S. Mizuno, and A. Yoshise, *A polynomial algorithm for a class of linear complementarity problems*, Math. Programming, 44 (1989), pp. 1-26.

[8] ———, *An $O(\sqrt{n}\, L)$ iteration potential reduction algorithm for linear complementarity problems*, Math. Programming, 50 (1991), pp. 331-342.

[9] K. A. McShane, C. L. Monma, and D. Shanno, *An implementation of a primal-dual interior point method for linear programming*, ORSA J. Comput., 1 (1989), pp. 70-83.

[10] C. B. Moler, J. Little, S. Bangert, and S. Kleinman, *Pro-Matlab User's Guide*, MathWorks, Sherborn, MA, 1987.

[11] R. C. Monteiro and I. Adler, *Interior path following primal-dual algorithms, Part* I: *Linear programming*, Math. Programming, 44 (1989), pp. 27-41.

[12] ———, *Interior path following primal-dual algorithms, Part* II: *Convex quadratic programming*, Math. Programming, 44 (1989), pp. 43-66.

[13] G. Sonnevend, J. Stoer, and G. Zhao, *On the complexity of following the central path of linear programs by linear extrapolation*, Methods Oper. Res., 63 (1989), pp. 19-31.

[14] M. J. Todd, *A low complexity interior-point algorithm for linear programming*, SIAM J. Optimization, 2 (1992), pp. 198-209.

[15] M. J. Todd and Y. Ye, *A centered projective algorithm for linear programming*, Math. Oper. Res., 15 (1990), pp. 508-529.

[16] Y. Ye, *An $O(n^3L)$ potential reduction algorithm for linear programming*, Math. Programming, 50 (1991), pp. 239-258.

[17] Y. Zhang and R. Tapia, *A superlinearly convergent polynomial primal-dual interior-point algorithm for linear programming*, Res. Report 91-02, Department of Mathematics, University of Maryland, Catonsville, MD, 1991; SIAM J. Optimization, 3 (1993), to appear.

SIAM J. OPTIMIZATION
Vol 2, No. 3, pp. 360-375, August 1992

TRANSFER METHOD FOR CHARACTERIZING THE EXISTENCE OF MAXIMAL ELEMENTS OF BINARY RELATIONS ON COMPACT OR NONCOMPACT SETS*

JIANXIN ZHOU† AND GUOQIANG TIAN‡

Abstract. This paper systematically studies the existence of maximal elements for unordered binary relations on compact or noncompact sets in a general topological space. This is done by developing a method, called *transfer method*, to derive various necessary and sufficient conditions that characterize the existence of maximal elements for a binary relation in terms of: (1) (generalized) transitivity conditions under certain topological assumptions; (2) topological conditions under certain (generalized) transitivity assumptions; and (3) (generalized) convexity conditions under certain topological assumptions. There are two basic approaches in the literature to prove the existence by providing sufficient conditions. One assumes certain convexity and continuity conditions for a topological vector space and the other assumes certain weakened transitivity and continuity conditions for a general topological space. The results unify those two approaches and generalize almost all of the existing results in the literature.

Key words. binary relations, maximal elements, transfer continuities, transfer transitivities, transfer convexities

AMS(MOS) subject classifications. 49A27, 90C48, 90C31, 90B50

1. Introduction. Let Y be a topological space and $u: Y \rightarrow \mathbf{R}$ be a function. The classical Weierstrass theorem states that u attains its maximum on any nonempty compact set $X \subset Y$ if u is upper semicontinuous. As generalizations of the Weierstrass theorem, Tian and Zhou [18] proved two theorems that give necessary and sufficient conditions for u to attain its maximum on a nonempty compact set by introducing the notion of transfer continuities. The idea behind this is quite simple. To characterize the existence of maximal points for a function u, for given $u(x) > u(y)$, we really do not have to know the topological relations between x and a neighborhood $\mathcal{N}(y)$ of y. All we need to know is the topological relations between a neighborhood of y and a point x' in the upper part of $u(y)$, i.e., whether x can be transferred to x', a point in the upper part of $u(y)$ such that $u(x') > (\geqq) u(y')$: $y' \in \mathcal{N}(y)$, and if so, u is said to be *transfer* (*weakly*) *upper continuous* on X. However, in many cases in economics, decision analysis, optimization, and game theory, *a binary relation is not representable by a function even for an ordering.* Thus many results are given in the literature to prove the existence of maximal elements of a binary relation for this case. See, e.g., Yu [22], Borwein [4], Luc [10], and others who study the existence of maximal elements for a partial ordering induced by a convex cone in a topological vector space; and Fan [7],[1] Schmeidler [13], Sonnenschein [14], Shafer [11], Shafer and Sonnenschein [12], Bergstrom [3], Walker [19], Yannelis and Prabhakar [21], Campbell and Walker [5], Tian [16], [17], and others who study the existence of maximal elements for unordered binary relations by assuming either certain convexities or certain transitivities (at least acyclicity). Most of the results provide only sufficient conditions.

* Received by the editors August 13, 1990; accepted for publication (in revised form) July 26, 1991.

† Department of Mathematics, Texas A&M University, College Station, Texas 77843. This work was supported in part by Air Force Office of Scientific Research grant 87-0334.

‡ Department of Economics, Texas A&M University, College Station, Texas 77843.

[1] Fan [7, Lem. 4] does not phrase his results in terms of maximizing binary relations, but his results can be interpreted that way.

In this paper we systematically study maximization of binary relations. Note that there are two basic approaches to unordered binary relations in the literature: one through "weak" (i.e., reflexive) binary relations, and the other through "strict" (i.e., irreflexive) binary relations. Kim and Richter [9] made the connection between these two approaches and proved that *these two approaches are equally valid*: *definitions and theorems in one approach correspond in parallel to definitions and theorems in the other approach.* So in this paper we, without loss of generality, deal with only "strict" binary relations.

Let Y be a set and $X \subset Y$ be a subset. Denote by "$>$" an "irreflexive (strict) binary relation" on Y and "$\geqq$" the completion of "$>$," i.e., $x \geqq y$ means that $y > x$ does not hold, and thus "$\geqq$" is a "reflexive (weak) and complete binary relation." Here $y > x$ is read "y is *strictly preferred* (*or dominated*) to x" and y is said to be a dominator to x. Let A be a subset of Y and $y \in Y$. Denote by $y > (\geqq) A$ if $y > (\geqq) x$ for all $x \in A$ and y is said to be a dominator (maximizer) to A.

An element $x^* \in X$ is said to be a *maximal element* of the binary relation "$>$" on X if $x \geqq X$, i.e., x^* has no dominator in X.

Our objective in this paper is to study the existence of maximal elements for a binary relation "$>$" on a nonempty compact or noncompact set. We characterize the existence in terms of: (1) certain topological conditions, (2) certain (generalized) transitivity conditions, and (3) certain generalized convex (geometric) conditions. We extend the notion of transfer continuities further to *transfer transitivities* and *transfer convexities.* We call this notion the *transfer method.* The basic idea behind it is as follows. For topology, given $x > y$, the conventional continuity conditions describe topological behavior or relations between x and a neighborhood of y. For transitivity, given a finite subset $X_0 = \{x_1, x_2, \cdots, x_n\}$, conventional transitivities describe "relations" within the finite set X_0, i.e., the "internal relations." For geometry and algebra, given a finite subset $X_0 = \{x_1, x_2, \cdots, x_n\}$, conventional convexity conditions describe "relations" between this finite set and its convex hull. To characterize the existence of maximal elements for "$>$," when $x > y$, we do not have to know the topological relations between x and a neighborhood of y, the internal relations of the finite subset X_0, and the geometric and algebraic relations between the finite set and its convex hull. We only need to know, for topology, the topological behavior or relations between a neighborhood of y and an element x' in its "upper" part (so x can be transferred to a certain element x' in the "upper" part of a neighborhood of y); for transitivity, the relations between the finite subset X_0 and an element x' in the "upper" part of the finite subset X_0, i.e., the "external relation"; for geometry, the relations between the finite set X_0 and the convex hull of a corresponding finite subset in the part not "below" X_0. Conditions describing the topological relations between a neighborhood of y and an element in its "upper" part are called *transfer continuities*; conditions describing the relations between the finite subset X_0 and an element in its "upper" part are called *transfer transitivities*; and conditions describing the geometric relations between the finite subset X_0 and the convex hull of a corresponding finite set in the part not "below" X_0 are called *transfer convexities.*

This paper consists of four sections. In § 1, we introduce various transfer conditions and we explore their connections with conventional conditions and some of their properties as preliminaries for further development. In § 2, we characterize the existence of maximum elements for binary relations on nonempty compact sets by giving necessary and/or sufficient conditions in terms of: (1) transfer transitivity conditions under certain transfer continuity assumptions, (2) transfer continuity conditions under certain transfer transitivity assumptions, and (3) transfer convexity conditions under certain transfer continuity assumptions. In § 3, we first discuss some properties of the definitions in § 1, and then provide several theorems to characterize the existence of maximum elements for binary relations on nonempty noncompact sets by also giving necessary and/or sufficient conditions in terms of various transfer conditions. In § 4, as concluding remarks, we first indicate that our results can be used to give conditions under which the maximum correspondence in Walker's Maximum Theorem is nonempty valued, which is required for many applications in decision analysis and game theory and serves as part of our motivation for this work. Then we show how a maximization problem, with respect to a (weak) binary relation, can be converted to a maximization problem, with respect to a (strict) binary relation, so our approach can be applied.

1.1. Transfer transitivities. In the following definition, whenever $K = X$, "to K" will be replaced by "on X" or omitted.

DEFINITION 1. Let K be a subset of a set X. A binary relation "$>$" defined on X is said to be:

(1) *Transfer n-maximal* to K, if for each finite subset $\{x_1, x_2, \cdots, x_n\} \subset X$ there exists $x' \in K$ such that $x' \geqq \{x_1, x_2, \cdots, x_n\}$;

(2) *transfer finitely maximal* to K, if it is transfer n-maximal to K for all $n = 1, 2, \cdots$;

(3) *n-acyclic* on X, if $x_1 > x_2 > \cdots > x_k$ implies $x_1 \geqq x_k$ for all $k = 1, 2, \cdots, n$ (1-acyclic just means $x \geqq x$ for all $x \in X$);

(4) *acyclic* on X, if it is n-acyclic for all $n = 1, 2, \cdots$;

(5) *transfer n-strict maximal* to K, if for all y_i, x_i in X with $y_i > x_i$, $i = 1, 2, \cdots, n$ there exists $x' \in K$ such that $x' > \{x_1, x_2, \cdots, x_n\}$;

(6) *transfer finitely strict maximal* to K, if it is transfer n-strict maximal to K for all $n = 1, 2, \cdots$;

(7) *n-link transitive* on X, if $y > x_0 \geqq x_1 \geqq \cdots \geqq x_n > z$ implies $y > z$;

(8) *link transitive* on X, if it is n-link transitive on X for all $n = 0, 1, 2, \cdots$;

(9) *fully transitive* on X, if its completion "$\geqq$" is transitive on X, i.e., $x \geqq y \geqq z$ imply $x \geqq z$.

Remark 1. Here we can see that many definitions in the literature have been unified. The way we define those transitivities makes it easier for us to save terminologies and to see implications among different transitivities. For instance, in Definitions 1(1), 1(3), 1(5), and 1(7), the case for $n+1$ implies the same case for n. Conventionally:

(1) a 1-acyclic "$>$" is said to be irreflexive, i.e., not $x > x$ or $x \geqq x$ for all $x \in X$;

(2) a 2-acyclic "$>$" is said to be asymmetric, i.e., $x > y$ and not $y > x$, which implies $x \geqq y$, and is also said to be a "preference" relation;

(3) a 0-link transitive "$>$" is said to be (weakly) transitive in [5]. Therefore, a 0-link transitive "$>$" induces a partial ordering;

(4) a 1-link transitive "$>$" is said to be extratransitive in [5].

Remark 2. Cone preference relations have been adopted very often in (multiple-criteria decision making) vector optimization (cf. Yu [22], Borwein [4], Tanaka [15], Ferro [8], and Luc [10]). Therein (weak) cone preferences are defined to induce partial orderings. Here we show that a cone preference is just a very special case of our approach.

Let X be a subset of a real topological vector space Y and let C be a convex cone in Y. Let $C^- = -C$. Define (see, e.g., [10]) the (weak) cone preference "$\geqq_c$" in Y by $y \geqq_c x$ if and only if $y - x \in C$. Thus its asymmetric part of $\geqq_c$, denoted by $>_c$, i.e., $y >_c x$ whenever $y \geqq_c x$ and not $x \geqq_c y$, defines a strict preference relation. Then a point $x^* \in X$ is said to be an *efficient* (*or minimal*) point of X with respect to C if no $x \in X$ such that $x^* >_c x$, i.e., either $x^* - x \notin C$ or $x^* - x \in C \cap C^-$ for all $x \in X$. For such a (weak) cone preference, we can define a (strict) cone preference relation "$>$" in Y by $y > x$ if and only if $x - y \in C \backslash (C \cap C^-)$, and write its completion "$\geqq$" by $x \geqq y$ whenever $y > x$ does not hold, i.e., $x \geqq y$ if either $x - y \notin C$ or $x - y \in C \cap C^-$. Now following our definition, a maximal element of "$>$" on X is an element $x^* \in X$ such that $x^* \geqq x$, for all $x \in X$, i.e., in this case $x^* - x \notin C$ or $x^* - x \in C \cap C^-$ for all $x \in X$. Thus x^* is an efficient point of "$\geqq_c$" if and only if it is a maximal point of "$>$."

It may be remarked that the above-defined (strict) cone binary relation "$>$" is 0-link transitive on X, i.e., $z > y > x$ implies $z > x$. To see this we only need to show that $x' \in C \backslash (C \cap C^-)$ and $y' \in C \backslash (C \cap C^-)$ imply $x' + y' \in C \backslash (C \cap C^-)$. Since C is a convex cone, $x' + y' \notin C \backslash (C \cap C^-)$ implies $x' + y' \in (C \cap C^-)$ and $y' \in C$ implies $-y' \in C^-$. Then $x' = (x' + y') + (-y') \in C^-$and leads to a contradiction.

Thus our approach is very general and includes the cone preference as a special case. It then frees us, in considering vector optimization, from using linear structures and from restricting a binary relation to being defined by a cone. We believe our *transfer method* can be applied to cone preference to both derive and characterize the existence of maximal elements.

Remark 3. Campbell and Walker [5] overlooked the fact that the pseudotransitivity in [5], defined by "$x_1 > x_2 \geqq x_3 > x_4$ implies $x_1 > x_4$ when $x_2 \neq x_3$," is weaker than the 1-link transitivity when $>$ is asymmetric. The pseudotransitivity and 1-link transitivity are equivalent by noting that the pseudotransitivity implies the 0-link transitivity (since the pseudotransitivity and the 0-link transitivity together clearly imply the 1-link transitivity). To see this, suppose that $x > y > z$ (which implies $x \neq z$ by the asymmetricity), but $z \geqq x$. Then we have $y > z \geqq x > y$. The pseudotransitivity implies $y > y$ but this is impossible.

Since our objective is to characterize the existence of *maximal* elements for a binary relation, to better understand those transitivities stated in Definition 1, it is beneficial for us to restate some of these transitivities in terms of maximization terminologies.

LEMMA 1. *Let "$>$" be a binary relation on a set X.*

(1) *For any fixed integer $n = 1, 2, 3, \cdots$, the binary relation is n-acyclic on X if and only if any n elements $\{x_1, x_2, \cdots, x_n\} \subset X$ have an internal maximal element, i.e., there exists $x_i \in \{x_1, x_2, \cdots, x_n\}$ such that $x_i \geqq \{x_1, x_2, \cdots, x_n\}$. Consequently, the binary relation is acyclic on X if and only if for any integer n, any finite subset $\{x_1, x_2, \cdots, x_n\} \subset X$ has an internal maximal element.*

(2) *The binary relation is acyclic if it is 0-link transitive.*

Proof. (1) The second part of (1) follows from the first part, so we only need to prove the first part. The cases $n = 1, 2$ are obvious. For $n > 2$, to prove the "only if" part, we assume that the binary relation is n-acyclic and that there exists n elements $\{x_1, x_2, \cdots, x_n\} \subset X$ without an internal maximal element. These elements therefore form a $k+1$ cycle for some integer k with $3 \leqq k \leqq n$. Without loss of generality, we assume that the $k+1$ cycle is of the form $x_1 > x_2 > \cdots > x_k > x_1$. Since the binary relation is k-acyclic, we have $x_1 \geqq x_k > x_1$, and this is impossible.

To prove the "if" part, we assume that for each fixed integer $n > 2$ the binary relation has an internal maximal element for any n elements in X. This implies that the binary relation has an internal maximal element for any k elements in X with $3 \leqq k \leqq n$. Let $\{x_1, x_2, \cdots, x_k\}$ be k elements in X with $x_1 > x_2 > \cdots > x_k$. Since $x_k > x_1$ will force a $k+1$ cycle to form, i.e., these k elements have no internal maximal elements, and reduce to a contradiction, we must have $x_1 \geqq x_k$ and thus the binary relation is k-acyclic for all $1 \leqq k \leqq n$.

(2) Without loss of generality, if there is a cycle of the form $x_1 > x_2 \cdots > x_k > x_1$ for some $1 \leqq k \leqq n$, 0-link transitivity will lead to $x_1 > x_1$—a contradiction. □

LEMMA 2. *The binary relation is* 1*-link transitive on X if and only if for any integer n and any x_i and y_i with $y_i > x_i$, $i = 1, 2, \cdots, n$, there exists $1 \leqq k \leqq n$ such that $y_k > \{x_1, x_2, \cdots, x_n\}$.*

Proof. The "if" part is obvious. For if $y > x_1 > x_2 > z$, then either $y > \{x_1, z\}$ or $x_2 > \{x_1, z\}$. But $x_1 \geqq x_2$, so $y > \{x_1, z\}$ and thus "$>$" is 1-link transitive. Now we prove the "only if" part by mathematical induction. When $n = 1$ it is obviously true. Suppose it is true for all $n \leqq m$. Now for $n = m+1$, if $y_i > x_i$, $1 \leqq i \leqq m+1$, then, according to the assumption on $n \leqq m$, there exists y_k, $1 \leqq k \leqq m$, such that $y_k > \{x_1, x_2, \cdots, x_m\}$. If $y_k > x_{m+1}$, it is done. Otherwise we have $y_{m+1} > x_{m+1} \geqq y_k > \{x_1, x_2, \cdots, x_m\}$. By the 1-link transitivity, we obtain $y_{m+1} > \{x_1, x_2, \cdots, x_m\}$. Then $y_{m+1} > \{x_1, x_2, \cdots, x_m, x_{m+1}\}$. □

Remark 4. Definitions 1(3), 1(4), 1(7), and 1(8) are of conventional types and Definitions 1(1), 1(2), 1(5), and 1(6) are of transfer types. By consulting Lemmas 1 and 2 we can see how we applied the transfer method to the conventional Definitions 1(3), 1(4), 1(7), and 1(8) (we simply allow the dominator or maximal element to n elements to exist inside or outside the n elements) to obtain, respectively, Definitions 1(1), 1(2), 1(5), and 1(6). Therefore, they are very natural generalizations of the conventional assumptions. It is these transfer conditions that enable us to avoid the asymmetric assumption. When $K = X$ we have the following implications among various transitivities stated in Definition 1, while none of their converses hold ($\oplus$ means that the binary relation is asymmetric):

$$\begin{tikzcd}
 & (9) \arrow[d, Rightarrow, "\oplus"'] & \\
 & (8) \arrow[d, Rightarrow] & \\
 & (7) \arrow[dl, Rightarrow] \arrow[dr, Rightarrow, "(n>0)"] & \\
(4) \arrow[d, Rightarrow] \arrow[dr, Rightarrow] & & (6) \arrow[dl, Rightarrow, "\oplus"] \arrow[d, Rightarrow] \\
(3) \arrow[dr, Rightarrow, "(\text{same } n)"'] & (2) \arrow[d, Rightarrow] & (5) \arrow[dl, Rightarrow, "\oplus(\text{same } n)"] \\
 & (1) &
\end{tikzcd}$$

where

(3)⇒(1) follows from Lemma 1(1);

(7)⇒(6) follows from Lemma 2;

(7)⇒(4) follows from Lemma 1(2);

(5)⇒(1) because if n elements have no maximal element, each one of them has a dominator; then (5) guarantees the existence of an outside dominator (a maximal element under asymmetry) to all these n elements;

(4)⇒(2) follows from Lemma 1(1).

Next we provide two examples to demonstrate that for a binary relation the acyclic condition strictly implies the transfer finitely maximal condition, while the acyclic condition is independent of the transfer finitely strict maximal condition.

Example 1. Let $Y=\mathbf{C}$, the complex plane. Define a binary relation "$>$" for any $z_1, z_2 \in Y$ by

$$(1)\qquad z_1 > z_2 \quad \text{iff} \begin{cases} \text{either} & |z_1|<|z_2| \text{ and } z_1, z_2 \text{ are on the same ray from the origin} \\ \text{or} & |z_1|=|z_2| \text{ but } z_1=e^{i\theta}z_2, \text{for } 0<\theta\leqq\pi/2. \end{cases}$$

Let X be the unit disk on the complex plane $\mathbf{C}$. Then for each r, $0<r\leqq 1$, we have a cycle

$$(r,0)>(0,-r)>(-r,0)>(0,r)>(r,0).$$

However, the origin is the unique maximal point on X, which is strictly preferred to any other point. So "$>$" is transfer finitely strict maximal (of course, transfer finitely maximal) on X.

Example 2. Let $Y=\mathbf{C}$, the complex plane. Define a binary relation "$>$" for any $z_1, z_2 \in Y$ by

$$(2)\qquad z_1 > z_2 \quad \text{iff} \begin{cases} \text{either} & |z_1|<|z_2| \text{ and } \arg(z_1)=\arg(z_2) \\ \text{or} & |z_1|=|z_2| \text{ but } z_1=e^{i\theta}z_2, \text{for } 0<\theta\leqq\pi/2. \end{cases}$$

Here the argument of the origin, arg (0), is regarded as zero. Let X be the unit disk on the complex plane $\mathbf{C}$. Then for each r, $0<r\leqq 1$, we have a cycle

$$(r,0)>(0,-r)>(-r,0)>(0,r)>(r,0).$$

However, the origin is the unique maximal point on X, thus "$>$" is transfer finitely maximal on X. Indeed, we have $(0,0)>(0,r)$ for all $0<r\leqq 1$, and $(0,0)\geqq$ any other points (where $>$ does not hold). If we let X be the upper half of the unit disk, including the bottom line, then it is easy to see that "$>$" is acyclic, but is not transfer finitely strict maximal on X. So we can see that the acyclic condition and the transfer finitely strict maximal condition are two independent conditions. We point out here that the 0-link transitive condition and the transfer finitely strict maximal condition are also independent.

1.2. Transfer continuities and convexities.

DEFINITION 2. Let X be a subset of a topological space Y and let z be any point in Y; denote $\mathcal{N}(z)$ a neighborhood of z. The binary relation "$>$" defined on Y is said to be:

(1) *upper continuous* on X, if for any $x\in X$ and $y\in Y$, $x>y$ implies that there exists $\mathcal{N}(y)$ such that $x>\mathcal{N}(y)$;

(2) *weakly upper continuous* on X, if for any $x\in X$ and $y\in Y$, $x>y$ implies that there exists $\mathcal{N}(y)$ such that $x\geqq\mathcal{N}(y)$.

For convenience, in further developments we define the weakly upper contour correspondence $U_w: X\to 2^Y$ by $U_w(x)=\{y\in Y: y\geqq x\}$ for each $x\in X$, and we define the strictly upper contour correspondence $U_s: Y\to 2^X$ by $U_s(x)=\{y\in X: y>x\}$ for each $x\in Y$.

Remark 5. Note that "$>$" is upper continuous on X if and only if U_s has (relatively) open lower sections on X, i.e., if and only if $U_s^{-1}(x)$ is open for all $x\in X$. The upper continuity is called the lower continuity in [3] and [20] and the weakly upper continuity in [5] is called the weakly lower continuity. The reason we call them "upper" is that when the binary relation "$>$" is represented by a real-valued function on Y our definitions coincide with the usual upper semicontinuities.

Let Z be a convex subset in a topological vector space E. A correspondence $P: Z\to 2^Z$ is said to be SS-*convex* (refer to Shafer and Sonnenschein) if $x\notin \operatorname{co} P(x)$ for all $x\in Z$. Here we used co A to denote the convex hull of a set A.

DEFINITION 3. Let Z be a convex subset of a topological vector space E and let $\emptyset \neq X \subset Z$. A correspondence $P: Z \to 2^X$ is said to be *generalized* SS-*convex* on X (cf. [17]) if for every finite subset $\{x_1, x_2, \cdots, x_m\}$ of X and $x_0 \in \operatorname{co}\{x_1, x_2, \cdots, x_m\}$, $x_j \notin P(x_0)$ for some $1 \leqq j \leqq m$.

Remark 6. Note that the SS-convexity implies the generalized SS-convexity. The converse statement may not be true unless $X = Z$.

Let Z be a convex subset in a topological vector space and let $\emptyset \neq X \subset Z$. A correspondence $G: X \to 2^Z$ is said to be FS-*convex* (refer to Fan [17] and Sonnenschein [14]) if for any finite set $\{x_1, x_2, \cdots, x_n\} \in X$, $\operatorname{co}\{x_1, x_2, \cdots, x_n\} \subset \bigcup_{i=1}^n G(x_i)$. Next we apply the transfer method to generalize the above conventional continuities and convexities. Once the definitions are compared, the ideas behind the transfer method become clear.

In the following definition, whenever $K = X$, "to K" will be replaced by "on X" or omitted.

DEFINITION 4. Let X be a set of a topological space Y and $K \subset X$ be a subset. The binary relation "$>$" on Y is said to be

(1) *transfer upper continuous* to K, if for any $x \in X$ and $y \in Y$, $x > y$ implies that there exist $x' \in K$ and $\mathcal{N}(y)$ such that $x' > \mathcal{N}(y)$;

(2) *transfer pseudoupper continuous* to K, if for any $x \in X$ and $y \in Y$, $x > y$ implies that there exist $x' \in K$ and $\mathcal{N}(y)$ such that $x' > y$ and $x' \geqq \mathcal{N}(y)$;

(3) *transfer weakly upper continuous* to K, if for any $x \in X$ and $y \in Y$, $x > y$ implies that there exist $x' \in K$ and $\mathcal{N}(y)$ such that $x' \geqq \mathcal{N}(y)$.

DEFINITION 5. Let X be a topological space and let Z be a convex subset in a topological vector space. A correspondence $G: X \to 2^Z$ is said to be *transfer* FS-*convex* on X if for any finite set $\{x_1, x_2, \cdots, x_n\} \subset X$ there exists a corresponding finite set $\{y_1, y_2, \cdots, y_n\} \subset Z$ such that for any subset $\{y_{i1}, y_{i2}, \cdots, y_{is}\}$ $(1 \leqq s \leqq n)$ of $\{y_1, y_2, \cdots, y_n\}$ we have

$$\operatorname{co}\{y_{i1}, y_{i2}, \cdots, y_{is}\} \subset \bigcup_{r=1}^{s} G(x_{ir}),$$

where $\{x_{i1}, x_{i2}, \cdots, x_{is}\}$ is a corresponding subset of $\{x_1, x_2, \cdots, x_n\}$.

DEFINITION 6. Let X be a topological space and let Z be a convex subset in a topological vector space. A correspondence $P: Z \to 2^X$ is said to be *transfer* SS-*convex* on X if for any finite set $\{x_1, x_2, \cdots, x_n\} \subset X$ there exists a corresponding finite set $\{y_1, y_2, \cdots, y_n\} \subset Z$ such that for any subset $\{y_{i1}, y_{i2}, \cdots, y_{is}\}$ $(1 \leqq s \leqq n)$ of $\{y_1, y_2, \cdots, y_n\}$ and $y_{i0} \in \operatorname{co}\{y_{i1}, y_{i2}, \cdots, y_{is}\}$ we have $x_{ir} \notin P(y_{i0})$.

DEFINITION 7. Let Z be a convex subset in a topological vector space and let $\emptyset \neq X \subset Z$. The binary relation $> (\geqq)$ is said to be *transfer* SS-*convex* (*transfer* FS-*convex*) on X if $U_s: Z \to 2^X$ $(U_w: X \to 2^Z)$ is transfer SS-convex (transfer FS-convex) on X.

Remark 7. Conventional convexity conditions give relations between a finite set $\{x_1, x_2, \cdots, x_n\}$ and its convex hull $\operatorname{co}\{x_1, x_2, \cdots, x_n\}$. Transfer convexity conditions give relations between a finite set $\{x_1, x_2, \cdots, x_n\}$ and the convex hull of a corresponding finite set $\{y_1, y_2, \cdots, y_n\}$, which may differ from $\{x_1, x_2, \cdots, x_n\}$.

DEFINITION 8. Let X and Y be two topological spaces. A correspondence $G: X \to 2^Y$ is said to be *transfer closed-valued* on X if for every $x \in X$, $y \notin G(x)$ implies that there exists $x' \in X$ such that $y \notin \operatorname{cl} G(x')$, i.e., $y \notin$ the closure of $G(x')$.

In the remainder of this section we prove several lemmas that give the interconnections between different definitions and that will be useful in later proofs.

LEMMA 3. (1) *Let Y be a topological space and let $\emptyset \neq X \subset Y$. Then the correspondence $U_w : X \to 2^Y$ is closed-valued on X if and only if "$>$" is upper continuous on X; the correspondence $U_w : X \to 2^Y$ is transfer closed-valued on X if and only if "$>$" is transfer upper continuous to X.*

(2) *Let Z be a convex subset in a topological vector space and let $\emptyset \neq X \subset Z$. Then the correspondence $U_w : X \to 2^Z$ is* FS-*convex on X if and only if $U_s : Z \to 2^X$ is generalized* SS-*convex on X, and the binary relation "$>$" is transfer* SS-*convex on X if and only if "$\geqq$" is transfer* FS-*convex on X.*

Proof. The proof follows immediately from the definitions. □

LEMMA 4. *Let Z be a nonempty convex subset of a topological vector space and let $\emptyset \neq X \subset Z$. Suppose "$>$" is a binary relation on Z such that $U_w : X \to 2^Z$ is finitely closed for each $x \in X$ (i.e., the intersection of $U_w(x)$ with any finite-dimensional subspace of Z is closed). Then "$>$" is transfer finitely maximal on X if and only if $>$ ($\geqq$) is transfer* SS-*convex (transfer* FS-*convex) on X.*

Proof. By [6], U_w has the finite intersection property if and only if U_w is transfer FS-convex on X and therefore if and only if U_s is transfer SS-convex on X. It is clear that U_w has the finite intersection property if and only if "$>$" is transfer finitely maximal on X. □

LEMMA 5. *Let Y be a topological space and let $\emptyset \neq X \subset Y$ and let "$>$" be a binary relation on Y. Then $\bigcap_{x\in X} \operatorname{cl} U_w(x) = \bigcap_{x\in X} U_w(x)$ if and only if U_w is transfer closed-valued or equivalently if and only if "$>$" is transfer upper continuous on X.*

Proof. Sufficiency. It is clear that $\bigcap_{x\in X} U_w(x) \subset \bigcap_{x\in X} \operatorname{cl} U_w(x)$. So we only need to show $\bigcap_{x\in X} \operatorname{cl} U_w(x) \subset \bigcap_{x\in X} U_w(x)$. Suppose $y \notin \bigcap_{x\in X} U_w(x)$. Then $y \notin U_w(z)$ for some $z \in X$. Since U_w is transfer closed-valued on X, there exists some $z' \in X$ such that $y \notin \operatorname{cl} U_w(z')$ and then $y \notin \bigcap_{x\in X} \operatorname{cl} U_w(x)$.

Necessity. Assume $\bigcap_{x\in X} \operatorname{cl} U_w(x) = \bigcap_{x\in X} U_w(x)$. If $y \notin U_w(x)$, then $y \notin \bigcap_{x\in X} \operatorname{cl} U_w(x) = \bigcap_{x\in X} U_w(x)$ and thus $y \notin \operatorname{cl} U_w(x')$ for some $x' \in X$. Thus U_w is transfer closed-valued on X. □

2. Maximization of binary relations on compact sets. There are two basic approaches in the literature to showing nonemptiness of the set of maximal elements on a nonempty compact set without assuming transitivity of the binary relation. One approach, under some convexity and continuity conditions, was developed by Fan [7], Sonnenschein [14], Shafer [11], Shafer and Sonnenschein [12], Yannelis and Prabhakar [21], and Tian [16], [17], among others. The other approach may be found in Bergstrom [3], Walker [19] (under acyclic and upper continuity assumptions), and Campbell and Walker [5] (under the 1-link transitivity, compactness for the space, and weakly upper continuity for the binary relation). In this section we generalize and unify the two approaches by giving several theorems that characterize the existence of maximal elements for a binary relation on a compact set. Theorem 1 characterizes the existence of maximal elements of a binary relation on a compact set in terms of transfer continuity (topological condition) for a given weakened transitivity condition. Theorem 2 characterizes the existence of maximal elements of a binary relation in terms of transfer transitivity for a given weakened topological condition (transfer continuity) and Theorem 3 characterizes the existence of maximal elements of a binary relation in terms of geometric conditions (transfer convexities) for a given weakened topological condition (transfer continuities).

LEMMA 6. *Let X be any subset of a topological space Y and let "$>$" be any binary relation on Y.*

(1) *If "$>$" has a maximal element on X, then "$>$" is transfer finitely maximal on X.*

(2) *If "$>$" has a maximal element on X, then "$>$" is transfer weakly upper continuous on X.*

(3) *If "$>$" is transfer upper continuous on X, then the set of all maximal elements on X is closed (possibly empty) in X. If "$>$" is fully transitive and the set of all maximal elements on X is nonempty and closed, then "$>$" is transfer upper continuous.*

Proof. The proof follows immediately from the definitions. □

THEOREM 1. *Let X be a nonempty compact topological space and let the binary relation "$>$" on X be transfer finitely strict maximal on X. Then "$>$" has a maximal element on X if and only if "$>$" is transfer weakly upper continuous on X.*

Proof. Sufficiency. Suppose, by way of contradiction, that "$>$" does not have a maximal element. Then for each $y \in X$, there exists $x \in X$ such that $x > y$. By the transfer weakly upper continuity of "$>$," there exist $x' \in X$ and a neighborhood $\mathcal{N}(y)$ such that $x' \geqq y'$ for all $y' \in \mathcal{N}(y)$. It follows that $X = \bigcup_{y \in X} \mathcal{N}(y)$. Since X is compact, there exist finite points $\{y_1, y_2, \cdots, y_n\}$ such that $X = \bigcup_{i=1}^{n} \mathcal{N}(y_i)$. Let x_i' be the associated point such that $x_i' \geqq y'$ for all $y' \in \mathcal{N}(y_i)$. Since we assume that there is no maximal element, for the finite subset $\{x_1', x_2', \cdots, x_n'\}$, by the transfer finitely strict maximal property there exists $x' \in X$ such that $x' > x_i'$, for all $i = 1, 2, \cdots, n$. However, $x' \in \mathcal{N}(y_j)$ for some $j = 1, 2, \cdots, n$. We have $x_j' \geqq x'$. It leads to a contradiction. So X has a maximal element.

Necessity. This follows from Lemma 6(1). □

THEOREM 2. *Let X be a nonempty compact topological space.*

(1) *Assume that the binary relation "$>$" is transfer upper continuous on X. Then the set of all maximal elements on X is nonempty and compact if and only if "$>$" is transfer finitely maximal on X.*

(2) *Assume that the binary relation "$>$" on X is asymmetric (i.e., 2-acyclic) and fully transitive. Then the set of all maximal elements on X is nonempty and compact on X if and only if "$>$" is transfer upper continuous on X.*

Proof of (1). The necessity follows from Lemma 6(1). We prove the sufficiency. Since "$>$" is transfer finitely maximal on X, for every finite subset $\{x_1, \cdots, x_n\}$, there is $x' \in X$ such that for each $i = 1, 2, \cdots, n, x' > x_i$ or $x' \geqq x_i$. Define $U_w(x) = \{y \in X : y \geqq x\}$. Thus U_w and then cl U_w have the finite intersection property. By Lemma 5, $\bigcap_{x \in X} U_w(x) = \bigcap_{x \in X} \text{cl } U_w(x) \neq \emptyset$ on the compact set X. So the set of all maximal elements on X, which is $\bigcap_{x \in X} U_w(x) = \bigcap_{x \in X} \text{cl } U_w(x)$, is nonempty and compact.

Proof of (2). The sufficiency follows from part (1) and we only need to prove the necessity. Notice that under the full transitivity, for any nonmaximal element y and any maximal element x we have $x > y$. Since the set of all maximal elements is closed, any nonmaximal element y has an $\mathcal{N}(y)$ that contains no maximal element. Therefore, for any maximal element x we have $x > y'$ for all $y' \in \mathcal{N}(y)$. That is, "$>$" is transfer upper continuous. □

Thus Theorem 1 generalizes the results of Campbell and Walker [5] by relaxing the weakly upper continuity and the 1-link transitivity (pseudotransitivity) of "$>$." It also generalizes the results of Tian and Zhou [18] by relaxing the full transitivity of "$>$." Theorem 2(1) generalizes the results of Fan [7], Sonnenschein [14], Shafer [11], Shafer and Sonnenschein [12], Yannelis and Prabhakar [21], and Tian [17] by relaxing the upper continuity and (generalized) SS-convexity of "$>$" and the convexity of X. Theorem 2(1) also generalizes the results of Bergstrom [3] and Walker [19] by relaxing the upper continuity and acyclicity of $>$. Thus our results unify two basic approaches to the existence of maximal elements by giving necessary and sufficient conditions.

Remark 8. If we compare the conditions of Theorems 1 and 2(1), we can find that there is a trade-off between the transfer transitivities and the transfer continuities (a trade-off between transitivity conditions and topological conditions): If one condition is weakened, then the other must be strengthened and vice versa. Many theorems we give below will also have this trade-off relation.

Theorem 3 below, which is obtained in Tian [16], is a special case of Theorem 2 (which needs to assume that X is a subset of a topological vector space). We state it here as an alternative.

THEOREM 3. *Let Z be a nonempty convex compact subset of a topological vector space and let $\emptyset \neq X \subset Z$. Let "$>$" be a transfer upper continuous binary relation on X. Then the set of all maximal elements of "$>$" on X is nonempty and compact if and only if "$>$" is transfer* SS-*convex on X.*

Lemma 3(2) and Lemma 4 give partial interconnections between Theorem 2(1) and Theorem 3.

Remark 9. At this point, it is quite natural to conjecture that the transfer finitely strict maximal condition in Theorem 1 might be further weakened, or to ask, for a binary relation on a compact set: What is the weakest possible transitivity condition under which the existence of maximal elements is equivalent to the transfer weakly upper continuity? This question is related to our understanding of the fundamental structures of mathematics, namely, topology, transitivity, and their interconnections. So far it is still an open question. However, Campbell and Walker [5] provide a clue. They construct an example [5] in which a binary relation is weakly upper continuous (and thus is transfer weakly upper continuous) and 0-link transitive but fails to have a maximal element on a nonempty compact set. Therefore, under the transfer weakly upper continuity, any transitivity condition proposed, other than the transfer finitely strict maximal condition, must be weaker than the 1-link transitive condition and independent of or stronger than the 0-link transitive condition.

For any function u, we can define a fully transitive binary relation "$>$" as follows: $x > y$ if and only if $u(x) > u(y)$. Thus the transfer continuities of a function u can be similarly defined. As direct consequences of the above theorems, we provide two corollaries that are generalizations of the classical Weierstrass theorem and are obtained in Tian and Zhou [18].

COROLLARY 1 [18]. *Let X be a nonempty compact topological space and let u $X \to \mathbf{R} \cup \{-\infty\}$ be a function. Then u attains its maximum on X if and only if u is transfer weakly upper continuous on X.*

COROLLARY 2 [18]. *Let X be a nonempty compact topological space and let $u: X \to \mathbf{R} \cup \{-\infty\}$ be a function. Then the set of maximum points of u on X is nonempty and compact if and only if u is transfer upper continuous on X.*

The following examples show that the above corollaries are very useful for us to see whether or not the maximum points of functions exist—even though these functions are very discontinuous.

Example 3. Consider a function u defined on the interval $X = [0, 1]$ by

$$u(x) = \begin{cases} 1 + x & \text{if } x \text{ is a rational number,} \\ x & \text{otherwise.} \end{cases} \tag{3}$$

We can easily see that u is not upper semicontinuous. In order to see that u is transfer upper continuous, for any neighborhood $\mathcal{N} \subset [0, 1]$, we may choose any rational number x' such that $\sup\{x \mid x \in \mathcal{N}\} \leqq x' \leqq 1$. In addition, by Corollary 2, we know the set of all maximal points is nonempty and compact. In fact, $x = 1$ is a unique maximum point of u on $[0, 1]$.

Example 4. Consider the so-called Dirichlet function u defined on the interval $X=[0,1]$ by

$$u(x)=\begin{cases}1 & \text{if } x \text{ is an irrational number,}\\ 0 & \text{if } x \text{ is a rational number.}\end{cases} \tag{4}$$

Note that u defined by (4) is clearly not transfer upper continuous. However, it is transfer weakly upper continuous by choosing x' as any irrational number. Thus, by Corollary 1, u has a maximum point. We can also easily see that the set of maximum points of u on $[0,1]$ is a set consisting of all irrational numbers and thus is not compact.

Example 5. Now if a function u is defined on the interval $X=[0,1]$ by

$$u(x)=\begin{cases}x & \text{if } 0\leqq x<1,\\ 0 & \text{if } x=1,\end{cases} \tag{5}$$

then u is not transfer weakly upper continuous on X. This is because for $y=1$ and $x\in(0,1)$, we cannot find any $x'\in X$ and neighborhood $\mathcal{N}(y)$ of y such that $u(z)\leqq u(x')$ for all $z\in\mathcal{N}(y)$. Thus, by Corollary 1, we know that u does not have any maximum point. In fact, we can easily see that u does not have a maximum point on $[0,1]$.

3. Maximization of binary relations on noncompact sets. For application purposes the compactness assumption of a set is sometimes too restrictive, especially when solving problems with data in infinite dimension. In this section we prove several theorems that give necessary and/or sufficient conditions for the existence of maximal elements of a binary relation on noncompact sets in terms of topological conditions (transfer continuities) for given transitivities, or in terms of transfer transitivities for given topological conditions (transfer continuities), or in terms of transfer convexities for given topological conditions. Thus, by applying our transfer method, we generalize almost all of the results in the literature and all results in the last section. Furthermore, by using the "transfer" feature of our transfer method, we are able to provide an approach with potential applications in constrained maximization.

Recall that the function u in Example 4 is not transfer upper continuous, but it is easy to see that it is transfer pseudo-upper continuous. So the transfer upper continuity strictly implies the transfer pseudo-upper continuity. To show that the transfer pseudo-upper continuity strictly implies the transfer weakly upper continuity, we set up the following example.

Example 6. Let $X=K$ be the unit disk in the complex plane $\mathbf{C}$. Define a binary relation "$\succ$" on X by

$$z_1\succ z_2 \quad \text{if } \arg(z_1)>\arg(z_2),$$

for all $z_1, z_2\in Z$. Since for the origin 0, its argument $\arg(0)$ is not defined, $0\succeq z$ for all $z\in Z$. So "$\succ$" is transfer weakly upper continuous. But when we observe the behavior of "$\succ$" around the point $z=(x,0)$ for $x>0$, we can see that "$\succ$" is not transfer pseudo-upper continuous. Obviously, "$\succ$" is 0-link transitive. As a matter of fact, "$\succ$" is also 1-link transitive. For if $z_1\succ z_2\succeq z_3\succ z_4$, then we have $\arg(z_1)>\arg(z_2)$ and $\arg(z_3)>\arg(z_4)$, which also implies that $z_3\neq 0$. It follows that $\arg(z_2)\geqq\arg(z_3)$. Therefore, $\arg(z_1)>\arg(z_4)$ or $z_1\succ z_4$. Thus "$\succ$" is 1-link transitive.

The above example shows that for a binary relation "$\succ$" under the 1-link transitivity, the transfer pseudo-upper continuity strictly implies the transfer weakly upper continuity. However, when a preference relation "$\succ$" is fully transitive, it is easy to see that the transfer pseudo-upper continuity is equivalent to the transfer weakly upper continuity. So the question may be asked: What is the weakest possible transitivity condition under which the transfer pseudo-upper continuity is equivalent to the transfer weakly upper continuity. We need to consider another question before we can answer this one.

For a binary relation "$>$" defined on a set X, if $x \in X$ is not a maximal element of "$>$" on X, then there exists an element $y \in X$ such that $y > x$. We may be concerned with the question of whether or not there exists a maximal element $x^* \in X$ such that $x^* > x$. In general, the answer is no. But under certain transitivity conditions, the answer is yes.

LEMMA 7. *If the binary relation "$>$" is 2-link transitive on X and $>$ has a maximal element, then for any nonmaximal element $x \in X$ there exists a maximal element $x^* \in X$ such that $x^* > x$.*

Proof. Suppose, by way of contradiction, that there is a nonmaximal element $y \in X$ such that $y \geqq x^*$ for every maximal element x^* on X. Then there is an element $z \in X$ such that $z > y$. Note that z must be a nonmaximal element since $y \geqq x^*$ for all maximal elements x^* on X. So there is an element $x \in X$ such that $x > z > y$. Let x^* be any maximal element on X. Then we have $z > y \geqq x^* \geqq x > z$, which implies that $z > z$ by the 2-link transitivity—a contradiction. □

It can be seen in the above proof that 2-link transitivity can be replaced by a weaker condition "$x_1 \geqq x_2 \geqq x_3 > x_4 \Rightarrow x_1 \geqq x_4$."

LEMMA 8. *Let "$>$" be a 2-link transitive binary relation on a topological space X. Then "$>$" is transfer weakly upper continuous if and only if it is transfer pseudo-upper continuous.*

Proof. We only need to show that under the assumption, the transfer weakly upper continuity implies the transfer pseudo-upper continuity. When "$>$" is transfer weakly upper continuous and $x > y$, there exists $x' \in X$ and $\mathcal{N}(y)$ of y such that $x' \geqq \mathcal{N}(y)$. If "$>$" has a maximal element on X, by Lemma 7, there exists a maximal element $x^* \in X$ such that $x^* > y$ and $x^* \geqq \mathcal{N}(y)$. If "$>$" does not have a maximal element on X, notice that the 2-link transitivity implies the transfer finitely strict maximal condition; then there exists $x_0 \in X$ such that $x_0 > \{x, x'\}$, or $x_0 > x > y$ and $x_0 > x' \geqq \mathcal{N}(y)$. Then it follows that $x_0 > y$ and $x_0 \geqq \mathcal{N}(y)$, by noting that the 2-link transitivity implies the 0-link transitivity. Therefore, "$>$" is transfer pseudo-upper continuous. □

In the following, we provide various necessary and sufficient conditions to characterize the existence of binary relations on noncompact sets.

THEOREM 4. *Let X be a topological space.*

(1) *The set of all maximal elements of the binary relation "$>$" on X is nonempty and closed if there exists a nonempty compact set $K \subset X$ such that "$>$" is transfer finitely maximal and transfer upper continuous to K.*

(2) *Assume that the binary relation "$>$" on X is 2-acyclic and fully transitive. Then the set of all maximal elements on X is nonempty and closed if and only if there exists a nonempty compact set $K \subset X$ such that "$>$" is transfer upper continuous to K.*

(3) *Assume that the binary relation "$>$" on X is 2-acyclic and fully transitive. Then the set of all maximal elements on X is nonempty and compact if and only if there exists a nonempty compact set $K \subset X$ such that "$>$" is transfer upper continuous to K and for each $y \in X \backslash K$, there exists $x \in K$ such that $x > y$.*

Proof of (1). By the existence result in Theorem 2(1), we can see that "$>$" has a maximal element x^* on K. Suppose "$>$" has no maximal element on X. By the transfer upper continuity to K, there exists $x \in K$ such that $x > x^*$. This leads to a contradiction. The closedness of the set of all maximal elements follows Lemma 6(3).

Proof of (2). The sufficiency follows from (1) and the necessity is similar to that of Theorem 2(2). Just let $K = \{x^*\}$, where x^* is any maximal element on X.

Proof of (3). The sufficiency is similar to that of (2) and we only need to note that all the maximal elements must be in K. The proof of the necessity is similar to that of Theorem 2(2) by letting K be the set of maximal elements on X. □

COROLLARY 3. *Let X be a topological space and $u: X \to \mathbf{R}$ be a function. Then*

(1) *the set of all maximal elements of u on X is nonempty and closed if and only if there exists a nonempty compact set $K \subset X$ such that u is transfer upper continuous to K;*

(2) *the set of all maximal elements of u on X is nonempty and compact if and only if there exists a nonempty compact set $K \subset X$ such that u is transfer upper continuous to K and for each $y \in X \backslash K$ there is $x \in K$ with $u(x) > u(y)$.*

THEOREM 5. *Let X be a topological space. Suppose that the binary relation "$>$" on X is* 1-*link transitive. Then X has a maximal element if there exists a nonempty compact set $K \subset X$ such that "$>$" is transfer pseudo-upper continuous to K.*

Proof. By Theorem 1, "$>$" has a maximal element x^* on K. Suppose that "$>$" has no maximal element on X. Then there exist $y \in X$ and $x \in K$ such that $x > y \geqq y > x^*$ by the transfer pseudo-upper continuity to K. But this implies $x > x^*$ by the 1-link transitivity, which contradicts the fact that x^* is a maximal element of "$>$" on K. So "$>$" has a maximal element on X. □

Remark 10. Note that in the example provided by Campbell and Walker [5] the binary relation is not only 0-link transitive and weakly upper continuous but also transfer pseudo-upper continuous on a compact set. However, it fails to have a maximal element. So the 1-link transitive assumption in Theorem 5 cannot be replaced by the 0-link transitivity. □

THEOREM 6. *Let X be a topological space and let the binary relation "$>$" on X be such that there exists a nonempty compact set $K_1 \subset X$ such that "$>$" is transfer finitely strict maximal to K_1. Then X has a maximal element if and only if there exists a nonempty compact subset $K_2 \subset X$ such that "$>$" is transfer weakly upper continuous to K_2.*

Proof. The necessity is trivial. Just let $K_2 = \{x^*\}$, where x^* is any maximal element on X. We only need to prove the sufficiency.

When there exists a nonempty compact subset $K_1 \subset X$ such that "$>$" is transfer finitely strict maximal to K_1, let $K = K_1 \cup K_2$. Then "$>$" is transfer finitely strict maximal and transfer weakly upper continuous to K. By Theorem 1, there is a maximal element x^* on K. Suppose, by way of contradiction, that "$>$" has no maximal element on X. Then there exists an element $y \in X$ such that $y > x^*$. But "$>$" is transfer finitely strict maximal to K; for the nonmaximal element $x^* \in X$ there exists $x \in K$ such that $x > x^*$, a contradiction. So X has a maximal element in K_1. □

The following corollary is a complete characterization for a function to attain its maximum values.

COROLLARY 4. *Let X be a topological space and let $u: X \to \mathbf{R}$ be a function. Then the set of all maximal elements of u on X is nonempty if and only if there exists a nonempty compact set $K \subset X$ such that u is transfer weakly upper continuous to K.*

THEOREM 7. *Let X be a topological space and let "$>$" be a binary relation on X. Assume that*

(1) *there is an element $x_0 \in X$ such that* cl $U_w(x_0)$ *is compact in X;*

(2) *U_w is transfer upper continuous on X.*

Then the set of all maximal elements of "$>$" on X is nonempty and compact if and only if "$>$" is transfer finitely maximal on X.

Proof. The necessity follows from Lemma 6(1). We only need to show the sufficiency. Since "$>$" is transfer finitely maximal on X, U_w has a finite intersection property on X and so does cl U_w. Now cl $U_w(x) \cap$ cl $U_w(x_0)$ is compact and has a finite intersection property as well. So $\bigcap_{x \in X}$ cl $U_w(x) \neq \emptyset$ and is compact. Since condition (2) is equivalent to $\bigcap_{x\in X} U_w(x)=\bigcap_{x\in X} \mathrm{cl}\, U_w(x)$, $\bigcap_{x\in X} U_w(x)$ is nonempty and compact. □

Similarly, we can extend Theorem 3 to cover binary relations on sets that are not convex or compact.

THEOREM 8 [16]. *Let Z be a nonempty convex subset of a topological vector space, X a nonempty subset of Z, and “$>$” a binary relation on Z. Assume that*

(1) *there is a vector $x_0\in X$ such that* $\mathrm{cl}\, U_w(x_0)$ *is compact in Z;*

(2) *U_w is transfer upper continuous on X;*

(3) *for each $y\in Z\backslash X$, there exists $x\in X$ such that $x>y$.*

Then the set of all maximal elements of “$>$” on X is nonempty and compact if and only if “$>$” is transfer SS*-convex on X.*

THEOREM 9. *Let X be a topological space and “$>$” be a transfer upper continuous binary relation on X. Then the set of all maximal elements on X is nonempty and closed if and only if there exists a nonempty compact subset $K\subset X$ such that “$>$” is transfer finitely maximal to K.*

Proof. The necessity is trivial. Just let $K=\{x^*\}$, where x^* is any maximal element on X. We only need to prove the sufficiency. First we show that

$$\bigcap_{x\in X} \mathrm{cl}\, U_w(x)\cap K\neq\emptyset. \tag{6}$$

In fact, since “$>$” is transfer finitely maximal to K, for any finite subset $\{x_1, x_2, \cdots, x_n\}\subset X$, there exists $y\in K$ such that $x\geqq x_i$, $i=1,2,\cdots,n$. That is,

$$\bigcap_{i=1}^{n} U_w(x_i)\cap K\neq\emptyset.$$

It follows that

$$\bigcap_{i=1}^{n} \mathrm{cl}\, U_w(x_i)\cap K\neq\emptyset.$$

However, for each $x\in X$, the set $\mathrm{cl}\, U_w(x)\cap K$ is compact and therefore

$$\bigcap_{x\in X} \mathrm{cl}\, U_w(x)\cap K\neq\emptyset.$$

Due to the assumption that “$>$” is transfer upper continuous, Lemma 5 reads

$$\bigcap_{x\in X} U_w(x)=\bigcap_{x\in X} \mathrm{cl}\, U_w(x),$$

which is a nonempty closed subset in X. This completes the proof. □

Remark 11. Theorems 7 and 9 are generalizations of Theorem 2(1). They coincide if X is compact. Note that there is a trade-off between Theorem 7 and Theorem 9. Assumption (1) in Theorem 7 has been removed in Theorem 9, but the condition that “$>$” is transfer finitely maximal on X in Theorem 7 has been strengthened to the condition that “$>$” is transfer finitely maximal to a compact subset $K\subset X$. As a result, the conclusion in Theorem 7 that the set of all maximal elements is nonempty and compact becomes weaker in Theorem 9. □

4. Concluding remarks. In this section we give some further remarks.

Let E (environment space) and Y (action space) be two topological spaces; let $F: E\to 2^Y$ be a nonempty-valued correspondence; and let “$>_e$” be a family of the binary relations on Y that depends on the parameter $e \in E$. Define a binary correspondence $P: E \times Y \to 2^Y$ by

$$P(e, y) = \{x \in Y: x \succ_e y\}$$

for $(e, y) \in E \times Y$. To study a family of maximization problems with respect to the parameterized binary relation "$\succ_e$," we define the maximum (marginal) correspondence $M: E \to 2^Y$, for each $e \in E$, as

$$M(e) = \{y \in F(e): P(e, y) \cap F(e) = \emptyset\}.$$

Berge [1], [2, p. 116] first studied various continuity properties of the maximum correspondence $M(e)$ for a simple case where

$$M(e) = \{y \in F(e): u(e, y) \geqq u(e, x), \forall x \in F(e)\}$$

for some function $u: E \times Y \to \mathbf{R}$. He proved that if u is a continuous function and F is a nonempty compact-valued continuous correspondence, then the maximum correspondence M is nonempty compact-valued and upper semicontinuous. Since then, this theorem, called Berge's Maximum Theorem, has become one of the most useful and powerful theorems in economics, optimization, and game theory. Walker [20] extended Berge's Maximum Theorem to maximization with respect to binary relations. He gave conditions under which M is an upper semicontinuous correspondence with compact (but possibly empty) values. In [18], a further generalization is obtained by giving necessary and sufficient conditions, but M is still possibly empty-valued. Just as Berge's Maximum Theorem can be used to prove the existence of Nash equilibrium and equilibrium for the generalized game with payoff functions, Walker's Maximum Theorem can be used to prove the existence of Nash equilibrium and equilibrium for the generalized game without ordered binary relations if the nonemptiness of the maximum correspondence $M(e)$ can be guaranteed. It is worth indicating that our work in this paper is partially motivated by this problem and the results established here can be applied to giving various conditions under which $M(e)$ are nonempty valued.

Let Y be a topological space and $X \subset Y$ be a subset. For a given (weak) binary relation "$\succ^*$" on Y, if a maximal element on X with respect to "$\geqq^*$" is defined as an element $x^* \in X$ such that for each $x \in X$, either $x^* \geqq^* x$ or x^* and x cannot be compared, then we can define a (strict) binary relation "$\succ$" as the *asymmetric part* of "$\geqq^*$," i.e., $y \succ x$ whenever $y \geqq^* x$ and not $x \geqq^* y$ and write the completion "$\geqq$" of "$\succ$" by $y \geqq x$ whenever $x \succ y$ does not hold. Then follow our definition that a maximal element of "$\succ$" on X is an element $x^* \in X$ such that $x^* \geqq x$ for all $x \in X$, which reads: For each $x \in X$ either $x^* \geqq x$ or x^* and x cannot compare. So these two definitions for maximal elements on X coincide and a maximization problem with respect to the (weak) binary relation can be converted to a maximization problem with respect to the (strict) binary relation. Note that the above-defined (strict) binary relation "$\succ$" is always asymmetric (2-acyclic).

Finally, we would like to mention that the results stated in the above sections can also be used to prove the existence of greatest elements for a weak (reflexive) binary relation "$\geqq^*$." Let Y be a topological space and $X \subset Y$ be a subset. For a weak binary relation "$\geqq^*$" on Y, a point $x^* \in X$ is said to be a *greatest element* of $\geqq^*$ on X if $x^* \geqq^* x$ for all $x \in X$. For this weak binary relation "$\geqq^*$," we can define a strict binary relation "$\succ^*$" as follows. $x \succ^* y$ if and only if not $y \geqq^* x$. Then we can easily see that $x^* \in X$ is a greatest element of $\geqq^*$ on X if and only if x^* is a maximal element of "$>^*$" on X. Thus proving the existence of a greatest element of a weak binary relation is equivalent to proving the existence of a maximal element of the reduced strict binary relation.

Acknowledgments. The authors are grateful to the three anonymous referees and to Professor J. Zowe for their valuable comments, suggestions, and corrections.

REFERENCES

[1] C. BERGE, *Espaces Topologiques et Fonctions Multivoques*, Donod, Paris, 1959.

[2] ———, *Topological Spaces*, E. M. Patterson, trans., Macmillan, New York, 1963.

[3] T. X. BERGSTROM, *Maximal elements of acyclic relations on compact sets*, J. Econom. Theory, 10 (1975), pp. 403-404.

[4] J. M. BORWEIN, *On the existence of Pareto efficient points*, Math. Oper. Res., 8 (1983), pp. 64-73.

[5] D. E. CAMPBELL AND M. WALKER, *Optimization with weak continuity*, J. Econom. Theory, 50 (1990), pp. 459-464.

[6] S. S. CHANG AND Y. ZHANG, *Generalized* KKM *theorem and variational inequalities*, J. Math. Anal. Appl., 159 (1991), pp. 208-223.

[7] K. FAN, *A generalization of Tychonoff's fixed point theorem*, Math. Ann., 142 (1961), pp. 305-310.

[8] F. FERRO, *A minimax theorem for vector-valued functions*, J. Optim. Theory Appl., 60 (1989), pp. 18-31.

[9] T. KIM AND M. K. RICHTER, *Nontransitive-Nontotal Consumer Theory*, J. Econom. Theory, 38 (1986), pp. 324-363.

[10] D. T. LUC, *An existence theorem in vector optimization*, Math. Oper. Res., 14 (1989), pp. 693-699.

[11] W. SHAFER, *The nontransitive consumer*, Econometrica, 42 (1974), pp. 913-919.

[12] W. SHAFER AND H. SONNENSCHEIN, *Equilibrium in abstract economies without ordered preferences*, J. Math. Econom., 2 (1975), pp. 345-348.

[13] D. SCHMEIDLER, *Competitive equilibrium in markets with a continuum of traders and incomplete preferences*, Econometrica, 37 (1969), pp. 578-585.

[14] H. SONNENSCHEIN, *Demand theory without transitive preferences, with application to the theory of competitive equilibrium*, in Preferences, Utility, and Demand, J. S. Chipman, L. Hurwicz, M. K. Richter, and H. Sonnenschein, eds., Harcourt Brace Jovanovich, New York, 1971.

[15] T. TANAKA, *Some minimax problems of vector-valued functions*, J. Optim. Theory Appl., 59 (1988), pp. 504-524.

[16] G. TIAN, *Necessary and sufficient conditions for the existence of maximal elements of preferences relations*, Working Paper 90-15, Texas A&M University, College Station, TX, 1990.

[17] ———, Generalizations of the FKKM *theorem and Ky-Fan minimax inequality, with applications to maximal elements, price equilibrium, and complementarity*, J. Math. Anal. Appl., to appear.

[18] G. TIAN AND J. ZHOU, *Transfer continuities, generalizations of the Weierstrass and maximum theorems—a characterization approach*, Working Paper 90-28, Texas A&M University, College Station, TX, 1990.

[19] M. WALKER, *On the existence of maximal elements*, J. Econom. Theory, 16 (1977), pp. 470-474.

[20] ———, *A generalization of the maximum theorem*, Internat. Econom. Rev., 20 (1979), pp. 267-270.

[21] N. C. YANNELIS AND N. D. PRABHAKAR, *Existence of maximal elements and equilibria in linear topological spaces*, J. Math. Econom., 12 (1983), pp. 223-245.

[22] P. L. YU, *Cone convexity, cone extreme points, and nondominated solutions in decision problems with multiobjectives*, J. Optim. Theory Appl., 14 (1974), pp. 319-377.

SIAM J. OPTIMIZATION
Vol. 2, No. 3, pp. 376-404, August 1992

ON THE CHOICE OF THE REGULARIZATION PARAMETER IN NONLINEAR INVERSE PROBLEMS*

K. ITO† AND K. KUNISCH‡

Abstract. This paper focuses on regularization techniques for nonlinear ill-posed inverse problems. Tikhonov regularization and regularization due to the use of norm constraints are analyzed. A model function technique is proposed to iteratively determine an optimal regularization parameter or the parameter characterizing the norm constraint, and to estimate the error in the data if it is not known a priori.

Key words. nonlinear least squares, Tikhonov regularization, sensitivity analysis, model functions

AMS(MOS) subject classifications. 49B, 65K, 35R30

1. Introduction. In this paper we study regularization of nonlinear ill-posed inverse problems. Here, ill posedness refers to the lack of continuous dependence of the solutions of the problem on its data. In a numerical ad hoc approach this may cause serious difficulties or failure of the algorithm. A common method for solving ill-posed problems in a stable manner is to replace the original problem by a family of "nearby" problems that have more amenable properties. This is referred to as regularization. In actual realizations of a regularization technique, the notion of "nearby" is expressed in one (or several) "regularization" parameters. One of the essential problems in the use of a regularization technique is the appropriate choice of the regularization parameter. The purpose of this paper is to describe, analyze, and test new techniques for the choice of the regularization parameter. Let us explain the approach that we propose by means of a specific example. We consider the estimation of the diffusion coefficient a in

$$-\operatorname{div}(a \operatorname{grad} u) + cu = f \quad \text{in } \Omega, \tag{1.1}$$

where Ω is a bounded domain in $\mathbb{R}^n$, c and f are known, and a feasible boundary condition is assumed to be satisfied by u on the boundary $\partial\Omega$ of Ω. The problem consists in determining the scalar-valued functional parameter a from an observation z of the system for which (1.1) is assumed to be a model. Problems of this nature arise, for instance, in groundwater flow modeling; see, e.g., [Y]. The problem can be formulated as inverting the parameter-to-solution mapping $a \to u(a)$ at z, i.e., to solve

$$u(a) = z \tag{1.2}$$

for a solution a^*. Since the solutions of (1.1) generally satisfy certain regularity properties, it is easy to see that (1.1) may have no solution, e.g., if z is not sufficiently smooth. Even if (1.2) had a solution, it would not depend continuously on z, unless a very weak and practically useless (distributional) norm was chosen; see, e.g., [CK], [EKN]. In addition, (1.2) is not a good starting point for numerical computations. For these reasons ill-posed inverse problems in general, and parameter estimation

* Received by the editors May 21, 1990; accepted for publication (in revised form) July 18, 1991.

† Department of Mathematics, University of Southern California, Los Angeles, California 90024. This research was supported in part by National Science Foundation grants UINT-8521208 and DMS-8818530 and by Air Force Office of Scientific Research grant F49620-86-C-0111.

‡ Institut für Mathematik, Technische Universität Graz, A-8010 Graz, Austria. This research was supported in part by the "Fonds zur Förderung der wissenschaftlichen Forschung" under grant P6005 and by Air Force Office of Scientific Research grant F49620-86-C-0111.

problems like the one above in particular, are frequently formulated as optimization problems. For the problem under consideration this may be done by considering

$$\min_{a \in Q_{ad}} \tfrac{1}{2}|u(a)-z|^2_{\mathscr{X}}, \tag{1.3}$$

where $\mathscr{X}$ denotes a practically relevant topology for the observation space, and Q_{ad} describes the set of feasible diffusion coefficients, and incorporates, e.g., regularity properties and a positivity constraint $a(x) \geqq \alpha > 0$. In (1.3), the state u of (1.1) is considered as a function of the variable a. We have recently developed an approach in which the state variable u and the unknown parameter a are both independent variables, and the partial differential equation, formally expressed as $e(a, u)=0$, is considered as an explicit constraint [IK], [IKK]. Since the technique was numerically very successful, we also specify it formally here, and consider

$$\min \tfrac{1}{2}|\mathrm{u}-\mathrm{z}|^2_{\mathscr{X}}, \quad \mathrm{e(a, u)}=0, \quad \mathrm{a} \in \mathrm{Q}_{\mathrm{ad}}. \tag{1.4}$$

In the case in which z is attainable, i.e., if there exists $a^* \in Q_{ad}$ such that $u(a^*)=z$, a^* is a solution of (1.3) and $(a^*, u(a^*))$ is a solution of (1.4). Since the solution of (1.2) does not depend continuously on z, the same is true for (1.3) and (1.4). The optimization problems, however, readily lend themselves to various regularization techniques. We consider two such techniques, one of which is based on adding a Tikhonov-type regularization term to the cost functional, and another that employs a (semi)norm constraint of the unknown parameter. Combining both techniques in one formulation, (1.3) and (1.4) become, respectively,

$$\min \frac{1}{2}|u(a)-z|^2_{\mathscr{X}} + \frac{\beta}{2}\langle a, Pa\rangle, \quad a \in Q_{ad}, \quad \langle a, Pa\rangle \leqq \gamma, \tag{1.5}$$

and

$$\begin{gathered} \min \frac{1}{2}|u-z|^2_{\mathscr{X}} + \frac{\beta}{2}\langle a, Pa\rangle, \\ a \in Q_{ad}, \quad e(a, u)=0, \qquad \langle a, Pa\rangle \leqq \gamma, \end{gathered} \tag{1.6}$$

where P is a bounded linear self-adjoint nonnegative operator describing a norm or a seminorm on the coefficient space. For $\beta \geqq 0$ and $\gamma \leqq \infty$ let $a^{\beta,\gamma}$ denote the solution of (1.5) and let $(a^{\beta,\gamma}, u^{\beta,\gamma})$ denote the solution of (1.6). The question that we address is the choice of β and/or γ. If the data z were error free and attainable, and if we could solve (1.5) or (1.6) with infinite precision, then we could take $\beta=0$ and $\gamma=\infty$. Since this is not the case, we choose $\beta>0$ or $\gamma<\infty$ to stabilize an otherwise unstable problem; see, e.g., [CK], [IK], and [EKN]. Our plan, generally speaking, is to determine β or γ iteratively, starting with a large value of β and/or a small value of γ (which allows for a stable solution of (1.5) or (1.6) but introduces a large regularization error) and to then iteratively decrease β or to increase γ until the problem approaches ill posedness. We will stop just before ill posedness would render the solution of (1.5) or (1.6) infeasible. To accomplish this goal we introduce model functions. Let us consider the special case (1.5) with $\gamma=\infty$ and let $F:\mathbb{R}^+ \to \mathbb{R}^+$ be given by

$$F:\beta \to \frac{1}{2}|u(a^\beta)-z|^2_{\mathscr{X}} + \frac{\beta}{2}\langle a^\beta, Pa^\beta\rangle,$$

with a^β a solution of (1.5) with $\gamma=\infty$. Under appropriate conditions this solution is locally unique. Using a simplifying assumption, it will be shown that F is the solution of a two-parameter family of second-order ordinary differential equations. Their solution is a four-parameter family of functions that we will denote by m and that describes F. Determining the four parameters characterizing m requires solving (1.5) for two values of β. Once m is determined we require further principles to obtain an appropriate value for the regularization parameter. We may use the well-known Morozov principle, which requires us to choose the regularization parameter β_{M} such that

$$|u(a^{\beta M})-z|^2_{\mathscr{X}}=\delta^2, \tag{1.7}$$

where δ is the expected error level for the observation z. We will see that (1.7) can be expressed as $F(\beta_M)-\beta_M F'(\beta_M)=\delta^2/2$, which is approximated, employing the model function, by

$$m(\beta_M)-\beta_M m'(\beta_M)=\frac{\delta^2}{2}. \tag{1.8}$$

In our experiments (with and without the use of model function) the choice of the regularization parameter gave somewhat conservative results, and we will therefore also introduce an alternative to the Morozov principle given by

$$m(\beta_P)-(\beta_P^{\gamma_1}-\beta_P)m'(\beta_P)=\frac{\gamma_2}{2}\,\delta^2 \tag{1.9}$$

for constants $\gamma_1\in[1,2]$ and $\gamma_2\in[\frac{1}{2},1]$.

If m coincided with F, then our algorithm could stop here. But m is only an approximation to F and hence we repeat the above sequence of calculating a model function and updating the regularization parameter according to some principle, e.g., the Morozov principle, several times, until a stopping criterion is reached.

As already seen in (1.8), knowledge of the error level δ was used in the algorithm to determine the regularization parameter. If the error level is not available, then our algorithm allows us to estimate δ through evaluation of (1.5) for two appropriately chosen values of β.

While this paper is motivated by nonlinear inverse problems, some aspects of it may also be new for linear problems. The theory of linear ill-posed inverse problems has received a considerable amount of attention and we refer the reader to several books on this subject, e.g.,[TA], [B], [G], [L], and [M].

The paper is organized as follows. In § 2 we consider general nonlinear ill-posed problems with either Tikhonov regularization or regularization due to constraints. An existence result in a seminorm setting is given and the relationship between the solution using Tikhonov-type regularization and the solution using norm constraints is analyzed. As mentioned above, the model functions are obtained as the solutions of ordinary differential equations. These equations can be obtained due to the differentiability properties of the optimal value function, as well as of the solutions to regularized nonlinear least squares problems such as (1.5) and (1.6) with respect to β and γ. In § 3 we present results on the directional differentiability of the solutions of the regularized problems with respect to regularization parameters. These results will also provide first-order information on the optimal value function from knowledge of the solution to (1.5), respectively, (1.6) (see Theorem 3.1). Once these analytical preliminaries are established, we develop our results on the optimal choice of the regularization parameter. The case of Tikhonov regularization is treated in § 4 and regularization by constraints of the parameters is treated in § 5. Both these sections contain numerical results illustrating the proposed methods.

2. The $(\mathscr{P}^\beta)$ and $(\mathscr{P}^\gamma)$ problems and differentiability. We are concerned with the minimization problem

$$\begin{aligned} &\min f(a, u)+\frac{\beta}{2}\langle Pa, a\rangle \\ (2.1) \qquad & \\ &\text{subject to } e(a, u)=0, \quad \langle Pa, a\rangle \leqq \gamma, \quad 1(a) \in K, \end{aligned}$$

where

$$\begin{aligned} & f: Q \times X \rightarrow \mathbb{R}^{+}=[0, \infty), \\ & P: Q \rightarrow Q, \\ & e: Q \times X \rightarrow Y, \\ & l: Q \rightarrow Z, \\ & \beta \in \mathbb{R}^{+}, \qquad \gamma \in \mathbb{R}^{+} \cup\{\infty\}, \\ & K \text{ is a closed convex cone with vertex at zero in } Z, \end{aligned}$$

and Q, X, Y, and Z are real Hilbert spaces, P is a bounded, linear, selfadjoint nonnegative operator, and l is affine and continuous. Under the above assumptions, every element $a \in Q$ can be decomposed uniquely as

$$a=a^{(1)}+a^{(2)},$$

where $a^{(1)} \in \ker P$ and $a^{(2)} \in (\ker P)^{\perp}$.

The following additional hypotheses will be used.

(H1) There exists $m>0$ such that

$$\langle Pa, a\rangle \geqq m|a|_Q^2 \quad \text{for all } a \in(\ker P)^{\perp}.$$

(H2) For every $a \in Q$ with $l(a) \in K$ there exists a unique element $u(a) \in X$ such that $e(a, u(a))=0$.

(H3) For any weakly convergent sequence a_n with w-lim $a_n=a$ and $l(a_n) \in K$, we have $\lim_{n \rightarrow \infty} f(a_n, u(a_n)) \geqq f(a, u(a))$ and $\lim_{n \rightarrow \infty} e(a_n, u(a_n))=e(a, u(a))$.

(H4) If $\{a_n\}$ is a sequence in Q with $l(a_n) \in K$ and $\langle Pa_n, a_n\rangle \leqq \gamma$ for all n, and with $\{|a_n^{(2)}|_Q\}$ bounded and $\{|a_n^{(1)}|_Q\}$ unbounded, then $\{a_n\}$ cannot be a minimizing sequence for (2.1).

Throughout it is assumed that the set

$$Q_{ad}=\{a \in Q\colon l(a) \in K, \langle Pa, a\rangle \leqq \gamma\}$$

is nonempty.

THEOREM 2.1. *Let* (H1)-(H4) *hold and let* $\beta>0$ *or* $\gamma<\infty$. *Then there exists a solution* (a_0, u_0) *of* (2.1).

Proof. Let (a_n, u_n) be a minimizing sequence for (2.1). Due to (H2) we can equivalently consider $\{a_n\}$ as a minimizing sequence. Clearly, $\{Pa_n\}$ is bounded in Q and hence $\{a_n^{(2)}\}$ is bounded by (H1). Condition (H4) then implies that $\{a_n^{(1)}\}$ is bounded as well. Hence $\{a_n\}$ is bounded and there exists a weakly convergent subsequence, again denoted by $\{a_n\}$, and $a_0 \in Q$, such that w-lim $a_n=a_0$ and $l(a_0) \in K$. Finally, (H3) and weak lower semicontinuity of the norm in Q imply that $(a_0, u_0)=(a_0, u(a_0))$ is a solution of (2.1).

We turn to a brief discussion of the hypotheses (H1)-(H4). Hypothesis (H1) holds, for example, if P has closed range. In this case there exists $\hat{m} \geq 0$ such that $|Pa|_Q \geqq \hat{m}|a|_Q$ for all $a \in \ker P$ and it follows that (H1) holds with $m = \sqrt{\hat{m}}$. With (H2) holding, the equality constraint could be eliminated from the constraints in (2.1) and incorporated in the function f. We prefer to keep $e(a, u) = 0$ as an explicit constraint since it allows greater flexibility, for example, in the numerical implementation of (2.1) by Lagrangian techniques. In view of (H2) and Theorem 2.1 we will henceforth refer to both a_0 and the pair $(a_0, u_0) = (a_0, u(a_0))$ as a solution of (2.1). (H3) is one of several possible choices of a compactness-type assumption that is required to quarantee the existence of a solution to (2.1). Regarding (H4) we have the following result.

PROPOSITION 2.2. *Let* (H2) *hold and assume that* $\ker P_1 \cap Q_{ad}$ *is not empty. If f has the property that for any sequence $\{a_n\}$ in $Q_{ad} \subset Q$ with $\{|a_n^{(2)}|_Q\}$ bounded and $\{|a_n^{(1)}|_Q\}$ unbounded,*

$$\liminf_n f(a_n, u(a_n)) > \inf \{f(a, u(a)) \colon a \in \ker P \cap Q_{ad}\}$$

is satisfied, then (H4) *holds.*

Proof. Assume that $\{a_n\}$ is a sequence in Q_{ad} with $\{|\mathrm{a}_n^{(2)}|_Q\}$ bounded and $\{|a_n^{(1)}|_Q\}$ unbounded, and that it is also a minimizing sequence. Then we have

$$\begin{aligned}\liminf_n f(a_n, u(a_n)) &> \inf \{f(a, u(a)) \colon a \in \ker P \cap Q_{ad}\}\\ &\geqq \inf \{f(a, u(a)) \colon a \in Q_{ad}\}\\ &= \liminf_n \left(f(a_n, u(a_n)) + \frac{\beta}{2} |Pa_n|_Q^2 \right)\\ &\geqq \liminf_n f(a_n, u(a_n)).\end{aligned}$$

This is a contradiction and hence $\{a_n\}$ cannot be a minimizing sequence.

Remark 2.3. The hypothesis on f of Proposition 2.2 is applicable to parameter estimation problems; see § 4 and [IK2]. Another condition on f that implies (H4) is given by requiring that $f(a_n, u(a_n))$ is unbounded for any sequence $\{a_n\}$ in Q_{ad} with $\{a_n^{(2)}\}$ bounded and $\{a_n^{(1)}\}$ unbounded. It corresponds to an analogous assumption in the theory of linear problems, which requires that f be radially unbounded on $\ker P$; see, e.g., [G].

Remark 2.4. In applications, P may be defined through

$$\langle Px, y\rangle_Q = \langle\!\langle x, y\rangle\!\rangle \quad \text{for all } x, y \in Q, \tag{2.2}$$

where $\langle\!\langle \cdot, \cdot \rangle\!\rangle$ is a nonnegative continuous sesquilinear form on Q, [K], which has the property that

$$\langle\!\langle x, x\rangle\!\rangle \geqq \tilde{m}|x|_Q^2 \tag{2.3}$$

for some $\tilde{m} > 0$ independent of $x \in N^{\perp}$, with $N = \{x \colon \langle\!\langle x, x\rangle\!\rangle = 0\}$.

In this case P defined through (2.2) is a bounded, linear, nonnegative, self-adjoint operator and $\ker P = N$. It is obvious that $\ker P \subset N$. To show the converse inclusion let $x \in N$ and $y \in Q$ be arbitrary. We find

$$0 \leqq |\langle\!\langle x, y\rangle\!\rangle| \leqq \sqrt{\langle\!\langle x, x\rangle\!\rangle}\,\sqrt{\langle\!\langle y, y\rangle\!\rangle} = 0$$

and therefore $\langle\!\langle x, y\rangle\!\rangle = 0$ for all $y \in Q$. This implies that $x \in \ker P$ and that $\ker P = N$. Hypothesis (H1) now follows from (2.3). A specific example for this setup is given by

$$Q = H^1(0, 1) \quad \text{and} \quad \langle\!\langle x, y\rangle\!\rangle = \langle Dx, Dy\rangle_{L^2(0,1)},$$

where D denotes differentiation. In this case $N = \{x \colon x \text{ is constant}\}$ and $\tilde{m} = \pi^{-2}$.

The use of the regularization term $(\beta/2)\langle Pa, a\rangle$ and/or the constraint $\langle Pa, a\rangle \leqq \gamma$ not only guarantees existence of a solution to (2.1), but also stabilizes problem (2.1) in the sense of guaranteeing the continuous dependence of the solutions of (2.1) on perturbations in f, e, and l. For the case of quadratic problems (i.e., $a \to f(a, u(a))$) is quadratic) we refer to [TA] and [G], for general nonlinear problems, to [EKN], [SV], and [V], and for parameter estimation problems, to [CK], [IK2], and [KS], for example. If the problem with $\beta = 0$ and $\gamma = \infty$ is not well posed, then the introduction of a regularization term or of a constraint on the unknowns enhances well posedness, while at the same time a new error resulting from these terms is introduced. From a practical point of view, therefore, one of the most important questions in solving (2.1) in a stable manner is the choice of β and/or γ. Before we address this question we study the solutions of (2.1) as functions of β and γ. For future reference we specify the two problems in which only the regularization term or the constraint $\langle Pa, a\rangle \leqq \gamma$ are used:

$$(\mathcal{P}^\beta) \qquad \begin{aligned} &\min f(a, u) + \frac{\beta}{2}\langle Pa, a\rangle \\ &\text{subject to } e(a, u) = 0, \quad l(a) \in K, \end{aligned}$$

and

$$(\mathcal{P}^\gamma) \qquad \begin{aligned} &\min f(a, u) \\ &\text{subject to } e(a, u) = 0, \quad \langle Pa, a\rangle \leqq \gamma, \quad l(a) \in K. \end{aligned}$$

Under the assumptions of Theorem 2.1 there exist solutions $(a^\beta, u^\beta) = (a^\beta, u(a^\beta))$ of $(\mathcal{P}^\beta)$ for any $\beta > 0$ and $(a^\gamma, u^\gamma) = (a^\gamma, u(a^\gamma))$ for any $\gamma < \infty$ (provided, of course, that $Q_{ad} \neq \emptyset$).

Concerning the relationship between $(\mathcal{P}^\beta)$ and $(\mathcal{P}^\gamma)$ it is easy to check that any solution $(a^\beta, u(a^\beta))$ of $(\mathcal{P}^\beta)$ is also a solution of $(\mathcal{P}^\gamma)$ if $\gamma = \langle Pa^\beta, a^\beta\rangle$. For the converse, additional hypotheses will be needed. Let (a_0, u_0) be a solution of (2.1).

(H5) f and e are continuous and twice continuously differentiable on $\{(a, u(a)) : l(a) \in K\}$.

(H6) (a_0, u_0) is a regular point with respect to the constraints in (2.1), i.e.,

$$\begin{pmatrix} e'(a_0, u_0)(Q, X) \\ \langle Pa_0, \cdot\rangle(Q) \\ 1'(Q) \end{pmatrix} + \begin{pmatrix} 0 \\ \mathbb{R}_+ \\ -K \end{pmatrix} + \mathbb{R}\begin{pmatrix} 0 \\ \langle Pa_0, a_0\rangle \\ l(a_0) \end{pmatrix} = \begin{pmatrix} Y \\ \mathbb{R} \\ Z \end{pmatrix}.$$

Here $e'(a, u)$ denotes the Fréchet derivative of e at (a, u) and l' stands for the derivative of l, which is independent of the point where it is taken. The Lagrangian $\mathcal{L} : Q \times X \times Y \times \mathbb{R} \times Z \to \mathbb{R}$ is given by

$$\begin{aligned} \mathcal{L}(a, u, \lambda, \mu, \eta) = f(a, u) + \frac{\beta}{2}\langle Pa, a\rangle + \langle \lambda, e(a, u)\rangle \\ + \frac{\mu}{2}(\langle Pa, a\rangle - \gamma) + \langle \eta, l(a)\rangle. \end{aligned} \tag{2.4}$$

If (a_0, u_0) is a regular point, then there exists a Lagrange multiplier $(\lambda_0, \mu_0, \eta_0) \in Y \times \mathbb{R}_+ \times K_+$, with $K_+ = \{z \in Z : \langle z, k\rangle \leqq 0 \text{ for all } k \in K\}$, the dual cone of K such that

$$\begin{aligned} &\mathcal{L}'(a_0, u_0, \lambda_0, \mu_0, \eta_0) = 0, \\ &e(a_0, u_0) = 0, \mu_0(\langle Pa_0, a_0\rangle - \gamma) = 0, \qquad \langle \eta_0, l(a_0)\rangle = 0, \end{aligned} \tag{2.5}$$

where $\mathscr{L}'$ denotes the derivative with respect to (a, u). The following second-order sufficient optimality condition will be used frequently throughout this paper [MZ].

(H7) There exists $\kappa > 0$ such that

$$\mathscr{L}''(a_0, u_0, \lambda_0, \mu_0, \eta_0)((h, v), (h, v)) \geqq \kappa|(h, v)|^2_{Q\times X}$$

for all $(h, v) \in \ker e'(a_0, u_0)$.

Without change of notation we will refer to (H6) and (H7) in the context of the solution (a^β, u^β) and (a^γ, u^γ) of the special problems $(\mathscr{P}^\beta)$ and $(\mathscr{P}^\gamma)$. The Lagrange multipliers will be denoted by $(\lambda^\beta, \mu^\beta)$ and $(\lambda^\gamma, \mu^\gamma, \eta^\gamma)$ in this case.

THEOREM 2.5. *Let* (H1)-(H4) *hold. Then any solution* (a^β, u^β) *of* $(\mathscr{P}^\beta)$ *is also a solution of* $(\mathscr{P}^\gamma)$ *with* $\gamma = \langle Pa^\beta, a^\beta\rangle$. *Conversely, if* (a^γ, u^γ) *is a solution of* $(\mathscr{P}^\gamma)$ *and* (H5)-(H7) *hold at* (a^γ, u^γ), *then* (a^γ, u^γ) *is a solution of* $(\mathscr{P}^\beta)$ *with* $\beta = \mu^\gamma$.

Proof. The first assertion is obvious. To verify the converse claim, let (a^γ, u^γ) be a solution of $(\mathscr{P}^\gamma)$. Due to (H5)-(H7) there exists a Lagrange multiplier $(\lambda^\gamma, \mu^\gamma, \eta^\gamma) \in Y \times \mathbb{R}^+ \times K_+$ for $(\mathscr{P}^\gamma)$ such that

$$f''(a^\gamma, u^\gamma)(z, z) + \langle\lambda^\gamma, e''(a^\gamma, u^\gamma)(z, z)\rangle + \mu^\gamma\langle Ph, h\rangle \geqq \kappa|z|^2_{Q\times X} \tag{2.6}$$

for all $z = (h, v) \in \ker e'(a^\gamma, u^\gamma)$. It is easy to check that $(\lambda^\gamma, \eta^\gamma)$ is a Lagrange multiplier for $(\mathscr{P}^\beta)$ with $\beta = \mu^\gamma$. Moreover, (2.6) is a second-order sufficient optimality condition for (a^γ, u^γ) to be a minimum for $(\mathscr{P}^\beta)$ with $\beta = \mu$ [MZ]. This concludes the proof.

3. Differentiability properties. In this section we investigate differentiability properties of the Lagrange functionals associated with $(\mathscr{P}^\beta)$ and $(\mathscr{P}^\gamma)$, and of the solutions (a^β, u^β) and (a^γ, u^γ) with respect to β and γ, respectively. The results we will present will be essential in justifying the model functions to be developed in §§ 4 and 5. They follow from the general sensitivity analysis developed in [IK2]. We specify some additional notation and hypotheses for a solution (a_0, u_0) of (2.1). Without change of notation we will use these concepts for solutions of the special problems $(\mathscr{P}^\beta)$ and $(\mathscr{P}^\gamma)$. We define by $B: Q\times X \to Q\times X$ the operator representation of

$$f''(a_0, u_0) + \langle\lambda_0, e''(a_0, u_0)(\cdot, \cdot)\rangle \quad \text{such that}$$

$$\langle By, z\rangle_{Q\times X} = f''(a_0, u_0)(y, z) + \langle\lambda_0, e''(a_0, u_0)(y, z)\rangle$$

for all y and z in $Q\times X$, and by $E: Q\times X \to Y$ the operator

$$E = e''(a_0, u_0).$$

Since the discussion in this section is of a local nature, the inequality constraint $\langle Pa, a\rangle \leqq \gamma$ can be dropped if $\langle Pa_0, a_0\rangle < \gamma$, and we therefore consider only the case $\langle Pa_0, a_0\rangle = \gamma$ here.

(H8) The operator $\mathscr{B}: Q\times X\times\mathbb{R} \to Y\times\mathbb{R}\times Z$ given by

$$\mathscr{B}(h, v, r) = \begin{pmatrix} E(h, v) \\ \frac{1}{2}\langle Pa_0, h\rangle \\ Lh + rl(a_0) \end{pmatrix}$$

is surjective.

Assume next that (H2) and (H5)-(H8) hold at a solution $(a^{\hat\beta}, u^{\hat\beta})$ of $(\mathscr{P}^{\hat\beta})$ for some $\hat\beta > 0$, and let $(\lambda^{\hat\beta}, \eta^{\hat\beta})$ be an associated Lagrange multiplier. Then by Theorem 2.1 of [IK2], there exist neighborhoods $V(\hat\beta)$ of $\hat\beta$ and $V(w(\hat\beta))$ of $w(\hat\beta) = (a^{\hat\beta}, u^{\hat\beta}, \lambda^{\hat\beta}, \eta^{\hat\beta})$ such that for all $\beta \in V(\hat\beta)$ there exists a quadruple $(a^\beta, u^\beta, \lambda^\beta, \eta^\beta) \in V(w(\hat\beta))$ of a solution (a^β, u^β) to $(\mathscr{P}^\beta)$ and of an associated Lagrange multiplier

$(\lambda^\beta, \eta^\beta)$, which depends Lipschitz continuously on $\beta \in V(\hat\beta)$. In particular, for $\beta \in V(\hat\beta)$ the solution $w(\beta)$ to $(\mathcal{P}^\beta)$ is unique within $V(w(\hat\beta))$. Henceforth, when referring to a (unique) solution $w(\beta)$ of $(\mathcal{P}^\beta)$ it is always understood in this local sense. The Lagrangian for the $(\mathcal{P}^\beta)$ problem is denoted by $\mathcal{L}(a, u, \lambda, \eta)$ and is defined as in (2.4) with the term $(\mu/2)(\langle Pa, a\rangle - \gamma)$ deleted. Similarly, if (H2) and (H5)-(H8) hold at a solution $(a^{\hat\gamma}, u^{\hat\gamma})$ of $(\mathcal{P}^{\hat\gamma})$ for some $\hat\gamma < \infty$, and if $(\lambda^{\hat\gamma}, \mu^{\hat\gamma}, \eta^{\hat\gamma})$ is an associated Lagrange multiplier, then there exist neighborhoods $V(\hat\gamma)$ of $\hat\gamma$ and $V(w(\hat\gamma))$ of $w(\hat\gamma) = (a^{\hat\gamma}, u^{\hat\gamma}, \lambda^{\hat\gamma}, \mu^{\hat\gamma}, \eta^{\hat\gamma})$ such that for all $\gamma \in V(\hat\gamma)$ there exists a quintuple $(a^\gamma, u^\gamma, \lambda^\gamma, \mu^\gamma, \eta^\gamma) \in V(w(\hat\gamma))$ consisting of a solution to $(\mathcal{P}^\gamma)$ and an associated Lagrange multiplier, which depends Lipschitz continuously on $\gamma \in V(\hat\gamma)$. We introduce the optimal value functions for $(\mathcal{P}^\beta)$ and $(\mathcal{P}^\gamma)$:

$$F(\beta) = \min\left\{ f(a, u) + \frac{\beta}{2}\langle Pa, a\rangle : e(a, u) = 0, l(a) \in K \right\} \tag{3.1}$$

and

$$F(\gamma) = \min\{f(a, u): e(a, u) = 0, \langle Pa, a\rangle \leqq \gamma, l(a) \in K\}. \tag{3.2}$$

THEOREM 3.1. *Let* (H2) *and* (H5)-(H8) *hold at a solution* $(a^{\hat\beta}, u^{\hat\beta})$ *of* $(\mathcal{P}^{\hat\beta})$, $\hat\beta > 0$, *or* $(a^{\hat\gamma}, u^{\hat\gamma})$ *of* $(\mathcal{P}^{\hat\gamma})$, $\hat\gamma < \infty$, *respectively. Then the functions defined in* (3.1) *or* (3.2) *are differentiable at* $\hat\beta$ *or* $\hat\gamma$, *respectively, and*

$$\frac{d}{d\beta} F(\hat\beta) = \frac{d}{d\beta} \mathcal{L}(a^{\hat\beta}, u^{\hat\beta}, \lambda^{\hat\beta}, \eta^{\hat\beta}) = \frac{1}{2}\langle Pa^{\hat\beta}, a^{\hat\beta}\rangle$$

or

$$\frac{d}{d\gamma} F(\hat\gamma) = \frac{d}{d\gamma} \mathcal{L}(a^{\hat\gamma}, u^{\hat\gamma}, \lambda^{\hat\gamma}, \mu^{\hat\gamma}, \eta^{\hat\gamma}) = -\frac{1}{2}\mu^{\hat\gamma}.$$

This result follows from Proposition 3.1 in [IK2]. The result in [IK2] in turn depends on the Lipschitz continuity of the mappings $\beta \to w(\beta)$ at $\hat\beta$ or $\gamma \to w(\gamma)$ at $\hat\gamma$, on the uniqueness of the Lagrange multipliers due to (H8), and on the sensitivity analysis of Lempio and Maurer [LM]. The significance of this theorem is given by the fact that the sensitivity of the optimal value functions $F(\beta)$ and $F(\gamma)$ at specific values $\hat\beta$ and $\hat\gamma$ can be expressed in terms of the solutions of $(\mathcal{P}^{\hat\beta})$ and $(\mathcal{P}^{\hat\gamma})$, respectively.

Remark 3.2. Let us point out that under the assumptions of Theorem 3.1 the conclusions of this theorem remain valid in a neighborhood $\hat V(\hat\beta)$ of $\hat\beta$ or $\hat V(\hat\gamma)$ of $\hat\gamma$, respectively. Since the arguments for the $(\mathcal{P}^\beta)$ and $(\mathcal{P}^\gamma)$ problems are analogous, we only give it for the former. First, (H2) and (H6) are global assumptions, and (H6) is used only to guarantee existence of a Lagrange multiplier, which we know to exist for β sufficiently close to $\hat\beta$ from the discussion above. Since, moreover, $w(\beta)$ depends continuously on β for β sufficiently close to $\hat\beta$, hypotheses (H7) and (H8) hold in a neighborhood of $\hat\beta$. Thus, under the assumptions of Theorem 3.1, $\beta \to F(\beta)$ is differentiable with $(d/d\beta)F(\beta) = \frac{1}{2}\langle Pa^\beta, a^\beta\rangle$ in a neighborhood of $\hat\beta$.

We now describe differentiability properties of the mappings $\beta \to w(\beta)$ and $\gamma \to w(\gamma)$ for $\beta \in V(\hat\beta)$ and $\gamma \in V(\hat\gamma)$. For the simplicity of the presentation we assume that the infinite-dimensional inequality constraint is inactive, i.e., that the interior of K, denoted by int K, is nonempty and that $l(a^{\hat\beta}) \in \text{int } K$, respectively, $l(a^{\hat\gamma}) \in \text{int } K$. The general case will be treated in Remark 3.5 below. A Hilbert space-valued function $g(t)$, $t \in \mathbb{R}$, is called directionally differentiable at t_0 with directional derivative $\dot g$ (or

$(d/dt)\dot{g}(t_0))$ if

$$\lim_{t\to 0^+}\left|\frac{g(t_0+t)-g(t_0)}{t}-\dot{g}\right|=0.$$

THEOREM 3.3. *Let* (H2), (H5)-(H8) *hold at a solution* $(a^{\hat\beta}, u^{\hat\beta})$ *of* $(\mathcal{P}^{\hat\beta})$, $\hat\beta>0$ *or* $(a^{\hat\gamma}, u^{\hat\gamma})$ *of* $(\mathcal{P}^{\hat\gamma})$, $\hat\gamma<\infty$, *respectively, and assume that* $l(a^{\hat\beta})\in \operatorname{int} K$, *respectively, that* $l(a^{\hat\gamma})\in \operatorname{int} K$ *with* $\operatorname{int} K\neq\emptyset$. *Then* $w(\beta)$ *and* $w(\gamma)$ *are directionally differentiable at* $\hat\beta$ *and* $\hat\gamma$, *respectively, and the following equations hold*:

$$0=B(\dot a^{\hat\beta}, \dot u^{\hat\beta})+E^*\dot\lambda^{\hat\beta}+\begin{pmatrix}Pa^{\hat\beta}+\hat\beta P\dot a^{\hat\beta}\\ 0\end{pmatrix},\qquad 0=E(\dot a^{\hat\beta}, \dot u^{\hat\beta}), \tag{3.3}$$

respectively,

$$\begin{aligned}&0=B(\dot a^{\hat\gamma}, \dot u^{\hat\gamma})+E^*\dot\lambda^{\hat\gamma}+\begin{pmatrix}\mu^{\hat\gamma}P\dot a^{\hat\gamma}+\dot\mu^{\hat\gamma}Pa^{\hat\gamma}\\ 0\end{pmatrix},\\ &0=E(\dot a^{\hat\gamma}, \dot u^{\hat\gamma}),\\ &0=+\tfrac{1}{2}-\langle Pa^{\hat\gamma}, \dot a^{\hat\gamma}\rangle,\quad \text{if } \mu^{\hat\gamma}>0.\end{aligned} \tag{3.4}$$

The proof of Theorem 3.3 can easily be derived from Theorem 3.4 of [IK2]. An analogous result holds for the weak limits of the backward difference quotients of $w(\beta)$ at $\hat\beta$ and $w(\gamma)$ at $\hat\gamma$. For the former, the term $Pa^{\hat\beta}$ in the first equation of (3.3) has to be replaced by $-Pa^{\hat\beta}$, and for the latter the term $\frac{1}{2}$ in the third equation of (3.4) is replaced by $-\frac{1}{2}$.

Remark 3.4. The conclusions of Theorem 3.3 remain correct in neighborhoods of $\hat\beta$ and $\hat\gamma$, respectively.

Remark 3.5. Here we provide the results on the differentiability of $w(\beta)$ and $w(\gamma)$ without the assumption that the infinite-dimensional inequality constraint is inactive. These results follows from Theorems 3.4 and 3.5 of [IK2]. Additional hypotheses and the concept of polyhedricity of a cone K with respect to $z\in Z$ are required.

The cone K is called polyhedric with respect to $z\in Z$ if

$$\overline{\bigcup_{\lambda>0}\lambda(K-Pz)\cap[z-Pz]^\perp}=\overline{\bigcup_{\lambda>0}\lambda(K-Pz)}\cap[z-Pz]^\perp,$$

where P denotes the projection onto K and $[z-Pz]^\perp$ stands for the orthogonal complement of the subspace spanned by $z-Pz$.

(H9) K is polyhedric at $l(a_0)+\eta_0$.

(H10) The operator

$$(h, v)\to\begin{pmatrix}E(h, v)\\ L(h)\end{pmatrix}$$

from $Q\times X$ to $Y\times Z$ is surjective.

(H11) The operator

$$(h, v, r)\to\begin{pmatrix}E(h, v)\\ \tfrac{1}{2}\langle Pa_0, h\rangle\\ L(h)+rL(\dot a_0)\end{pmatrix}$$

from $Q\times X\times\mathbb{R}$ to $Y\times\mathbb{R}\times Z$ is surjective.

Here L denotes the Fréchet derivative of the affine function l and the index 0 with a_0 and η_0 is used to denote (a^β, η^β) in case of the $(\mathcal{P}^\beta)$ problem and (a^γ, η^γ) for the $(\mathcal{P}^\gamma)$ problem.

Now let us assume that (H2) and (H5)–(H11) hold at a solution $(a^{\hat\beta}, u^{\hat\beta})$ of $(\mathscr{P}^{\hat\beta})$, $\hat\beta > 0$. Then $w(\beta)$ is directionally differentiable at $\hat\beta$ and

$$\begin{aligned} &0 = B(\dot a^{\hat\beta}, \dot u^{\hat\beta}) + E^*\dot\lambda^{\hat\beta} + \begin{pmatrix} P\dot a^{\hat\beta} + \hat\beta P\dot a^{\hat\beta} + L^*\dot\eta^{\hat\beta} \\ 0 \end{pmatrix}, \\ &0 = E(\dot a^{\hat\beta}, \dot u^{\hat\beta}), \\ &0 \in -L\dot a^{\hat\beta} + \partial\psi_{\hat K_+}(\dot\eta^{\hat\beta}), \end{aligned} \tag{3.5}$$

where $\hat K_+$ is the dual cone of

$$\overline{\bigcup_{\lambda>0} \lambda(K - l(a^{\hat\beta}, u^{\hat\beta}))} \cap [\eta^{\hat\beta}]^\perp$$

and $\partial\psi_{\hat K}(x)$ is the subdifferential of the indicator function ψ of $\hat K$ at x, i.e.,

$$\partial\psi_{\hat K}(x) = \begin{cases} \{y \in Z : \langle y, c - x\rangle_Z \leqq 0 \text{ for all } c \in \hat K\} & \text{if } x \in \hat K, \\ \phi & \text{if } x \notin \hat K. \end{cases}$$

Analogously, if (H2) and (H5)–(H11) hold at a solution $(a^{\hat\gamma}, u^{\hat\gamma})$ of $(\mathscr{P}^{\hat\gamma})$, $\hat\gamma < \infty$, then $w(\gamma)$ is directionally differentiable at $\hat\gamma$ and

$$\begin{aligned} &0 = B(\dot a^{\hat\gamma}, \dot u^{\hat\gamma}) + E^*\dot\lambda^{\hat\gamma} + \begin{pmatrix} \mu^{\hat\gamma} P\dot a^{\hat\gamma} + \dot\mu^{\hat\gamma} Pa^{\hat\gamma} + L^*\dot\eta^{\hat\gamma} \\ 0 \end{pmatrix}, \\ &0 = E(\dot a^{\hat\gamma}, \dot u^{\hat\gamma}), \\ &0 \in +\tfrac{1}{2} - \langle Pa^{\hat\gamma}, \dot a^{\hat\gamma}\rangle + \begin{cases} 0 & \text{if } \mu^{\hat\gamma} > 0 \\ \partial\psi_{\mathbb{R}_+}(\dot\mu^{\hat\gamma}) & \text{if } \mu^{\hat\gamma} = 0, \end{cases} \\ &0 \in -L\dot a^{\hat\gamma} + \partial\psi_{\hat K_+}(\dot\eta^{\hat\gamma}), \end{aligned} \tag{3.6}$$

where $\hat K_+$ is the dual cone of

$$\overline{\bigcup_{\lambda>0} \lambda(K - l(a^{\hat\gamma}, u^{\hat\gamma}))} \cap [\eta^{\hat\gamma}]^\perp.$$

4. Model function for $(\mathscr{P}^\beta)$. In this section we concentrate on the regularized problem $(\mathscr{P}^\beta)$. For the class of problems that we have in mind, solving the unregularized problem $(\mathscr{P}^0)$ would be numerically infeasible due to lack of continuous invertibility of f. The regularization term will guarantee that $(\mathscr{P}^\beta)$ can be solved in a numerically stable way if β is sufficiently large. The use of the regularization term, however, introduces error, the regularization error, since $(\mathscr{P}^\beta)$ is solved instead of $(\mathscr{P}^0)$. Increasing the regularization parameter implies an increase in the regularization error, while decreasing the regularization parameter to 0^+ increases the error due to lack of continuous invertibility of f. The problem of the optimal choice of the regularization parameter arises. It has received a considerable amount of attention for linear problems; see, e.g., [B], [G], [M], and [L]. The use of model functions for the $(\mathscr{P}^\beta)$ and $(\mathscr{P}^\gamma)$ problems was inspired by a similar technique developed by Hebden, Moré, and Reinsch for quasi-Newton methods (see [DS, p. 136]).

In this research we propose a four-parameter family of functions $m(\beta)$, which describes the minimal value function $F(\beta)$. The family of model functions $m(\beta)$ will subsequently be used for two purposes:

(i) To estimate the value $F(0)$ of the (unstable) unregularized problem $(\mathscr{P}^0)$, from evaluations of the stable regularized problem $(\mathscr{P}^\beta)$, $\beta > 0$.

(ii) To determine a "best" parameter value β^* to solve the regularized problem $(\mathscr{P}^{\beta^*})$ for $(a^{\beta^*}, u^{\beta^*})$. Here additional techniques, such as the Morozov principle, will be used.

Throughout this section it will be convenient to think of u as a dependent variable defined through $e(a, u) = 0$. The problem under investigation is thus

$$(\mathcal{P}^{\beta}) \qquad \min f(a) + \frac{\beta}{2}\langle Pa, a\rangle \quad \text{subject to } l(a) \in K.$$

We give some preliminary results. Let (H1)-(H4) hold so that by Theorem 2.1 there exist solutions a^{β} of $(\mathcal{P}^{\beta})$, $\beta > 0$. Then $\beta \to F(\beta)$ is monotonically increasing, $\beta \to f(a^{\beta})$ is monotonically increasing in a multivalued sense, and $\beta \to \langle Pa^{\beta}, a^{\beta}\rangle$ is monotonically decreasing in a multivalued sense. For the precise statement we define $A_{\beta} = \{a^{\beta}\colon a^{\beta}$ is a solution of $(\mathcal{P}^{\beta})\}$, for $\beta > 0$.

PROPOSITION 4.1. *Let* (H1)-(H4) *hold and let* $\beta > \beta_0 > 0_0$. *Then we have*

(i) $0 \leqq F(\beta_0) \leqq F(\beta)$,

(ii) $\sup_{A_{\beta 0}} f(a^{\beta 0}) \leqq \inf_{A_{\beta}} f(a^{\beta})$,

(iii) $\sup_{A_{\beta}} \langle Pa^{\beta}, a^{\beta}\rangle \leqq \inf_{A_{\beta 0}} \langle Pa^{\beta 0}, a^{\beta 0}\rangle$.

Assumptions (H1)-(H4) guarantee the existence of solutions to $(\mathcal{P}^{\beta})$. The proof of the proposition is quite analogous to that of Lemma 3.2 in [CK].

PROPOSITION 4.2. *Assume that* int $K \neq \emptyset$ *and that f is twice continuously Fréchet differentiable on* $\{a\colon l(a) \in K\}$. *For* $\hat{\beta} > 0$ *let* $a^{\hat{\beta}}$ *be a solution of* $(\mathcal{P}^{\hat{\beta}})$ *such that* $l(a^{\hat{\beta}}) \in$ int K *and*

$$f''(a^{\hat{\beta}})(h, h) + \hat{\beta}\langle Ph, h\rangle \geqq \kappa |h|_Q^2 \tag{4.1}$$

for some $\kappa > 0$ *independent of* $h \in Q$. *Then there exists a neighborhood* $V(\hat{\beta})$ *of* $\hat{\beta}$ *such that for* $\beta \in V(\hat{\beta})$

(i) $\beta \to a^{\beta}$ *is continuously differentiable,*

(ii) $F'(\beta) = \frac{1}{2}\langle Pa^{\beta}, a^{\beta}\rangle$, $F''(\beta) = \langle Pa^{\beta}, \dot{a}^{\beta}\rangle$,

(iii) $F''(\beta) = -\langle (B + \beta P)\dot{a}^{\beta}, \dot{a}^{\beta}\rangle \leqq -(\kappa/2)|\dot{a}^{\beta}|_Q^2$,

where $B = B(\beta)$ *is the operator representation of the bilinear form* $f''(a^{\beta})$.

Under the assumption of Proposition 4.2, the extremal value function associated with $(\mathcal{P}^{\beta})$ is monotonically nondecreasing and convex downward in the neighborhood of β-values where the second-order sufficient optimality condition (4.1) is satisfied.

Proof of Proposition 4.2. Theorem 3.3 is applicable, since (H7) is implied by (4.1) and the remaining hypotheses are clearly satisfied. From (3.3) and Remark 3.4 we find for the directional derivative from the right,

$$0 = Pa^{\beta} + (B(\beta) + \beta P)\dot{a}^{\beta}, \tag{4.2}$$

for β in a neighborhood of $\hat{\beta}$. The directional derivative from the left satisfies (4.2) with Pa^{β} replaced by $-Pa^{\beta}$ and hence $\beta \to a^{\beta}$ is differentiable for β in a neighborhood of $\hat{\beta}$. The smoothness property of f together with (4.1) and (4.2) imply that $B + \beta P$ has a bounded inverse for β in a neighborhood of $\hat{\beta}$ and that $\beta \to \dot{a}^{\beta}$ is continuous. Thus (i) is verified. Theorem 3.1 implies (ii), and (iii) is a consequence of (4.2) and (ii). This ends the proof.

The starting point for the derivation of the model function for $F(\beta)$ is (4.2). Assuming that the hypotheses of Proposition 4.2 hold at $\hat{\beta} > 0$, taking the inner product of (4.2) with a solution a^{β} of $(\mathcal{P}^{\beta})$ with $\beta \in V(\hat{\beta})$ leads to

$$0 = \langle Pa^{\beta}, a^{\beta}\rangle + \langle (B + \beta P)\dot{a}^{\beta}, a^{\beta}\rangle,$$

which further implies

$$0 = 2F' + \beta F'' + \langle B\dot{a}^{\beta}, a^{\beta}\rangle. \tag{4.3}$$

The following step is the heuristic one in the derivation of our model function. If P were invertible so that $\langle B\dot{a}^\beta, a^\beta\rangle = \langle BP^{-1}P\dot{a}^\beta, a^\beta\rangle$, and BP^{-1}, which depends on a^β, had a Riesz representation of the type $c_1\beta + c_2$, with $c_i \in \mathbb{R}$, then (4.3) would become

$$0 = 2F' + \beta F'' + (c_1\beta + c_2)F'' \quad \text{for } \beta \in V(\hat{\beta}). \tag{4.4}$$

Solving the ordinary differential equation (4.4) we obtain

$$m(\beta) = e - \frac{b}{(\beta + c)^d} \tag{4.5}$$

for $F(\beta)$, where $d = 2/(1 + c_1) - 1$, $c = c_2/(c_1 + 1)$, and b and e are integration constants.

This is the desired class of functions. For future reference we specify the first and second derivatives of m:

$$m'(\beta) = \frac{bd}{(\beta + c)^{d+1}}, \quad m''(\beta) = \frac{-bd(d+1)}{(\beta + c)^{d+2}} \quad \text{for } d \neq -1, \quad d \neq -2. \tag{4.6}$$

While $m(\beta)$ was derived as a local model function, we hope that it is a good approximation to F for a large range of β values. We proceed with some comments and observations.

(i) Solving $(\mathcal{P}^\beta)$ once for some $\beta > 0$, we obtain values for $F(\beta)$ as well as for $F'(\beta) = \frac{1}{2}\langle Pa^\beta, a^\beta\rangle$, which correspond to values for $m(\beta)$ and $m'(\beta)$. Consequently, solving $(\mathcal{P}^\beta)$ for two values of β gives four conditions that can be used to determine (e, b, c, d) for the model function m.

(ii) The model function will be used in an iterative process starting with a "large" value for β and decreasing β until a stopping criterion has been reached. For the problems that we have in mind, solving $(\mathcal{P}^\beta)$ for a large value of β is stable and computationally easy. Moreover, for a large value of β, the value of $F(\beta)$ should be a good approximation to the parameter e in m.

(iii) The model function will be used in an iterative scheme to determine the "best" regularization parameter. At each stage of the iteration procedure we use the model function together with a criterion which will be specified below to determine a new value for β. This new β-value, together with a combination of old β values, is used to update the model function. In general, we expect that only a few iterations are required.

(iv) In our numerical experiments with specific problems the evaluation of the global model function (one of the β-values for determining m was large) at 0 gave a good approximation to $F(0)$.

Let us next consider the range in which the parameters (b, c, d, e) determining m should vary. At least two aspects have to be kept in mind: first, the absolute value of β from which, by evaluation of $F(\beta)$ and $F'(\beta)$, the parameters determining m are calculated; second, the relative distance between these β values. If the β values are "large" with "large" relative distance, we expect that the model function possesses the properties of F described in Propositions 4.1 and 4.2 and that the parameters satisfy

$$b > 0, \quad c > 0, \quad d > 0, \quad e \geqq b/c^d. \tag{4.7}$$

These conditions are based on the assumption that $c_1 < 1$ (which is quite reasonable, e.g., for linear inverse problems and large values of β). The assumption $c_1 < 1$ implies that $d > 0$. We exclude the case in which m has a pole for $\beta \geqq 0$ and require that $c > 0$. From Propositions 4.1 and 4.2 it is known that $F'(\beta) \geqq 0$ for all $\beta \geqq 0$. Hence we require $m' \geqq 0$. We also assume that m is not a constant. In view of (4.6) this leads to $b > 0$. The condition $e > 0$ is obvious from the expected asymptotic behavior of $m(\beta)$ as $\beta \to \infty$, and the fact that the range of m should be in $\mathbb{R}^+$. The requirement $m(0) \geqq 0$ leads to $e \geqq bc^{-d}$. As β decreases with the relative distance between the β-value still sufficiently large, the parameters in m are expected to satisfy

$$bd > 0, \qquad d \geqq -1. \tag{4.8}$$

Here the requirement $c_1 < 1$ is dropped and we use Proposition 4.2 to argue that m' and m'' should be positive, respectively, negative, for sufficiently large β. This leads to $bd > 0$ and $d + 1 > 0$. As the β values decrease further and become relatively close, we expect that m is still monotonically increasing, i.e.,

$$bd > 0. \tag{4.9}$$

We next describe two possibilities for choosing an optimal regularization parameter. This is done in the context of inverse problems that were already discussed in § 1. Let z and z^δ denote the error-free and error-corrupted observation, with $|z - z^\delta| = \delta$, and let $a \to u(a)$ denote the parameter-to-output mapping. Note that we have not distinguished between z and z^δ, as in the introduction. The problem consists of determining a best parameter to fit the data. The classical regularized least squares formulation for this problem is

$$\min \frac{1}{2} |u(a) - z^\delta|_X^2 + \frac{\beta}{2} \langle Pa, a\rangle \quad \text{over } a \in Q_{ad}, \tag{4.10}$$

where Q_{ad} is the set of admissible parameters. In this case we have

$$f(a) = \tfrac{1}{2} |u(a) - z^\delta|_X^2.$$

According to the Morozov principle, the regularization parameter is chosen such that the (output) error due to regularization equals the error level in the data, i.e., β_m is chosen such that

$$|u(a^{\beta M}) - z^\delta|_X^2 = \delta^2. \tag{4.11}$$

This principle has received a considerable amount of attention for linear inverse problems [M] and was also recently studied in the context of nonlinear inverse problems in [EKN] and [N], where convergence and rate of convergence of a^{β_M} in terms of the error level δ are studied. Expressing (4.11) in terms of the model function leads to

$$m(\beta_M) - \beta_M m'(\beta_M) = \tfrac{1}{2}\delta^2. \tag{4.12}$$

In our experience with the use of Morozov's principle for parameter estimation problems, we found that it worked well, but that it gave a too conservative estimate for the regularization parameter when compared to the best possible choice. One possible explanation is that it takes into account only the image space of u. We have therefore also carried out experiments with another principle by choosing the regularization parameter according to

$$|u(a^{\beta_P}) - z^\delta|_X^2 + \beta_P^{\gamma_1} \langle Pa^{\beta_P}, a^{\beta_P}\rangle = \gamma_2 \delta^2, \tag{4.13}$$

for some $\gamma_1 \in [1, 2]$ and $\gamma_2 \in [\frac{1}{2}, 1]$. In terms of the model function this can be expressed as

$$m(\beta_P) + (\beta_P^{\gamma_1} + \beta_P) m'(\beta_P) = \frac{\gamma_2}{2} \delta^2 \quad \text{for some } \gamma_1 \in [1, 2], \quad \gamma_2 \in [\tfrac{1}{2}, 1]. \tag{4.14}$$

Theoretical aspects of this principle are discussed in [Ku2]. Let us point out that for both the Morozov principle and the principle in (4.13), information about the error level δ is required. For inverse problems where the unperturbed observation z is attainable so that there exists $a^* \in Q_{ad}$ with $u(a^*) = z$, we have

$$F(0) = \min \tfrac{1}{2}|u(a) - z^\delta|_X^2 \leqq \tfrac{1}{2}|u(a^*) - z^\delta|_X^2 = \tfrac{1}{2}\delta^2.$$

It follows that

$$F(0) \leqq \tfrac{1}{2}\delta^2, \tag{4.15}$$

and $F(0)$ can be used to obtain a lower bound on the error level. In some of our calculations, assuming that δ was unknown, we successfully replaced δ^2 in (4.12) or (4.14) by $2m(0)$. This was motivated by (4.15) and the success was possibly due to the fact that for practical computations due to approximation and numerical error, $m(0)$ (which is calculated from evaluations of $(\mathscr{P}^\beta)$ for $\beta > 0$) actually overestimates $F(0)$.

We now present a pseudoalgorithm for the solution of (4.10) with an optimal regularization parameter that is based on iterative use of model functions and the Morozov principle or, alternatively, the principle described in (4.13). Let us define the functions

$$M(\beta) = m(\beta) - \beta m'(\beta) - \tfrac{1}{2}\delta^2 \tag{4.16}$$

and

$$P(\beta) = m(\beta) + (\beta^{\gamma_1} - \beta)m'(\beta) - \frac{\gamma_2}{2}\delta^2. \tag{4.17}$$

PSEUDOALGORITHM.

```
INPUT    $\beta_0 \in \mathbb{R}$   startup value for regularization parameter
         $u$        nonlinear function
         $z^\delta$       observation
         $\delta$        error level (if available)
         $\varepsilon$        parameter that distinguishes a "local" from a
                  "global" model

OUTPUT   $\beta^*$    best regularization parameter
         $a^{\beta^*}$   regularized solution

FEST                  estimate for $F(0)$
$(e_1, b_1, c_1, d_1)$:     first model function
$(e^*, b^*, c^*, d^*)$:     converged model function
$k$:                    number of iterations

1. Initialization of model function.
1.1. Solve $(\mathscr{P}^{\beta 0})$ to obtain $F(a^{\beta 0})$, $F'(a^{\beta 0})$,
1.2. $e_0 := F(a^{\beta 0})$,
     $b_0 := 2e_0$.
1.3. Calculate $c_0$, $d_0$ from
     $m'(\beta_0) = F'(\beta_0)$,
     $m(0) = 0$.
1.4. Calculate the tangent to $m$ at $\beta_0$ and intersect the tangent with the $m(\beta)$ axis to
     obtain $(0, \sigma)$ in the $(\beta, m(\beta))$ plane. If $\sigma < 0$ give error message and stop.
1.5. Calculate $\beta_1$ from $m(\beta_1) = (\sigma/2)$; if no solution exists, give error message.
1.6. Solve $(\mathscr{P}^{\beta 1})$ to obtain $F(a^{\beta 1})$, $F'(a^{\beta 1})$.
```

```
1.7. Calculate $(e_1, b_1, c_1, d_1)$ from
     $m'(\beta_0) = F'(\beta_0)$,
     $m'(\beta_1) = F'(\beta_1)$,
     $m(\beta_0) = F(\beta_0)$,
     $m(\beta_1) = F(\beta_1)$.
1.8. FEST$:= e_1 - (b_1/c_1^{d_1})$ (estimate for $F(0)$).
1.9. FLAG$:=$ GLO,
     $k := 1$.
2. Iteration
2.1. $k = k+1$.
2.2. Calculate $\beta_+$ as root of $M(\beta_+) = 0$ or $P(\beta_+) = 0$. If this equation has no solution
     and $k = 2$ then choose $\beta_+$ such that $M(\beta_+) = .3084c_1$ (respectively, $P(\beta_+) =
     .3084c_1$), otherwise put $\beta_+ = \beta_1/2$. In case $\delta$ is not known replace it by FEST in
     $M$, respectively, $P$.
2.3. If FLAG = LOC and
         $\beta_+ \in [\min(\beta_0, \beta_1) - \frac{1}{2}|\beta_0 - \beta_1|, \max(\beta_0, \beta_1) + \frac{1}{4}|\beta_0 - \beta_1|]$,
     then create OUTPUT and STOP (convergence achieved).
2.4. If $|\beta_+ - \beta_1| < \varepsilon\beta_1$ put FLAG = LOC; otherwise put FLAG = LOC.
2.5. $\beta_{-1} := \beta_0, \beta_0 := \beta_1 := \beta_+$.
2.6. If FLAG = GLO or (FLAG = LOC and $|\beta_{-1} - \beta_1| \geqq |\beta_0 - \beta_1|$) then calculate
     $(e_k, b_k, c_k, d_k)$ from the equations in step 1.7. Otherwise exchange the roles of $\beta_0$
     and $\beta_{-1}$ and calculate $(e_k, b_k, c_k, d_k)$ from the equations in step 1.7.
2.7. GOTO 2.1.
```

Comments on the pseudoalgorithm.

ad 1.2. We suggest choosing β_0 large. The resulting problem $(\mathcal{P}^{\beta_0})$ will be stable and the second summand in the model function is small, which suggests choosing $e_0 = F(a^{\beta_0})$. The choice $b_0 := 2e_0$ is heuristic.

ad 1.3. The requirement $m(0) = 0$ means that for the zeroth-order model function we ignore possible error in the data. If the error level is known with high confidence, we might want to use it. At the end of step 1.3 the zeroth-order model function, characterized by (e_0, b_0, c_0, d_0), is available. It is used in steps 1.4 and 1.5 to calculate a problem-dependent second choice for the regularization parameter. Alternatively, we could have simply chosen $\beta_1 = \beta_0/2$ and skipped steps 1.2–1.5.

ad 1.7. There are several possibilities for approximately solving the equations in 1.7. We eliminated, analytically, the variables e, b, and d and obtained a single equation for c, which was solved numerically.

ad 1.8. As explained above, FEST is an estimate for $F(0)$ and a lower bound for the error in the data. In the numerical experiments where δ was assumed to be unavailable, δ was replaced by FEST.

ad 2.2. For the origin of the factor .3084, we refer to the discussion further below.

ad 2.4. The condition $|\beta_+ - \beta_1| < \varepsilon\beta_1$ is the criterion that determines whether the next model function is considered to be a global or a local approximation to F. In our calculations we took $\varepsilon = .1$. The stopping criterion is fulfilled if the next β value, β_+, calculated in step 2.6 satisfies the criterion of step 2.3 while FLAG = LOC.

Illustrative examples of numerical results obtained with an implementation of the above algorithm will be presented. Prior to that, let us explain an alternative heuristic approach to quickly obtaining an acceptable regularization parameter, which requires only the initialization stage of our pseudoalgorithm. Let us note that in view of

$F'(\beta)=\frac{1}{2}\langle Pa^\beta, a^\beta\rangle$,

$$\tilde{J}(\beta)=2(m(\beta)-\beta m'(\beta))$$

is an approximation to the function

$$J(\beta)=|u(a^\beta)-z^\delta|_X^2,$$

where a^β is a solution to $(\mathcal{P}^\beta)$. We find that

$$\tilde{J}'(\beta)=\frac{2\beta bd(d+1)}{(c+\beta)^{d+2}},\quad \tilde{J}''=\frac{2bd(d+1)}{(c+\beta)^{d+2}}\left(1-\beta\frac{d+2}{c+\beta}\right),$$

and if (e, b, c, d) satisfy the constraints of (4.7) or (4.8) with $d>-1$, then

$$\tilde{J}'(\beta)\geqq 0$$

and $\tilde{J}''$ has precisely one zero β_s given by

$$\beta_s=\frac{c}{d+1},$$

which characterizes the maximum of $\tilde{J}'$. Let β_I be defined by

$$\tilde{J}(\beta_I)=\tilde{J}(0)+\tfrac{1}{2}(\tilde{J}(\beta_s)-\tilde{J}(0)). \tag{4.18}$$

The idea behind this choice of β_I is that we allow for half of the total decrease of $\tilde{J}$ between β_s and 0. Solving (4.18) approximately by setting $d=1$, we obtain

$$\beta_I\sim\frac{c}{34}[2+\sqrt{72}]. \tag{4.19}$$

In many of our numerical experiments d_1 was close to 1 and these observations suggested the choice of β_+ in step 2.2 of the pseudoalgorithm. With our numerical results we will also specify $\beta_I\sim .3084c_1$, with c_1 calculated from the initialization phase. It will be seen to be a conservative estimate for the optimal regularization parameter, which is simple to obtain. We should also mention that we did not enforce the constraints in (4.7), (4.8), or (4.9), but rather we were interested in observing whether the model functions automatically behaved in the way expressed by these inequalities.

We next illustrate the applicability of the results of §§ 2 and 3 and the results of the above algorithm by means of a specific example. Consider the two-point boundary value problem

$$\begin{aligned}&-(au_x)_x+u=g \quad \text{on } (0,1),\\ &u_x(0)=u_x(1)=0,\end{aligned} \tag{4.20}$$

with $g\in L^2(0,1)$. The problem consists of determining the coefficient a from knowledge of u by solving the nonlinear least squares problem

$$\min_{a\in Q_{ad}} \tfrac{1}{2}|u(a)-z^\delta|_{L^2}^2+\frac{\beta}{2}|a_x|_{L^2}^2, \tag{4.21}$$

where $z^\delta\in L^2(0,1)$ denotes the perturbed observation and $Q_{ad}=\{a\in H^1(0,1)$: $a\geqq\alpha>0\}$. The error-free observation is assumed to be $z=(u, a^*)$, with a^* the "true" coefficient. In terms of the general framework of §§ 2 and 3 we have $Q=Z=H^1(0,1)$, $K=\{a\in H^1(0,1)$: $a(x)\geqq 0\}$, $\langle Pa, b\rangle=\langle a_x, b_z\rangle$, $l(a)=\alpha-a$, and $f\colon Q\to\mathbb{R}^+$ given by

$$f(a)=|u(a)-z^\delta|_{L^2}^2.$$

Here we treated the boundary value problem as an implicit constraint and hence X, Y, and e are not needed. This is due to the choice we made for the numerical solution of (4.21), for which we used a Levenberg-Marquardt algorithm, recalculating the value for $u(a)$ whenever an update for a was made. We refer to [IK2], where conditions similar to those of (H1)-(H8) were checked for a boundary value problem with Dirichlet boundary conditions.

It is easy to check that $\ker P = \{a \in H^1(0,1)\colon a = \text{constant}\}$ and that (H1)-(H3) hold. To discuss (H4) we use Proposition 2.2 and assume that $\{a_n\}$ is a sequence that satisfies the hypotheses specified there. It follows that $\{a_n\}$ is a sequence of positive functions with $\lim_n \inf_x a_n(x) = \infty$. This implies that $u_x(a_n) \to 0$ in $L^2(0,1)$. Moreover, $\{u(a_n)\}$ is bounded in $H^1(0,1)$. This implies that a subsequence of $\{u(a_n)\}$ converges weakly in $H^1(0,1)$ to a constant. Since integration of

$$-(a_n u_x(a^n))_x + u(a_n) = g$$

gives $\int_0^1 u(a_n)\,dx = \int_0^1 g\,dx$, we obtain weak convergence in $H^1(0,1)$ of the sequence $\{u(a_n)\}$ itself to the constant function $\int_0^1 \text{g dx}$. Hence, if

$$\left|\int g\,dx - z\right|_{L^2}^2 > \inf\{|u(a)-z|_{L^2}^2\colon a \geqq \alpha,\ a = \text{constant}\}, \tag{4.22}$$

then Proposition 2.2 implies (H4). The conditions (H5) and (H6) are simple to check. As for (H7), we refer to [CK], where it is shown that this second-order sufficient optimality condition holds, provided that $|z^\delta - z|_{L^2}$ is sufficiently small and that β is chosen appropriately. Conditions (H8), (H9), and (H11) are easy to check, and (H10) holds, since $H^1(0,1)$ is polyhedric at any of its elements [H]. Thus the results of §§ 2 and 3 are applicable to the parameter estimation problem (4.21).

Next we present a specific numerical example.

Example 1. We choose

$$a^*(x) = 2 + \sin \pi x \quad \text{and} \quad z(x) = u(a^*)(x) = -4x^3 + 6x^2.$$

The inhomogeneity is given by $g(x) = 1 - x$. In this case the singular set S of the observation z is found to be $S = \{0, 1\}$ and there exists a unique coefficient $a \in H^1$ satisfying $u(a) = z$ given by a^* [Ku1]. It is also the unique solution of the unregularized problem with error-free observation z:

$$\min_{a \in Q_{ad}} \tfrac{1}{2}|u(a) - z|_{L^2}^2.$$

The problem (4.21) is infinite dimensional. It was discretized by approximating the unknown coefficient a by linear splines on a uniform grid with mesh $\{i/(N+1)\}_{i=0}^{N+1}$ and by approximating (4.20) by a standard Galerkin procedure, also employing linear elements on the same grid as used for the coefficients. Thus we have finite-dimensional coefficients a^N and a finite-dimensional state $u^N = u^N(a^N)$. The finite-dimensional problem that we considered was

$$\begin{aligned}
&\min_{a^N} \frac{1}{2}\frac{1}{N+1}\left(\sum_{i=1}^{N} \left| u^N(a^N)\left(\frac{i}{N+1}\right) - z^\delta\left(\frac{i}{N+1}\right)\right|^2 + \frac{1}{2}|u^N(a^N(0)) - z^\delta(0)|^2\right.\\
&\qquad \left. + \frac{1}{2}|u^N(a^N(1)) - z^\delta(1)|^2\right)\\
&\qquad + \frac{\beta}{2}(N+1)\sum_{i=0}^{N}\left|a^N\left(\frac{i+1}{N+1}\right) - a^N\left(\frac{i}{N}\right)\right|^2,
\end{aligned} \tag{4.23}$$

where z^δ is calculated from z by fitting a cubic interpolation polynomial z^δ through

$$z(y_i) + \text{rand}\,(y_j) \cdot \tilde{\delta}, \tag{4.24}$$

with $\{y_i\}_{j=1}^n$ equidistant grid points in $[0,1]$, rand (y_j) a uniformly distributed random number in $[-1,1]$, and $\tilde{\delta} \in \mathbb{R}$. For the numerical results of this section we chose $n = 9$.

The noise level δ relevant to our subsequent computations is given by

$$\delta^2 = \frac{1}{M+1}\left(\sum_{i=1}^{N}\left|z\left(\frac{i}{N+1}\right) - z^{\delta}\left(\frac{i}{N+1}\right)\right|^2 + \frac{1}{2}|z(0) - z^{\delta}(0)|^2 + \frac{1}{2}|z(1) - z^{\delta}(1)|^2\right). \tag{4.25}$$

The initial choice for β was taken as $\beta_0 = .01$ and the parameter distinguishing the global from the local model, as $\varepsilon = .1$. In all calculations that we present here, $N = 16$. We did not realize the constraint $a(x) \geqq \alpha$, which is known from previous experiments to have no effect.

The first numerical question that we raised was whether there existed an optimal β-value. To obtain an answer, we solved (4.23) for 101 equidistant values of β in the interval $[0, 5 \times 10^{-5}]$ (and for several values of β larger than 10^{-3}). The results in terms of the L^2-distance $|a^{N,\beta} - a^*|_L^2$ for various values of $\hat{\delta}$ is shown in Fig. 1. Here $a^{N,\beta}$ denotes the solution of (4.23). We observe that, in the sense of L^2, there exists a unique optimal β-value that we hope to approximate reasonably well with our algorithm. It can also be noted that the optimal β-values increase as $\hat{\delta}$ (and δ) increase. With Fig. 1, as well as with the other figures, we ask the reader to observe the varying scales that are being used. In Fig. 2 we compare the optimal β-values (indicated by 0) of Fig. 1 to the β-values β_M obtained as the solution of the Morozov equation

$$\frac{1}{N+1}\left(\sum_{i=1}^{N}\left|u^N(a^{N,\beta_M})\left(\frac{i}{N+1}\right) - z^{\delta}\left(\frac{i}{N+1}\right)\right|^2 + \frac{1}{2}|u^N(a^{N,\beta_M})(0) - z^{\delta}(0)|^2 + \frac{1}{2}|u^N(a^{N,\beta_M})(1) - z^{\delta}(1)|^2\right) = \delta^2 \tag{4.26}$$

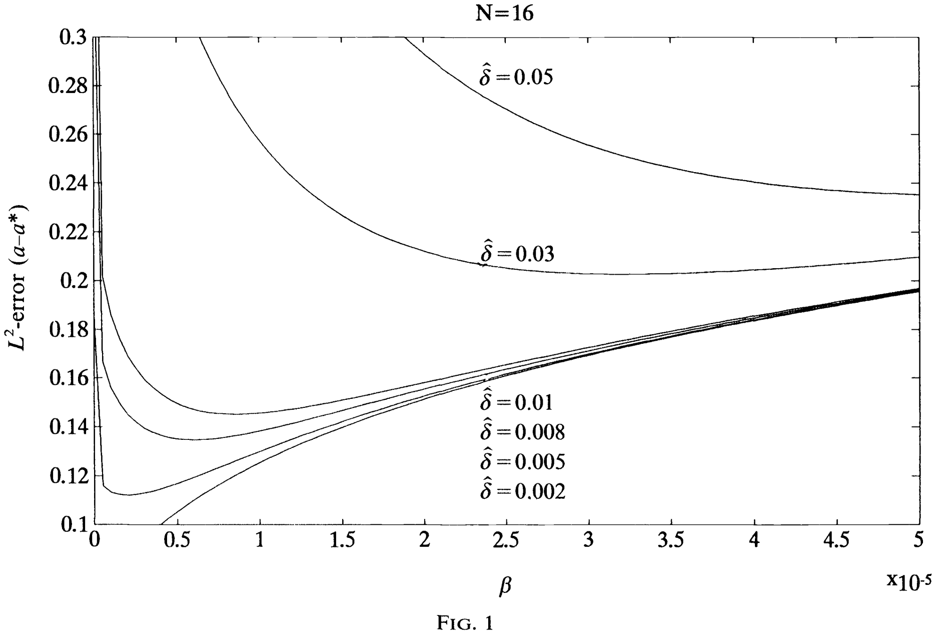


FIG. 1

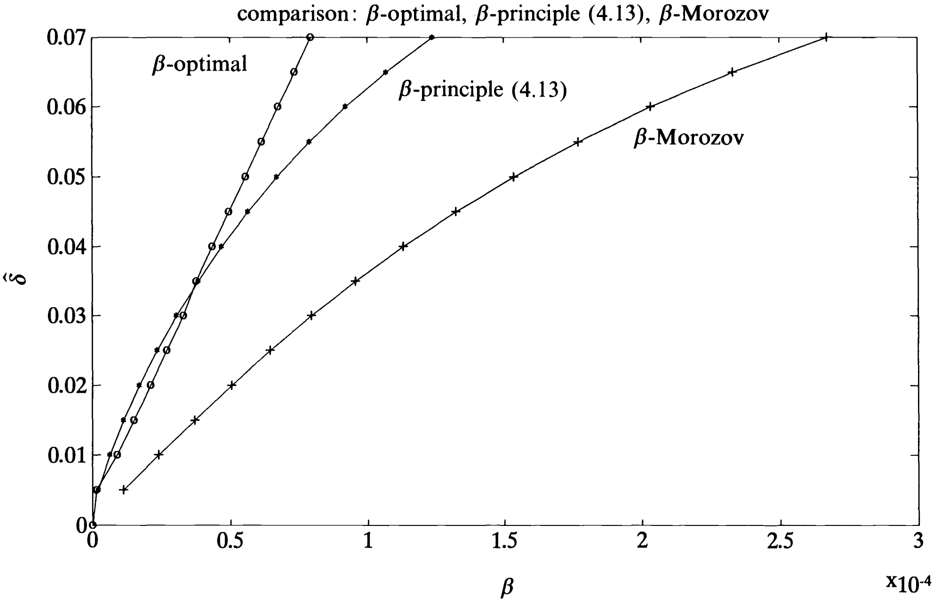


FIG. 2

(indicated by $*$). In order to obtain the exact value for β according to the Morozov principle we have not yet used our algorithm; we used instead the fact that

$$\frac{1}{2}\frac{1}{N+1}\left(\sum_{i=1}^{N}\left|u^N(a^{N,\beta})-z^\delta\left(\frac{i}{N+1}\right)\right|^2+\frac{1}{2}|u^N(a^{N,\beta_M})(0)-z^\delta(0)|^2\right.$$
$$\left.+\frac{1}{2}|u^N(a^{N,\beta_M})(1)-z^\delta(1)|^2\right)=F(\beta)-\beta F'(\beta);$$

the values for $F(\beta)$ and $F'(\beta)$ were available from the computations for Fig. 1 and thus $F(\beta)-\beta F'(\beta)=\frac{1}{2}\delta^2$ could be solved almost exactly. The purpose of this comparison was to study the behavior of the Morozov principle without the additional influence of the model function technique. We observe that the Morozov principle gives good, but somewhat conservative, estimates for the optimal β-value (in the sense of the L^2-error for a). A similar calculation was carried out with the principle described in (4.13) with $\gamma_1=\gamma_2=1$. The results are given by the +-line of Fig. 2. Larger values of γ_2 gave less favorable results. The results of Fig. 2 were obtained with the exact δ-value given by (4.25). We also carried out analogous calculations to those of Fig. 2 but with δ replaced by $\sqrt{2\times\text{FEST}}$, with FEST obtained from the initialization phase of our algorithm. The results for the Morozov principle and the principle according to (4.13) with $\gamma_1=1$, $\gamma_2=\frac{1}{2}$ are shown in Figs. 3 and 4. These results show the reliability of the initialization phase of our algorithm in using $\sqrt{2\times\text{FEST}}$ as an estimate for the error δ. The power of the initialization phase to estimate δ via FEST for the example under consideration is demonstrated independently in Fig. 5, in which we compare δ (dashed line) to $\text{FEST}=\sqrt{2m(0)}$ (solid line) for the values $\hat{\delta}=i\times.005$, $i=0,\cdots,14$.

Finally we show the iterates of the Morozov function M and of the function P defined in (4.16) and (4.17). For the Morozov function, δ was replaced by $\sqrt{2\times\text{FEST}}$,

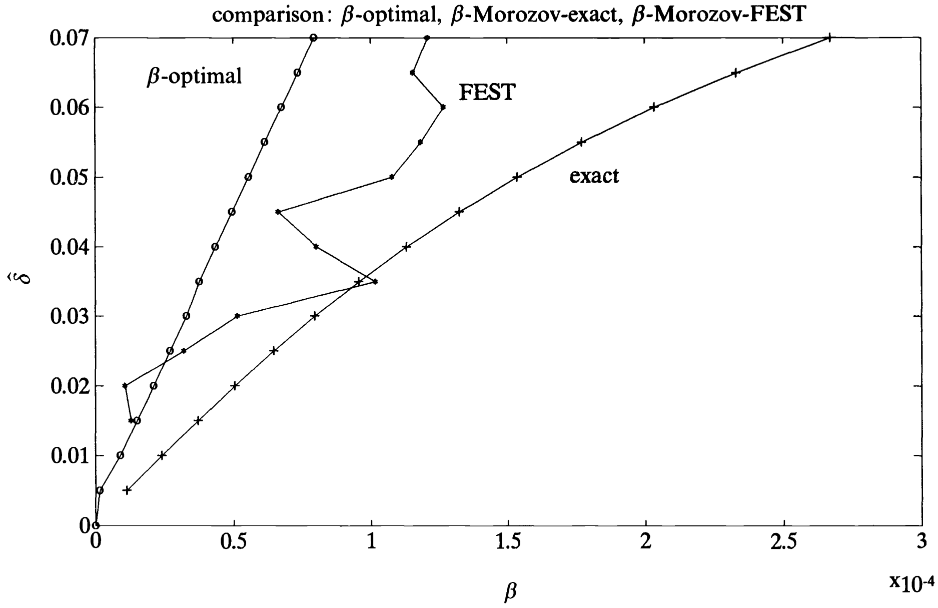


FIG. 3

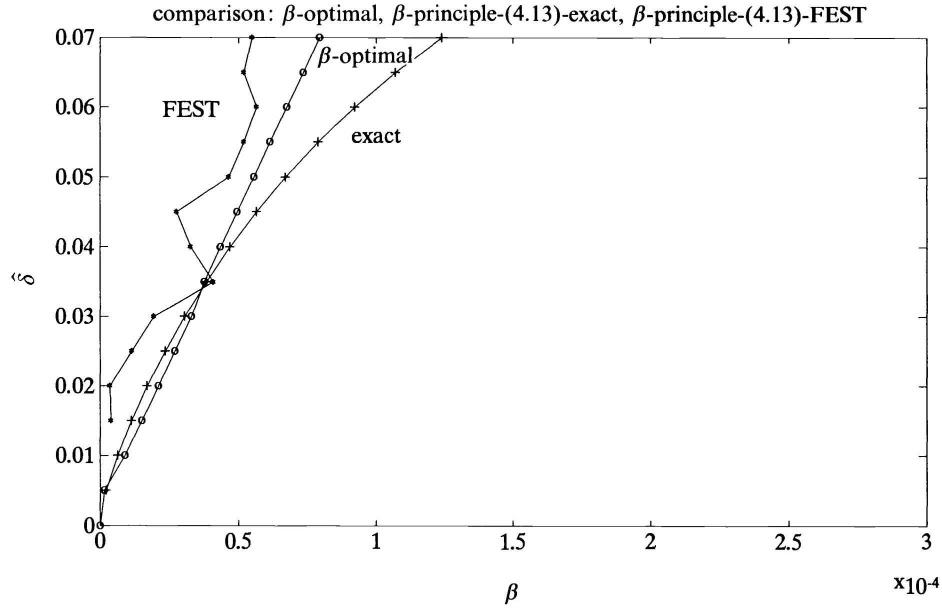


FIG. 4

which was obtained from the value (e_1, b_1, c_1, d_1) of the iterative stage of the algorithm. For these calculations, $\hat{\delta} = .03$. The dotted lines in Figs. 6 and 7 show the model function at various stages of the iteration. The solid line is obtained from (4.16) and (4.17) by replacing m and m' by F and F' (for which the values were calculated exactly with $N = 16$). Let us point out that the iterates approximate the exact Morozov function well, especially in the neighborhood of the ordinate value 0. This is also true for the

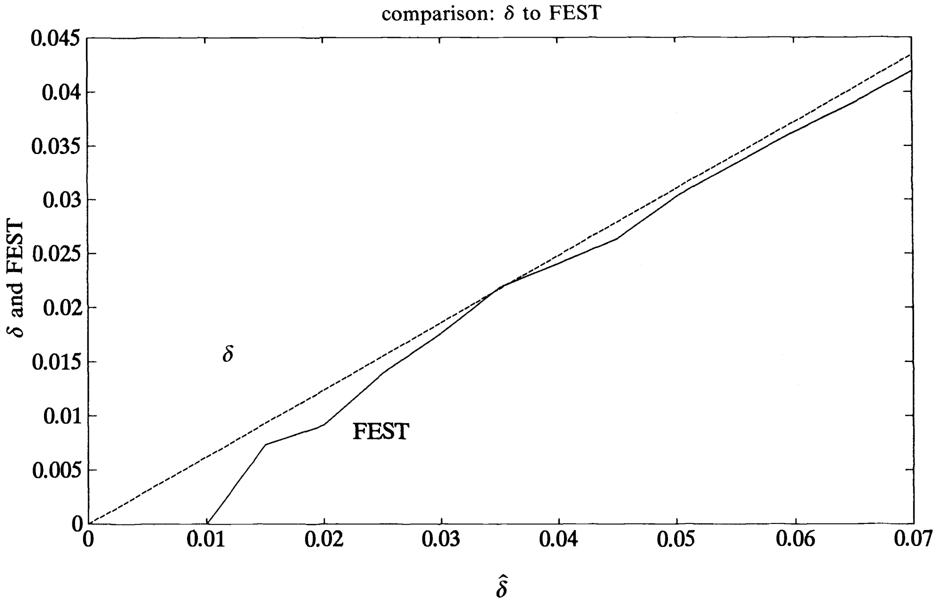


FIG. 5

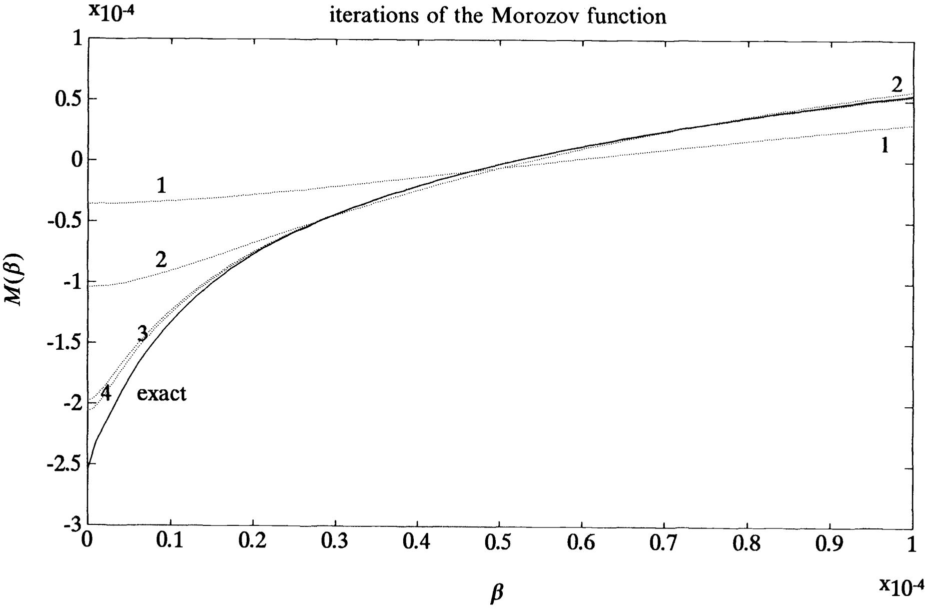


FIG. 6

iterates of Fig. 7 based on the principle of (4.13). The optimal β-value (in the sense of L^2) is $\beta_{\text{opt}} = 0.12 \times 10^{-4}$; the exact Morozov β-value is $\beta_M = 0.52 \times 10^{-4}$.

In Table 1 we give the values for the parameters (e, b, c, d) characterizing the model functions at consecutive iterations of the algorithm. In all cases the algorithm converged. Observe that FEST gives a good estimate for δ (Table 1(i), (ii)). In these examples the Morozov principle overestimates the best β-values, and principle (4.13)

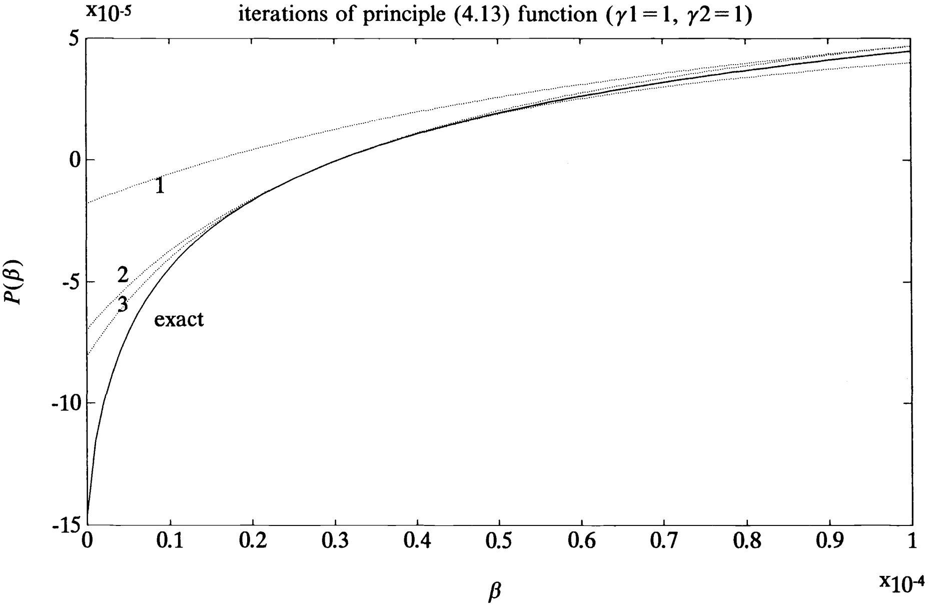


FIG. 7

with $\gamma_1=\gamma_2=1$ should be preferred. Comparing (ii) and (iii) of Table 1 we observe that due to the slight underestimate of FEST to δ, the converged value for β based on the Morozov principle is a little smaller in (ii) than in (iii). The exact value for the Morozov regularization parameter is $.796\times 10^{-4}$ if $\hat{\delta}=.03$ (Table 1(i)) and it is $.154\times 10^{-3}$ for $\hat{\delta}=.05$ (Table 1(ii)). The exact value for the regularization parameter according to principle (4.13), with $\gamma_1=\gamma_2=1$, is $.305\times 10^{-4}$ if $\hat{\delta}=.03$ (Table 1(iv)).

5. Model function for ($\mathcal{P}^\gamma$). In this section we consider the ($\mathcal{P}^\gamma$) problem. Although the equivalence between the β problem and γ problem under the second-order sufficient optimality condition (H7) is established in Theorem 2.5, they differ because in the γ problem the functional to be minimized is fixed while the constrained set is adjusted by changing $\gamma>0$. The parameter γ, which characterizes the constrained set $\{\langle a, Pa\rangle \leqq \gamma\}$, should be chosen so that a certain performance level is achieved (i.e., the minimal value function $F(\gamma)$ is smaller than an a priori chosen constant) and the stability of the minimization problem (i.e., the coercivity of the second Fréchet derivative of the Lagrangian) is maintained. In view of Theorem 3.1, the Lagrange multiplier $\mu(\gamma)$ provides a measure of how tightly the constraint is satisfied. In general, it is assumed that the constraint is active and $\mu(\gamma)>0$ and that condition (H7) holds only when $\mu_0>0$. Otherwise, the problem is well posed. In fact, for the case of the parameter estimation problem (1.5), the second derivative of the Lagrangian $\mathcal{L}$ is given by

$$\mathcal{L}''(a_0, u_0, \lambda_0, \mu_0, \eta_0)(h, v)^2 = f''(a_0)(h, v)^2 + \langle \lambda_0, e''(a_0, y_0)(h, v)^2\rangle$$
$$+\mu_0\langle Ph, h\rangle.$$

Because of lack of coercivity of $f''(a)$ (in some cases it may even be indefinite), condition (H7) requires that $\mu_0>0$ [CK], [IK1]. Our numerical calculations for (1.5) show that there exists a threshold value γ^* such that if $\gamma \geqq \gamma^*$, then the solution $a(\gamma)$ starts to oscillate. In the context of the above discussion, this can be understood as losing stability since under the assumptions of Theorem 3.3, $\mu(\gamma)$ is decreasing with respect to γ. This can be seen from (5.6).

TABLE 1
Example 1.

(i) Morozov in 2.2, $\hat{\delta} = .03$, δ replaced by FEST, corresponds to Fig. 6.

Iteration	e	b	c	d	β_+
1	$.282\times10^{-3}$	$.150\times10^{-7}$	$.931\times10^{-4}$	.974	$.287\times10^{-4}$
2	$.285\times10^{-3}$	$.102\times10^{-6}$	$.320\times10^{-4}$	.723	$.536\times10^{-4}$
3	$.492\times10^{-3}$	$.792\times10^{-4}$	$.310\times10^{-5}$	.135	$.515\times10^{-4}$
4	$.489\times10^{-3}$	$.773\times10^{-4}$	$.342\times10^{-5}$	.137	$.513\times10^{-4}$

The value for β_1 is given by $.287\times10^{-4}$, $\beta_{\text{opt}} = .12\times10^{-4}$, $\delta = .0186$, FEST = .0176.

(ii) Morozov in 2.2, $d = .05$, δ replaced by FEST.

Iteration	e	b	c	d	β_+
1	$.579\times10^{-3}$	$.195\times10^{-7}$	$.116\times10^{-3}$	.962	$.357\times10^{-4}$
2	$.582\times10^{-3}$	$.124\times10^{-6}$	$.157\times10^{-4}$	.713	$.107\times10^{-3}$
3	$.608\times10^{-3}$	$.140\times10^{-5}$	$.464\times10^{-5}$	.473	$.108\times10^{-3}$
4	$.604\times10^{-3}$	$.113\times10^{-5}$	$.471\times10^{-5}$	.492	$.108\times10^{-3}$

The value for β_I is given by $.358\times10^{-4}$, $\beta_{\text{opt}} = .55\times10^{-4}$, $\delta = .031$, FEST = .030.

(iii) Morozov in 2.2, $\hat{\delta} = .05$, δ exact.

Iteration	e	b	c	d	β_+
1	$.579\times10^{-3}$	$.195\times10^{-7}$	$.116\times10^{-3}$	.963	$.357\times10^{-4}$
2	$.582\times10^{-3}$	$.124\times10^{-6}$	$.157\times10^{-4}$	.713	$.149\times10^{-3}$
3	$.601\times10^{-3}$	$.906\times10^{-6}$	$.621\times10^{-5}$	.513	$.154\times10^{-3}$
4	$.595\times10^{-3}$	$.603\times10^{-6}$	$.676\times10^{-5}$	.552	$.154\times10^{-3}$

The value for β_I is given by $.358\times10^{-4}$

(iv) Principle (4.13) in 2.2, $\hat{\delta} = .03$, δ exact, corresponds to Fig. 7.

Iteration	e	b	c	d	β_+
1	$.282\times10^{-3}$	$.150\times10^{-7}$	$.931\times10^{-4}$	.974	$.287\times10^{-4}$
2	$.285\times10^{-3}$	$.102\times10^{-6}$	$.320\times10^{-4}$	.723	$.305\times10^{-4}$
3	$.240\times10^{-3}$	$.528\times10^{-9}$	$.333\times10^{-4}$	1.217	$.305\times10^{-4}$

The value for β_1 is given by $.287\times10^{-4}$.

In this study, based on the sensitivity analysis described in § 3, we construct a model function $m(\gamma)$ for the minimal value function $F(\gamma)$, which is characterized by four parameters. Using this model function we propose two procedures to determine the "best" value γ^*. Specifically, the family of model functions will be used for the following objectives:

(1) to accelerate the process of finding the value of γ satisfying, for instance, the Morozov principle (4.11);

(2) to estimate the value of $F(\gamma)$ at infinity, i.e., to evaluate the minimal value function without the seminorm constraint which allows us to estimate the noise in the data; and

(3) to obtain an estimate of the threshold value of $\mu^* = \mu(\gamma^*)$.

Assuming that the conditions of Theorem 3.3 hold at $\gamma > 0$, it follows from Theorem 3.1 and Remark 3.5 that

$$\frac{d}{d\gamma} F(\gamma) = -\tfrac{1}{2}\mu(\gamma), \tag{5.1}$$

$$B(\dot{a}, \dot{u}) + E^*\dot{\lambda} + \begin{bmatrix} \mu P\dot{a} + \dot{\mu} P a + L^*\dot{\eta} \\ 0 \end{bmatrix} = 0, \tag{5.2}$$

$$E(\dot{a}, \dot{u}) = 0, \tag{5.3}$$

$$\langle P\dot{a}, a\rangle = \tfrac{1}{2}, \tag{5.4}$$

$$L\dot{a} \in \hat{K} \quad \text{and} \quad \langle \dot{\eta}, z - L\dot{a}\rangle \leqq 0 \quad \text{for all } z \in \hat{K}, \tag{5.5}$$

where we used the notation of § 3.

Taking the inner product of (5.2) with $(\dot{a}, \dot{u})$, it follows from (5.3)-(5.5) that

$$\tfrac{1}{2}\dot{\mu} = -\langle P\dot{a}, \dot{a}\rangle\mu - \langle B(\dot{a}, \dot{u}), (\dot{a}, \dot{u})\rangle. \tag{5.6}$$

In order to construct model functions for μ and F we assume that the quadratic forms appearing on the right-hand side of (5.6) are constant. Then we obtain the differential equation for $(m, \tilde{\mu})$:

$$\frac{d}{d\gamma} m = -\tilde{\mu} \quad \text{and} \quad \frac{d\tilde{\mu}}{d\gamma} = -a(\tilde{\mu} + d) \tag{5.7}$$

where $m(\gamma)$ and $\tilde{\mu}$ represent approximations to $2F(\gamma)$ and $\mu(\gamma)$, respectively. This system of differential equations yields the model function m of the form

$$m(\gamma) = e^{-a\gamma + b} + c + d\gamma. \tag{5.8}$$

Although $(a, b, c, d) \in \mathbb{R}^4$ may vary with γ, we treat (a, b, c, d) as constants over the range of γ-values that we deal with.

Remark 5.1. (1) Since the constant a approximates $2\langle P\dot{a}, \dot{a}\rangle$, it should be positive.

(2) Note that $\tilde{\mu} = -(d/d\gamma)m = a\, e^{-a\gamma + b} - d$. Thus $\tilde{\mu}(\gamma)$ is decreasing, just like $\mu(\gamma)$, for values at which (H7) holds.

(3) Assume that $d > 0$. Then $\tilde{\mu}(\hat{\gamma}) = 0$ for $\hat{\gamma} = (b - \log(d/a))/a$. This may mean that for $\gamma > \hat{\gamma}$ the seminorm constraint is inactive. Consequently, $\hat{m} := m(\hat{\gamma})$ provides an estimate for $2F(+\infty)$.

(4) Assume that $d < 0$. Then $\tilde{\mu}(\gamma) \to -d$ as $\gamma \to \infty$. This may correspond to the case where the constraint is always active and $-d$ gives an estimate for the asymptotic value of the Lagrange multiplier $\mu(\gamma)$. Recall that $2ad$ describes the action of B. Since we assume that $a > 0$, the case $d > 0$ can be expected to hold when B is nonnegative and a discretization of the problem is made for practical computations. On the other hand, $d < 0$ describes the case where B is indefinite.

(5) The solution of (P^γ) for a given value of γ yields a pair $(F(\gamma), \mu(\gamma))$. Thus solving (P^γ) for two distinct values of γ gives four conditions that can be used to determine $(a, b, c, d) \in \mathbb{R}^4$ appearing in (5.8).

Based on the model function m of the form (5.7), we now propose two procedures to determine the "best" value of γ.

PROCEDURE 1. Assume that $d > 0$. From Remark 5.1(3) $\hat{m} = m(\hat{\gamma})$ with $\hat{\gamma} = (b - \log(d/a))/a$ can be used as an estimated noise level in the Morozov principle; i.e., γ^* will be determined as the root of $F(\gamma) = \frac{1}{2}\hat{m}$.

PROCEDURE 2. Assume that $d < 0$. From Remark 5.1(4) we may use $\hat{\mu} = -\lambda d$, $1 < \lambda < 2$ as a threshold value for $\mu(\gamma)$; i.e., γ^* will be determined as the root of $\mu(\gamma) = \hat{\mu}$. The larger the value of λ is, the more conservative (i.e., smaller) the solution of $\mu(\gamma) = \hat{\mu}$ will be. In our calculations we used $\lambda = 1.75$, which was determined empirically.

From Remark 5.1(5) two distinct values of γ are required in order to construct the model function. We proceed as follows.

ALGORITHM (construction of model function $m(\gamma)$).

Step 1. Assume that a good estimate of the lower bound of γ, say γ_1, is known. Then solving $(\mathscr{P}^{\gamma})$ with $\gamma=\gamma_1$, we obtain a pair $(F(\gamma_1), \mu(\gamma_1))$. Otherwise, set $\beta_1=1\times 10^{-2}$ and let $(a^{\beta_1}, u^{\beta_1})$ be a solution to $(\mathscr{P}^{\beta})$ with $\beta=\beta_1$. Then, set $\gamma_1=\langle Pa^{\beta_1}, a^{\beta_1}\rangle$, $\mu(\gamma_1)=\beta$, and $F(\gamma_1)=f(a^{\beta_1}, u^{\beta_1})$(by Theorem 2.5).

Step 2. Compute the Newton correction $p=2F(\gamma_1)/\mu(\gamma_1)$. If $p\leqq\gamma_1/2$, then let $\gamma_2=\gamma_1+\frac{1}{2}p$ and solve $(\mathscr{P}^{\gamma})$ with $\gamma=\gamma_2$. This yields a pair $(F(\gamma_2), \mu(\gamma_2))$. Otherwise, set $\beta_2=2.5\times\mu(\gamma_1)$ and solve $(\mathscr{P}^{\beta})$ with $\beta=\beta_2$ to obtain $\gamma_2=\langle Pa^{\beta_2}, a^{\beta_2}\rangle$, $\mu(\gamma_2)=\beta_2$, and $F(\gamma_2)=f(a^{\beta}, u^{\beta})$.

Step 3. Determine the parameters (a, b, c, d) in (5.8) by minimizing the functional

$$\sum_{i=1}^{2}\left(|2F(\gamma_i)-(e^{-a\gamma_i+b}+c+d\gamma_i)|^2+|\mu(\gamma_i)-(e^{-a\gamma_i+b}-d)|^2\right)$$

subject to $a>0$; i.e., $(a, b, c, d)\in\mathbb{R}^4$ is determined in the least squares sense.

We then combine the procedures to determine the best value of γ and the use of the model function $m(\gamma)$ in an iterative procedure as follows.

ITERATION

Step 4. From Step 3 we have the model function $m(\gamma)$.

Case 1. If $d>0$, then compute $\hat{\gamma}=(b-\log(d/a))/a$ and $\hat{m}=m(\hat{\gamma})$. Set $\gamma_3=\hat{\gamma}$ and $k=3$.

Step. 5. Solving $(\mathscr{P}^{\gamma})$ with $\gamma=\gamma_k$, we obtain a new pair $(F(\gamma_k), \mu(\gamma_k))$. Update the values of $(a, b, c, d)\in\mathbb{R}^4$ by minimizing the functional

$$\sum_{i=1}^{2}\left(|2F(\gamma_i)-(e^{+a\gamma_i+b}+c+d\gamma_i)|^2+|\mu(\gamma_i)-(e^{-a\gamma_i+b}-d)|^2\right) \tag{5.9}$$

subject to $a>0$, where the new conditions at γ_k are added to the least squares criterion.

Step 6. Calculate γ_{k+1} as the nearest root to γ_k of $m(\gamma)=\hat{m}$ where the updated model function obtained in Step 5 is used.

Step 7. If $|(\gamma_k-\gamma_{k+1})/\gamma_k|\leqq\varepsilon=1\times 10^{-3}$, set $\gamma^*=\gamma_k$ and stop. Otherwise set $k=k+1$ and return to Step 5.

Case 2. If $d<0$, then compute $\hat{\mu}=-1.75d$ and let γ_3 be the root of $\tilde{\mu}(\gamma)=\hat{\mu}$: i.e., $\gamma_3=(b-\log(-.75\times d)/a)/a$ and set $k=3$.

Step 5. This step is the same as Step 5 above.

Step 6. Calculate γ_{k+1} as the root of $\tilde{\mu}(\gamma)=\hat{\mu}$; i.e., $\gamma_{k+1}=(b-\log((\hat{\mu}+d)/a))/a$ where (a, b, c, d) are obtained through Step 5.

Step 7. This step is the same as Step 7 above.

Remark 5.2. The validity of the model function $m(\gamma)$ can be determined by checking the minimal value of the functional (5.9) for each k. The values of (a, b, c, d) obtained in Step 5 and the pairs $(F(\gamma_k), \mu(\gamma_k))$, $k\geqq 1$ along with a sequence of solutions $\{a(\gamma_k)\}_{k\geqq 1}$ can be used to analyze the behavior of the function $\gamma\to a(\gamma)$ and the validity of the procedure. In the case when the noise level δ^2 is known, the steps in Case 1 can be used to determine γ^* according to the Morozov principle.

We tested the proposed algorithm using the following parameter estimation problem, which consists of determining the positive coefficient $a(x)$ in the two-point boundary value problem:

$$-(au_x)_x=g \quad \text{in } (0.1), \qquad u(0)=u(1)=0,$$

knowing the measurement z of $u(a)$; i.e., finding an inverse of the solution map $a\in H^1(0,1)\to u(a)\in H_0^1(0,1)$. This is an ill-posed problem in the sense that a does not depend continuously on u [CK]. As in [IK1], the problem can be cast as a constrained minimization problem for $(a(x), u(x)) \in H^1(0,1) \times H_0^1(0,1)$:

$$\begin{aligned} &\text{minimize } \tfrac{1}{2}|u-z|_{H_0^1} \\ &\text{subject to } -(au_x)_x - g = 0, \quad \int_0^1 |a_x|^2\,dx \leqq \gamma, \quad a(x) \geqq \alpha > 0. \end{aligned} \tag{5.10}$$

It is of the form (1.6) and $(\mathscr{P}^\gamma)$, and the conditions (H1)-(H10) are satisfied (see [IK2] and the discussion in the previous section) and thus the results in §§ 2 and 3 can be applied to (5.10). In order to solve this minimization problem we employ the augmented Lagrangian method (see the detailed discussions in [IK1] and [IKK]). The problem is discretized using the standard finite element method, i.e., we represent

$$u^n(x) = \sum_{k=1}^{n-1} u_k B_k^{(n)}(x) \in H_0^1(0,1),$$

$$a^n(x) = \sum_{k=0}^{n} a_k B_k^{(n)}(x) \in H_0^1(0,1),$$

where $B_k^{(n)}(x)$ is a piecewise linear B-spline given by

$$B_k^{(n)}(x) = \begin{cases} n(x-x_{k-1}) & \text{on } [x_{k-1}, x_k], \\ n(x_{k+1}-x) & \text{on } [x_n, x_{k+1}], \\ 0 & \text{otherwise,} \end{cases}$$

with $x_k = k/n$, $0 \leqq k \leqq n$. That is, we solve the following minimization problem in (a^n, u^n):

$$\begin{aligned} &\text{minimize } \tfrac{1}{2}u^T Hu + u^T b \\ &\text{subject to } -H(a)u - g = 0, \quad a^T Wa \leqq \gamma^2, \quad a \geqq \alpha, \end{aligned} \tag{5.11}$$

where $a = \text{col}\,(a_0, \cdots, a_n) \in \mathbb{R}^{n+1}$ and $u = \text{col}\,(u_1, \cdots, u_{n-1}) \in \mathbb{R}^{n-1}$ are the coefficient vector of $a^n(x)$ and $u^n(x)$, respectively. H, $H(a)$, and W are symmetric tridiagonal matrices and they are given by

$$H_{i,j} = \int_0^1 (B_i^{(n)})_x (B_j^{(n)})_x\,dx,$$

$$H(a)_{i,j} = \int_0^1 a^n (B_i^{(n)})_x (B_j^{(n)})_x\,dx, \quad i,j = 1, \cdots, n-1,$$

$$W_{k,l} = \int_0^1 (B_i^{(n)})_x (B_j^{(n)})_x\,dx, \qquad k = 1 = 0, \cdots, n.$$

The vectors b and $f \in \mathbb{R}^{n-1}$ are given by

$$b_i = \int_0^1 (B_i^{(n)})_x z_x\,dx \quad \text{and} \quad g_i = \int_0^1 B_i^{(n)} g\,dx, \quad 1 \leqq i \leqq n-1.$$

The augmented Lagrangian method applied to (5.11) involves a sequence of minimizations of functionals of the form

$$\begin{aligned} L_{\hat{c}}(a, u; \lambda^k, \mu^k) = {} & \tfrac{1}{2}u^T Hu + u^T b + \lambda^{k^T} e(a,u) \\ & + \frac{\hat{c}}{2} e(a,u)^T H^{-1} e(a,u) \\ & + \frac{1}{2\hat{c}} \left| \max\left(0, \frac{\hat{c}}{2}(a^T Wa - \gamma) + \mu^k\right) \right|^2 \end{aligned} \tag{5.12}$$

over $(a, u) \in \mathbb{R}^{n-1} \times \mathbb{R}^n$ subject to $a \geqq \alpha$, where $e(a, u) = -H(a)u - g \in \mathbb{R}^{n-1}$. The sequence of Lagrange multipliers $\lambda \in \mathbb{R}^{n-1}$, $\mu \in \mathbb{R}^+$ is updated by

$$\lambda^{k+1} = \lambda^k + cH^{-1}e(a^k, u^k),$$

$$\mu^{k+1} = \max\left(0, \frac{c}{2}(a^{kT}Wa^k - \gamma) + \mu^k\right),$$

where (a^k, u^k) minimizes (5.12). We refer to [IK1] and [IKK] for a detailed discussion of the convergence properties of the augmented Lagrangian method.

We generated the test examples as follows. First, we chose the pair $(a^*, u^*) \in H^1(0, 1) \times H_0^1(0, 1)$ and set $g = -(a^* u_x^*)_x$. Then the measurement $z(x)$ is constructed as the linear interpolation of point measurements $\xi_i = u^*(x_i) + n_i$ at x_i, $i = 1, \cdots, n-1$, where $\{n_i\}$ are independent, uniformly distributed random variables in $[-\Delta, \Delta]$, i.e.,

$$z(x) = \sum_{i=1}^{n-1} \xi_i B_i^{(n)}(x) \in H_0^1(0, 1). \tag{5.13}$$

We varied the parameters n (equal to the number of elements for approximating $a(x)$ and $u(x)$) and Δ.

We used the following two examples for testing the proposed algorithm.

Example 2 (smooth a). We consider

$$u^*(x) = e^{-x} \sin 2\pi x, \qquad a^*(x) = 1 + x^2.$$

TABLE 2
Example 2, $n = 10$.

Δ		.01	.03	.05
No. of iterations		4	5	5
Initial set of parameters in (5.8)	a	2.8937×10^{-1}	2.7121	3.1762
	b	-5.3986×10^{-1}	-1.9551	-1.8582
	c	-5.4079×10^{-1}	4.7566×10^{-2}	1.5498×10^{-1}
	d	1.1077×10^{-1}	-1.1903×10^{-2}	-3.8191×10^{-2}
Final set of parameters in (5.8)	a	3.5932	2.7337	2.7138
	b	-1.9587	-1.9565	-1.8521
	c	4.0059×10^{-3}	4.8304×10^{-2}	1.4302×10^{-1}
	d	-1.5692×10^{-3}	-1.2362×10^{-2}	-3.0294×10^{-2}
γ updates	γ_2	1.20	0.57	0.42
	γ_3	1.43	1.39	1.24
	γ_4	1.36	1.261	1.175
	γ_5		1.265	1.177
Best γ		1.30	1.20	1.16
$\|a^* - a_\gamma\|_{L^2}$		3.17×10^{-2}	5.39×10^{-2}	7.229×10^{-2}
$\|a^* - a_{\gamma_{\text{best}}}\|_{L^2}$		3.14×10^{-2}	5.36×10^{-2}	7.227×10^{-2}
$\mu(\gamma_{\text{best}})$		6.31×10^{-3}	2.72×10^{-2}	4.88×10^{-2}
Noise level		6.06×10^{-3}	5.46×10^{-2}	1.52×10^{-1}

TABLE 3
Example 3, $n = 50$.

Δ		.01	.03	.05
No. of iterations		4	4	4
Initial set of parameters in (5.8)	a	1.2027×10^{-1}	1.0876×10^{-1}	1.0752×10^{-1}
	b	-1.0209	-1.0016	-7.2539×10^{-1}
	c	2.2913×10^{-1}	-1.4299	3.8563
	d	-1.4964×10^{-3}	-3.1721×10^{-3}	-5.3218×10^{-3}
Final set of parameters in (5.8)	a	1.1500×10^{-1}	1.0431×10^{-1}	1.0572×10^{-1}
	b	-1.0383	-1.0068	-7.2714×10^{-1}
	c	2.2312×10^{-1}	1.4227	3.8524
	d	-1.3460×10^{-3}	-2.9982×10^{-3}	-5.2273×10^{-3}
γ updates	γ_3	39.5	36.0	34.1
	γ_4	37.8	35.2	33.9
Best γ		49	51	57
$\|a^*-a_\gamma\|_{L^2}$		7.26×10^{-2}	1.42×10^{-1}	2.03×10^{-1}
$\|a^*-a_{\gamma_{\text{best}}}\|_{L^2}$		6.92×10^{-2}	1.39×10^{-1}	1.82×10^{-1}
$\mu(\gamma_{\text{best}})$		1.24×10^{-3}	2.73×10^{-3}	4.82×10^{-3}
Noise level		1.61×10^{-1}	1.45	4.03

The computations were carried out with $n = 10$, the startup value $\gamma = 1$, and the penalty parameter $\hat{c} = 1$ in the augmented Lagrangian method (5.12).

Table 2 summarizes our numerical findings. Plots comparing the minimal value function with the model function show very good agreement of these functions over the interval $[1, 1.5]$. For $\Delta = .01$, .03, and .05 we also plotted the values of $|a^* - a(\gamma)|_{L^2}$ with $a(\gamma)$ the solution of (5.10) against $\gamma \in [1, 1.15]$ and we observed that these functions have a distinct global minimum.

Example 3 (rapid change in the derivative of a^*). We consider

$$u^*(x) = e^{-x}\sin 2\pi x, \qquad a^*(x) = 1.5 + .5\tan^{-1}(1500(x - .4)).$$

The computations were carried out with $n = 50$ and the penalty parameter $\hat{c} = 5$ in (5.12). In order to obtain accurate solutions to the constrained minimization (5.10) for this example we used $\gamma_1 = 15$ and $\gamma_2 = 30$ in Step 2 of the algorithm. It was necessary to start the algorithm with a relatively small γ in order to calculate accurate Lagrange multipliers $\mu(\gamma)$ successively. The corresponding numerical results are shown in Table 3.

As a conclusion to our numerical studies, we observe that in both examples that we considered, the algorithm performed very well. The model function of form (5.8) provided a good approximation to $F(\gamma)$ over the range of γ-values that we dealt with. We used a priori knowledge for the lower bound of γ in both examples in this section. It is possible to obtain an estimate for such a bound by solving the β problem with an a priori chosen β_0 as described in Step 1 of the algorithm. A preliminary numerical study indicates that use of such an estimate as the startup value is very promising. A detailed study will be reported elsewhere.

REFERENCES

[B] J. BAUMEISTER. *Stable Solutions of Inverse Problems*, Vieweg, Braunschweig, Germany, 1987.

[CK] F. COLONIUS AND K. KUNISCH, *Output least squares stability in elliptic systems*, Appl. Math. Optim., 19 (1988), pp. 33-63.

[DS] J. E. DENNIS AND R. B. SCHNABEL, *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*, Prentice-Hall, Englewood Cliffs, NJ, 1983.

[EKN] H. W. ENGL, K. KUNISCH, AND A. NEUBAUER, *Tikhonov regularisation for the solution of nonlinear ill-posed problems*, Inverse Problems, 5 (1989), pp. 523-540.

[G] C. W. GROETSCH, *The Theory of Tikhonov Regularization for Fredholm Equations of the First Kind*, Pitman, Boston, 1984.

[H] A. HARAUX, *How to differentiate the projection on a convex set in Hilbert space. Some applications to variational inequalities*, J. Math. Soc. Japan, 29 (1977), pp. 615-631.

[IK1] K. ITO AND K. KUNISCH, *The augmented Lagrangian method for parameter estimation in elliptic systems*, SIAM J. Control Optim., 28 (1990), pp. 113-136.

[IK2] ———, *Sensitivity analysis of solutions to optimization problems in Hilbert spaces with applications to optimal control and estimation*, submitted.

[IKK] K. ITO, M. KROLLER, AND K. KUNISCH, *A numerical study of an augmented Lagrangian method for the estimation of parameters in elliptic systems*, SIAM J. Sci. Statist. Comput., 12 (1992), pp. 884-910.

[K] T. KATO, *Perturbation Theory for Linear Operators*, Springer-Verlag, Berlin, 1966.

[Ku1] K. KUNISCH, *Inherent identifiability of parameters in elliptic differential equations*, J. Math. Anal. Appl., 132 (1988), pp. 453-472.

[Ku2] K. KUNISCH, *On a class of damped Morozov principles*, submitted.

[KS] C. KRAVARIS AND J. H. SEINFELD, *Identification of parameters in distributed systems by regularization*, SIAM J. Control Optim., 23 (1985), pp. 217-241.

[L] A. K. LOUIS, *Inverse und schlecht gestellte Probleme*, Teubner, Stuttgart, 1989.

[LM] F. LEMPIO AND H. MAURER, *Differential stability in infinite-dimensional nonlinear programming*, Appl. Math. Optim., 6 (1980), pp. 139-152.

[M] V. A. MOROZOV, *Methods for Solving Incorrectly Posed Problems*, Springer-Verlag, New York, 1984.

[MZ] H. MAURER AND J. ZOWE, *First and second order necessary and sufficient optimality for infinite-dimensional programming problems*, Math. Programming, 16 (1979), pp. 98-110.

[N] A. NEUBAUER, *Tokhonov regularization for nonlinear ill-posed problems: Optimal convergence rates and finite dimensional approximation*, Inverse Problems, 5 (1989), pp. 541-559.

[SV] T. I. SEIDMAN AND C. R. VOGEL, *Wellposedness and convergence of some regularization methods for non-linear ill-posed problems*, Inverse Problems, 5 (1989), pp. 227-238.

[TA] A. N. TIKHONOV AND V. Y. ARSENIN, *Solutions of Ill-Posed Problems*, John Wiley, New York, 1977.

[V] C. R. VOGEL, *A constrained least squares method for nonlinear ill-posed problems*, SIAM J. Control Optim., 28 (1990), pp. 34-49.

[Y] W. W.-G. YEH, *Review of parameter identification procedures in groundwater hydrology: The inverse problem*, Water Res. Rev., 22 (1986), pp. 95-108.

SIAM J. OPTIMIZATION
Vol. 2, No. 3, pp. 405-424, August 1992

004

SECOND DERIVATIVES OF A CONVEX FUNCTION AND OF ITS LEGENDRE-FENCHEL TRANSFORMATE*

ALBERTO SEEGER†

Abstract. In 1977 Crouzeix established a simple relationship between the second-order differentials of a convex function $f:\mathbb{R}^n\to\mathbb{R}\cup\{+\infty\}$ and its Legendre-Fenchel transformate $f^*:\mathbb{R}^n\to\mathbb{R}\cup\{+\infty\}$. In the first part of this paper, the importance of Crouzeix's formula is enhanced by illustrating how it can be applied to a large number of classical and modern mathematical problems. In the second part, the result of Crouzeix is extended to the case in which the functions f and f^* are not necessarily smooth. This generalization is based on the works of Hiriart-Urruty and Seeger concerning the so-called second-order subdifferential of a convex function.

Key words. Legendre transformate, conjugate, subdifferential, second-order subdifferential, second-order directional derivative, Monge-Ampère measure, Monge-Ampère operator, curvature, umbilic point, infimal convolution, Cramer transform, canonical exponential family, maximum likelihood, second-order epidifferentiability, piecewise linear quadratic function

AMS(MOS) subject classifications. primary 52; secondary 44, 62E, 90C

1. Introduction. This paper turns around a result established by Crouzeix [Cr] concerning the second-order differentials of a convex function $f:\mathbb{R}^n\to\mathbb{R}\cup\{+\infty\}$ and of its Legendre-Fenchel transformate $f^*:\mathbb{R}^n\to\mathbb{R}\cup\{+\infty\}$. He showed that if f is twice continuously differentiable on a neighborhood of $\bar{x}\in\mathbb{R}^n$ and if the Hessian matrix $\nabla^2 f(\bar{x})$ of f at $\bar{x}$ is nonsingular, then f^* is twice continuously differentiable on a neighborhood of $\bar{y}=\nabla f(\bar{x})$ and, moreover, the Hessian matrix $\nabla^2 f^*(\bar{y})$ of f^* at $\bar{y}$ is given simply by

$$\nabla^2 f^*(\bar{y})=[\nabla^2 f(\bar{x})]^{-1}.$$

This formula reminds us of an analogous one, already mentioned in classical texts of variational calculus, for the Legendre transformate of a function that is smooth but not necessarily convex. In Crouzeix's paper, as throughout this work, the function f^* is given by

$$f^*(y):=\sup_{x\in\mathbb{R}^n}\{\langle x,y\rangle-f(x)\}\quad\forall y\in\mathbb{R}^n$$

and is called simply the conjugate of f.

In § 2 we introduce some notation and recall properties of the conjugacy operation $f\mapsto f^*$ we should keep in mind throughout the paper.

Section 3 illustrates how Crouzeix's formula can be applied to different types of situations and how its use considerably simplifies the proof of some well-known theorems. We should not underestimate the importance of this formula since it can be used as a basic tool toward a better understanding of several classical and modern mathematical problems. We will try to convince the reader of this fact by exhibiting some of its applications to domains as different as large deviations theory, convex analysis, differential geometry, and the statistical theory of canonical exponential families.

* Received by the editors January 24, 1990; accepted for publication (in revised form) July 1, 1991.

† Department of Mathematics, University of Washington, GN-50, Seattle, Washington 98195. Present address, Département de Mathématiques, Université d'Avignon, 33 rue Louis Pasteur, 84000 Avignon, France.

Section 4 extends Crouzeix's result to the case in which the functions f and f^* are not necessarily smooth. The concept of a Hessian matrix is, of course, meaningless in this more general setting and, therefore, it is replaced by the so-called second-order subdifferential of a convex function. The later notion, introduced by Hiriart-Urruty [H3] and further developed by Seeger [Se], is a suitable tool for formulating and extending Crouzeix's theorem. Indeed, the fact that the Hessian matrices $\nabla^2 f^*(\bar{y})$ and $\nabla^2 f(\bar{x})$ are inverse to each other can be expressed by saying that the second-order subdifferentials $\partial^2 f^*(\bar{y}) \subset \mathbb{R}^n$ and $\partial^2 f(\bar{x}) \subset \mathbb{R}^n$ are polar to each other. Similarly to what is done in § 3, the polarity relationship between the second-order subdifferentials can be used as a basic tool to handle some mathematical problems involving nonsmooth convex functions. However, in this already lengthy paper we will not repeat the same steps as before. Some of the examples of § 3, as well as many others not mentioned, constitute an open field of applications, which we encourage the reader to explore.

2. The conjugate of a convex function: Generalities. Throughout this paper $\Gamma_0(\mathbb{R}^n)$ denotes the set of functions from $\mathbb{R}^n$ into $\mathbb{R} \cup \{+\infty\}$ which are convex lower-semicontinuous and not identically equal to $+\infty$. We are concerned with the conjugacy operation

$$\begin{gathered}\Gamma_0(\mathbb{R}^n) \to \Gamma_0(\mathbb{R}^n),\\ f \mapsto f^*,\end{gathered}$$

where the conjugate f^* of f is given by

$$f^*(y) := \sup_{x \in \mathbb{R}^n} \{\langle x, y\rangle - f(x)\} \quad \forall y \in \mathbb{R}^n.$$

It is a well-known fact that the correspondence $f \mapsto f^*$ is an involution on $\Gamma_0(\mathbb{R}^n)$ and therefore the symmetric formula

$$f(x) = \sup_{y \in \mathbb{R}^n} \{\langle y, x\rangle - f^*(y)\} \quad \forall x \in \mathbb{R}^n$$

also holds. The above conjugacy operation can be found in the literature under different names, such as polarity correspondence [Mo], Fenchel transformation [At], Young transformation [Ku], or maximum transformation [Ir].

We are also concerned with a conjugation of the type

$$\begin{gathered}\Gamma_0(\mathbb{R}^n) \times P(\mathbb{R}^n) \to \Gamma_0(\mathbb{R}^n) \times P(\mathbb{R}^n),\\ (f, A) \mapsto (f^*, A^*),\end{gathered} \tag{2.1}$$

where $P(\mathbb{R}^n)$ stands for the class of all subsets of $\mathbb{R}^n$. For the sake of symmetry and in order to fix some notation, let us define the set $A^* \subset \mathbb{R}^n$ in a rather indirect way. Let π_X and π_Y be the projection mappings from $\mathbb{R}^n \times \mathbb{R}^n$ into $\mathbb{R}^n$ defined by

$$\pi_X(x, y) = x \quad \text{and} \quad \pi_Y(x, y) = y,$$

respectively. Following a procedure that can be found in Kiselman [Ki], we write

$$G := \{(x, y) \in \mathbb{R}^n \times \mathbb{R}^n : f(x) + f^*(y) - \langle x, y\rangle = 0\}$$

and then define

$$A^* := \pi_Y[G \cap \pi_X^{-1}(A)] \quad \forall A \subset \mathbb{R}^n. \tag{2.2}$$

The relationship between A^* and A can be expressed, of course, in terms of the subdifferential mapping

$$x \mapsto \partial f(x) := \{y \in \mathbb{R}^n : f(x') \geqq f(x) + \langle x' - x, y\rangle, \ \forall x' \in \mathbb{R}^n\}$$

or, alternatively, in terms of $y \mapsto \partial f^*(y)$. It is obvious that A^* can be defined equivalently by

$$A^* = \partial f(A) := \bigsqcup \{\partial f(x) : x \in A\} \tag{2.3}$$

or

$$A^* = (\partial f^*)^{-1}(A) := \{y \in \mathbb{R}^n : \partial f^*(y) \text{ intersects } A\}, \tag{2.4}$$

respectively. The equivalence between (2.2), (2.3), and (2.4) is an immediate consequence of the following result, which should be kept in mind throughout this paper.

PROPOSITION 2.1 (cf., for instance, [R1, Thm. 23.5]). *Let $f, f^* \in \Gamma_0(\mathbb{R}^n)$ be a couple of conjugate functions. Then the following statements are equivalent*:

(a) $(x, y) \in G$,

(b) $y \in \partial f(x)$,

(c) $x \in \partial f^*(y)$.

It is possible, of course, to repeat the same procedure once again in order to obtain the conjugate (f^{**}, A^{**}) of the pair (f^*, A^*). Although the equality $f^{**} = f$ holds for each $f \in \Gamma_0(\mathbb{R}^n)$, the set

$$A^{**} := \pi_X[G \cap \pi_Y^{-1}(A^*)] = \partial f^*(A^*)$$

could include A strictly.

The correspondence (2.1) is intimately connected with the classical Legendre transformation for functions that are smooth but not necessarily convex. If f is a differentiable real-valued function on an open subset Ω of $\mathbb{R}^n$ and A is an arbitrary subset of Ω, then the Legendre conjugate $(\tilde{f}, \tilde{A})$ of the pair (f, A) is defined by

$$\tilde{A} := \{\nabla f(x) : x \in A\},$$

$$\tilde{f}(y) := \langle (\nabla f)^{-1}(y), y \rangle - f((\nabla f)^{-1}(y)) \quad \forall y \in \tilde{A}.$$

Note that $\nabla f : A \to \tilde{A}$ is not necessarily one-to-one since a set like

$$(\nabla f)^{-1}(y) := \{x \in \mathbb{R}^n : \nabla f(x) = y\}$$

could have more than one element. In such a case, $\tilde{f}$ should be interpreted as a multivalued mapping.

3. Second-order differentiability of the conjugate: The smooth case. For each $f \in \Gamma_0(\mathbb{R}^n)$ we write, as customary,

$$\operatorname{dom} f := \{x \in \mathbb{R}^n : f(x) < +\infty\}$$

and

$$\operatorname{dom} \nabla f := \{x \in \operatorname{dom} f : f \text{ is differentiable at } x\}.$$

The definition of the sets $\operatorname{dom} f^*$ and $\operatorname{dom} \nabla f^*$ is obvious. Recall that on the set $\operatorname{dom} \nabla f$, and only there, the subdifferential mapping ∂f is single valued and reduces to the gradient mapping ∇f (cf. [R1, Thm. 25.1]).

A natural question to ask in the context of the present section is: Which conditions should we impose on $f \in \Gamma_0(\mathbb{R}^n)$ to ensure that the conjugate f^* is twice differentiable at a given point? We would also like to know how the functions $y \mapsto \nabla^2 f^*(y)$ and $x \mapsto \nabla^2 f(x)$ are related to each other when the Hessian matrices involved are well defined. A full answer to this question can be found, for instance, in a short paper by Crouzeix [Cr].

THEOREM 3.1 [Cr]. *Let $f \in \Gamma_0(\mathbb{R}^n)$ be twice continuously differentiable on a neighborhood of $\bar{x} \in \mathbb{R}^n$ and assume that the Hessian matrix $\nabla^2 f(\bar{x})$ is nonsingular. Then the conjugate function $f^* \in \Gamma_0(\mathbb{R}^n)$ is twice continuously differentiable on a neighborhood of $\bar{y} = \nabla f(\bar{x})$ and we have*

$$\nabla^2 f^*(\bar{y}) = [\nabla^2 f(\bar{x})]^{-1}. \tag{3.1}$$

It is fair to say that an analogous version of the formula (3.1) for the Legendre transformate $\tilde{f}$ was anticipated by other authors. For instance, it was already mentioned in classical texts of variational calculus (cf. [Fu, p. 107], [GF, p. 72], [Ru, p. 18]). In the framework of the theory of elastic materials, Hill and Rice [HR, pp. 451-452] stated in 1973 that, under suitable assumptions, the fourth-rank tensors of elastic moduli and compliances

$$L := \nabla^2 f(x) \quad \text{and} \quad M := \nabla^2 \tilde{f}(y)$$

are inverse to each other if the symmetric strain tensor x and the symmetric conjugate stress tensor y are related by the law

$$y = \nabla f(x).$$

The functions f and $\tilde{f}$ are interpreted as a work potential and a complementary potential, respectively. The inner product $\langle \cdot, \cdot \rangle$, which appears in the definition of $\tilde{f}$ takes the form

$$\langle x, y \rangle = \sum_{i,j} x_{ij} y_{ij}$$

in this case. We have intentionally modified the notation of Hill and Rice in order to put in evidence the connection existing between their statement and Crouzeix's theorem. Although in a very particular setting, the formula

$$\nabla^2 \tilde{\mathrm{f}}(\mathrm{y}) = [\nabla^2 \mathrm{f}(\mathrm{x})]^{-1} \quad \text{with } y = \nabla f(x) \tag{3.2}$$

was also exploited by Sewell [S1, p. 148] in the context of the theory of elementary catastrophes (see also [S2, p. 285]). Despite the references quoted above, it seems to us that many authors are not completely aware of the large domain of applications for the formulae (3.1) and (3.2). It is the purpose of this section to enhance the importance of Crouzeix's theorem by illustrating how it can be applied to different types of situations and how its use considerably reduces the proof of some well-known theorems.

For convenience, let us introduce the following notation. For $f \in \Gamma_0(\mathbb{R}^n)$ we write

$$x \in C(f) \overset{\text{def}}{\Leftrightarrow} \begin{cases} f \text{ is twice continuously differentiable on a} \\ \text{neighborhood of } x \text{ and } \nabla^2 f(x) \text{ is nonsingular} \end{cases}$$

and set

$$\mathcal{L} := \{(f, A) : f \in \Gamma_0(\mathbb{R}^n) \text{ and } A \subset C(f)\}.$$

This allows us to state in a compact way most of the results given in this section, in particular, the next trivial corollary of Theorem 3.1.

COROLLARY 3.2. *The following equivalence holds:*

$$(f, A) \in \mathcal{L} \Leftrightarrow (f^*, A^*) \in \mathcal{L}.$$

Moreover, the conjugacy correspondence (2.1) *restricted to $\mathcal{L} \subset \Gamma_0(\mathbb{R}^n) \times P(\mathbb{R}^n)$ is an involution.*

Proof. The proof is trivial. □

We point out immediately that $(f, A) \mapsto (f^*, A^*)$ coincides over $\mathcal{L}$ with the Legendre transformation $(f, A) \mapsto (\tilde{f}, \tilde{A})$. Keeping in mind Crouzeix's theorem and the above corollary, let us consider then the following mathematical problems.

3.1. The Monge–Ampère measure associated to a convex function. The Monge-Ampère measure m_f associated to a convex function f has a long history. It is intimately connected with the multidimensional Monge-Ampère equation

$$\det[\nabla^2 f(x)] = \varphi(x) \quad \forall x \in \Omega, \tag{3.3}$$

where φ stands for a known continuous function that is nonnegative on the nonempty open convex set $\Omega \subset \mathbb{R}^n$. If $f: \Omega \subset \mathbb{R}^n \to \mathbb{R}$ is a twice continuously differentiable convex function which solves the above equation, then, of course, we can write

$$m_f(A) = \int_A \varphi(x)\,dx \quad \forall A \in B_\Omega, \tag{3.4}$$

where B_Ω denotes the class of Borel sets in Ω and

$$m_f(A) := \int_A \det \nabla^2 f(x)\,dx. \tag{3.5}$$

A convex function $f: \Omega \subset \mathbb{R}^n \to \mathbb{R}$ that verifies (3.4) is called a generalized or weak solution of the Monge-Ampère equation (cf., for instance, [P, p. 70] and [CY]). This concept is meaningful as long as m_f is well defined. The following well-known proposition (cf. [RT, p. 355]) gives a different characterization of m_f and shows that weak solutions do not actually need to be twice continuously differentiable.

PROPOSITION 3.3. *Let λ be the (n-dimensional) Lebesgue measure. If $(f, \Omega) \in \mathcal{L}$, then*

$$m_f(A) = \lambda(\{\nabla f(x): x \in A\}) \quad \forall A \in B_\Omega.$$

Proof. Let A be an arbitrary Borel set in Ω. Since $(f, \Omega) \in \mathcal{L}$, the function $\nabla f: \mathrm{A} \to \mathrm{A}^*$ is invertible and has $\nabla f^*: A^* \to A$ as its inverse. Moreover, the Jacobian determinant of the change of variables $x = \nabla f^*(y)$ is well defined and verifies

$$\det \nabla^2 f^*(y) > 0 \quad \forall y \in A^*.$$

The general formula for a change of variables in a multiple integral yields in this case

$$\int_A \det \nabla^2 f(x)\,dx = \int_{(\nabla f^*)^{-1}(A)} \det \nabla^2 f(\nabla f^*(y)) \det \nabla^2 f^*(y)\,dy$$
$$= \int_{\nabla f(A)} 1\,dy.$$

The last equality is obtained, of course, by using Crouzeix's formula (3.1) where the roles of f and f^* are exchanged. The proof is then completed. □

Note that in the statement of Proposition 3.3 there is no explicit mention of the Legendre transformation. This concept is used only as an auxiliary tool in the above demonstration. Note also that the assumption $(f, \Omega) \in \mathcal{L}$ arises in a natural way if we expect that $f: \Omega \subset \mathbb{R}^n \to \mathbb{R}$ solves the Monge-Ampère equation (3.3) for a positive continuous function φ.

3.2. Integrating a function of the Monge–Ampère operator. In some applications it is not the density

$$x \mapsto M_f(x) := \det \nabla^2 f(x)$$

we should integrate, but some function u of it. In other words, we are concerned with the evaluation of an integral of the type

$$\int_A u \circ M_f := \int_A u[M_f(x)]\, dx.$$

A typical one-dimensional example is the expression

$$\int_a^b [f''(x)]^2\, dx,$$

which appears in the variational characterization of the cubic interpolation spline (cf. [Li, Thm. 2.10]).

Suppose for simplicity that $u:]0, \infty[\to]0, \infty[$ is a continuous function. For convenience, let us introduce the "dual" function $\hat{u}:]0, \infty[\to]0, \infty[$ of u given by

$$\hat{u}(s) = su(1/s) \qquad \forall s \in]0, \infty[.$$

Note that $\hat{u}$ is also continuous and has u as its dual function, i.e.,

$$(\hat{u})\hat{} = u.$$

Important examples of pairs of dual functions are:

(a) $u(t) = t, \hat{u}(s) = 1$;

(b) $u(t) = \sqrt{t}, \hat{u}(s) = \sqrt{s}$;

(c) $u(t) = t^2, \hat{u}(s) = 1/s$;

(d) $u(t) = \log t, \hat{u}(s) = -s \log s$.

The following result is then a generalization of Proposition 3.3.

PROPOSITION 3.4. *Let $u:]0, \infty[\to]0, \infty[$ be continuous and $(f, \Omega) \in \mathscr{L}$. Then*

$$\int_A u \circ M_f = \int_{A^*} \hat{u} \circ M_{f^*} \quad \forall A \in B_\Omega.$$

Proof. The proof is similar to the demonstration of Proposition 3.3. □

Since the function $t \to u(t) = \sqrt{t}$ is dual of itself, we get, in particular, the following invariant property of the conjugacy operation $(f, A) \mapsto (f^*, A^*)$.

COROLLARY 3.5. *Let $(f, \Omega) \in \mathscr{L}$. Then, for all $A \in B_\Omega$,*

$$\int_A \sqrt{M_f} = \int_{A^*} \sqrt{M_{f^*}}.$$

3.3. Umbilic points of conjugate convex hypersurfaces. Let us consider the hypersurface

$$\Sigma_\Omega(f) := \{(x, f(x)) : x \in \Omega\} \subset \mathbb{R}^{n+1}$$

associated with a convex function $f: \Omega \subset \mathbb{R}^n \to \mathbb{R}$. There are some particular points in Ω that deserve special attention, namely, the points in the subset

$$U_f := \{x \in \Omega: \nabla^2 f(x) \text{ exists and } \lambda_{\max}(\nabla^2 f(x)) = \lambda_{\min}(\nabla^2(f(x)))\},$$

where $\lambda_{\max}(H)$ and $\lambda_{\min}(H)$ denote, respectively, the largest and smallest eigenvalue of the matrix H. If $x \in U_f$, then the quadratic form

$$h \mapsto \langle h, \nabla^2 f(x) h \rangle$$

is constant over the unit sphere

$$S := \{h \in \mathbb{R}^n: \|h\| = 1\}$$

and this means geometrically that the "curvature" of the hypersurface $\Sigma_\Omega(f)$ at the point $(x, f(x))$ is the same in all the directions $h \in \mathbb{R}^n$. For this reason a point $x \in U_f$ is simply called an umbilic point of f (although this is not the most common definition of this concept found in the literature).

Similar definitions and remarks apply, of course, to the conjugate hypersurface

$$\Sigma_{\Omega^*}(f^*) := \{(y, f^*(y)) \colon y \in \Omega^*\} \subset \mathbb{R}^{n+1}.$$

The next proposition establishes a relationship between the umbilic points of f and those of its conjugate function f^*.

PROPOSITION 3.6. *Let $(\bar{x}, \bar{y}) \in G$ (cf. Proposition 2.1). Then the following statements are equivalent:*

(a) *$\bar{x} \in C(f)$ is an umbilic point of f;*

(b) *$\bar{y} \in C(f^*)$ is an umbilic point of f^*.*

Proof. As an immediate consequence of Crouzeix's theorem, it follows that

$$\lambda_{\max}(\nabla^2 f^*(\bar{y})) = \lambda_{\max}([\nabla^2 f(\bar{x})]^{-1}) = [\lambda_{\min}(\nabla^2 f(\bar{x}))]^{-1}$$

and

$$\lambda_{\min}(\nabla^2 f^*(\bar{y})) = \lambda_{\min}([\nabla^2 f(\bar{x})]^{-1}) = [\lambda_{\max}(\nabla^2 f(\bar{x}))]^{-1}.$$

This proves, of course, the equivalence announced in the present proposition. □

Note that the points in $C(f) \cap U_f$ are necessarily "nonflat" umbilic points of f.

COROLLARY 3.7. *There is a biunivoque correspondence between the nonflat umbilic points of f and the nonflat umbilic points of f^*. More precisely,*

$$\nabla f \colon C(f) \cap U_f \to C(f^*) \cap U_{f^*}$$

is a bijection.

3.4. The Hessian matrix of the infimal convolution. There is a basic operation in convex analysis that has proven useful in many branches of applied mathematics. We are speaking about the infimal convolution

$$x \in \mathbb{R}^n \mapsto [f_1 \,\square\, f_2](x) := \inf_{x_1 + x_2 = x} \{f_1(x_1) + f_2(x_2)\} \tag{3.6}$$

of two functions $f_1, f_2 \in \Gamma_0(\mathbb{R}^n)$. The subdifferential of the convex function $f_1 \,\square\, f_2$ at $\bar{x} \in \mathbb{R}^n$ admits the characterization

$$\partial [f_1 \,\square\, f_2](\bar{x}) = \partial f(\bar{x}_1) \cap \partial f_2(\bar{x}_2), \tag{3.7}$$

where $(\bar{x}_1, \bar{x}_2)$ is any pair at which the infimum in (3.6) is attained, i.e.,

$$\begin{aligned} &\bar{x} = \bar{x}_1 + \bar{x}_2, \\ &[f_1 \,\square\, f_2](\bar{x}) = f_1(\bar{x}_1) + f_2(\bar{x}_2) \in \mathbb{R}. \end{aligned} \tag{3.8}$$

For a proof of this fact and for conditions ensuring the existence of such a pair $(\bar{x}_1, \bar{x}_2)$, see, for instance, [La. § 6] or [Mo]. These conditions are related to the lower-semicontinuity of $f_1 \,\square\, f_2$ at the point $\bar{x}$.

Note that the equality (3.7) takes the form

$$\nabla [f_1 \,\square\, f_2](\bar{x}) = \nabla f_1(\bar{x}_1) = \nabla f_2(\bar{x}_2) \tag{3.9}$$

if the functions f_1 and f_2 are differentiable at $\bar{x}_1$ and $\bar{x}_2$, respectively. The following result concerning the Hessian matrix of $f_1 \,\square\, f_2$ was established by Hiriart-Urruty [H1].

PROPOSITION 3.8. *Let $f_1, f_2 \in \Gamma_0(\mathbb{R}^n)$ and $(\bar{x}_1, \bar{x}_2)$ be a pair verifying* (3.8). *If $\bar{x}_1 \in C(f_1)$ and $\bar{x}_2 \in C(f_2)$, then $\bar{x} \in C(f_1 \,\Box\, f_2)$ and we can write*

$$\nabla^2[f_1 \,\Box\, f_2](\bar{x}) = \{[\nabla^2 f_1(\bar{x}_1)]^{-1} + [\nabla^2 f_2(\bar{x}_2)]^{-1}\}^{-1}.$$

Proof. The demonstration is based on the conjugacy formula

$$(f_1 \,\Box\, f_2)^* = f_1^* + f_2^*.$$

If $\bar{y} \in \mathbb{R}^n$ denotes the common vector in (3.9), then Crouzeix's theorem allows us to write

$$\bar{y} \in C(f_1^*) \cap C(f_2^*). \tag{3.10}$$

Therefore

$$\nabla(f_1^* + f_2^*)(\bar{y}) = \nabla f_1^*(\bar{y}) + \nabla f_2^*(\bar{y}) = \bar{x}_1 + \bar{x}_2 = \bar{x}$$

and

$$\nabla^2(f_1^* + f_2^*)(\bar{y}) = \nabla^2 f_1^*(\bar{y}) + \nabla^2 f_2^*(\bar{y}).$$

Using Crouzeix's formula (3.1) for each of the above matrices, we get finally

$$[\nabla^2(f_1^* + f_2^*)^*(\bar{x})]^{-1} = [\nabla^2 f_1(\bar{x}_1)]^{-1} + [\nabla^2 f_2(\bar{x}_2)]^{-1}.$$

But

$$(f_1^* + f_2^*)^* = [f_1 \,\Box\, f_2]^{**} = \mathrm{cl}\,[f_1 \,\Box\, f_2]$$

and it can be proven that, in this case, $f_1 \,\Box\, f_2$ and its lower-semicontinuous hull cl $[f_1 \,\Box\, f_2]$ coincide. For this it suffices to consider (3.10) and to apply [R1, Thm. 16.4]. The proof is completed in this way. □

3.5. The Hessian matrix of the Cramer transformate. The Laplace transformate of a probability measure μ on $\mathbb{R}^n$ is the function $L_\mu : \mathbb{R}^n \to]0, +\infty]$ given by

$$L_\mu(x) := \int e^{\langle x,t \rangle}\, d\mu(t) \quad \forall x \in \mathbb{R}^n.$$

The cumulant transformate $K_\mu : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ of μ is simply defined as

$$K_\mu(x) := \log L_\mu(x) \quad \forall x \in \mathbb{R}^n$$

and its conjugate

$$y \in \mathbb{R}^n \mapsto K_\mu^*(y) := \sup_{x \in \mathbb{R}^n} \{\langle x, y \rangle - \log L_\mu(x)\}$$

is usually referred to as the Cramer transformate of μ. The multiple uses of the transformates K_μ and K_μ^* explain the diversity of names attributed to each one of these functions. The reader who is interested in their different interpretations and their main properties may find Table 1 helpful.

TABLE 1

K_μ	K_μ^*	References
cumulant generating function	—	Rahman [Ra]
—	Cramer transformate	Azencott [Az]
cumulant transformate	sup-log-likelihood function	Barndorff-Nielsen [B-N]
free energy function	(level 1) entropy function	Ellis [E1], [E2]

For later use let us recall some basic properties of the functions L_μ and K_μ that are well known in the literature. Let us introduce first the set

$$D := \{x \in \mathbb{R}^n : L_\mu(x) < +\infty\}$$
$$= \{x \in \mathbb{R}^n : K_\mu(x) < +\infty\}$$

and denote by int D its interior.

LEMMA 3.9 (cf. Barndorff-Nielsen [B-N, Thm. 4.1]). *The Laplace transformate L_μ is infinitely often differentiable at every point $\bar{x} \in \operatorname{int} D$ and the derivatives may be computed by differentiation under the integration sign. In particular, the gradient $\nabla L_\mu(\bar{x})$ and the Hessian matrix $\nabla^2 L_\mu(\bar{x})$ are given by*

$$\frac{\partial}{\partial x_i} L_\mu(\bar{x}) = \int t_i e^{\langle \bar{x}, t\rangle} d\mu(t) \quad \forall i = 1, \cdots, n$$

and

$$\frac{\partial^2}{\partial x_i \, \partial x_j} L_\mu(\bar{x}) = \int t_i t_j e^{\langle \bar{x}, t\rangle} d\mu(t) \quad \forall i, j = 1, \cdots, n,$$

respectively.

LEMMA 3.10 (cf. Barndorff-Nielsen [B-N, Thm. 4.1] and Ellis [E1, Thm. 7.5.1]). *The cumulant transformate K_μ belongs to $\Gamma_0(\mathbb{R}^n)$ and is infinitely often differentiable at every point $\bar{x} \in \operatorname{int} D$. In particular*

$$\nabla K_\mu(\bar{x}) = [L_\mu(\bar{x})]^{-1} \nabla L_\mu(\bar{x})$$

and

$$\nabla^2 K_\mu(\bar{x}) = [L_\mu(\bar{x})]^{-2}[L_\mu(\bar{x}) \nabla^2 L_\mu(\bar{x}) - \nabla L_\mu(\bar{x}) \nabla L_\mu(\bar{x})^T].$$

Now we are ready to establish the following result.

PROPOSITION 3.11. *Let $\bar{x} \in \operatorname{int} D$ and assume that the matrix*

$$A(\bar{x}) := L_\mu(\bar{x}) \nabla^2 L_\mu(\bar{x}) - \nabla L_\mu(\bar{x}) \nabla L_\mu(\bar{x})^T$$

is nonsingular. Then the Cramer transformate K_μ^* is twice continuously differentiable on a neighborhood of $\bar{y} = [L_\mu(\bar{x})]^{-1} \nabla L_\mu(\bar{x})$ and we have

$$\nabla^2 K_\mu^*(\bar{y}) = [L_\mu(\bar{x})]^2 [A(\bar{x})]^{-1}.$$

Proof. The proof is an immediate consequence of Crouzeix's theorem and Lemma 3.10. □

The above proposition can be complemented with the use of the Sherman-Morrison formula [Ho, p. 123]

$$[\sigma H - pp^T]^{-1} = \sigma^{-1} \left\{ H^{-1} + \frac{H^{-1} pp^T H^{-1}}{\sigma - \langle p, H^{-1} p\rangle} \right\}.$$

Suitable assumptions could be made, then, on $\sigma = L(\bar{x})$, $p = \nabla L_\mu(\bar{x})$, and $H = \nabla^2 L_\mu(\bar{x})$ to ensure that $A(\bar{x}) = \sigma H - pp^T$ is a nonsingular matrix; namely, we could assume that H is nonsingular and $\langle p, H^{-1} p\rangle \neq \sigma$.

A very important particular case of Proposition 3.11 is the one in which $\bar{x} = 0$. The condition $0 \in \operatorname{int} D$ amounts to saying that L_μ, or equivalently K_μ, is finite on a neighborhood of 0. In such a case Lemmata 3.9 and 3.10 show that

$$\frac{\partial K_\mu}{\partial x_i}(0) = \int t_i \, d\mu(t) \quad \forall i = 1, \cdots, n$$

and

$$\frac{\partial^2 K_\mu}{\partial x_i\,\partial x_j}(0)=\int t_i t_j\,d\mu(t)-\int t_i\,d\mu(t)\int t_j\,d\mu(t)\quad \forall i,j=1,\cdots,n.$$

In other words, $\nabla K_\mu(0)$ and $\nabla^2 K_\mu(0)$ coincide, respectively, with the mean value vector m_μ and the covariance matrix V_μ of μ.

COROLLARY 3.12. *Let L_μ be finite on a neighborhood of* 0 *and assume that the covariance matrix V_μ is nonsingular. Then the Cramer transformate K_μ^* is twice continuously differentiable on a neighborhood of m_μ and we have*

$$\nabla^2 K_\mu^*(m_\mu)=V_\mu^{-1}.$$

Under the hypotheses of the above corollary it follows that K_μ^* has a second-order Taylor expansion around m_μ and it takes the form

$$K_\mu^*(y)=\tfrac{1}{2}\langle y-m_\mu,\,V_\mu^{-1}(y-m_\mu)\rangle+0(\|y-m_\mu\|^2)\quad \forall y\in\mathbb{R}^n.$$

3.6. Maximum likelihood estimation in canonical exponential families. Let us recall some basic facts concerning the canonical exponential families. For this material, the reader is referred to Barndorff-Nielsen [B-N]. Let μ be a probability measure on $\mathbb{R}^n$ and let $L_\mu:\mathbb{R}^n\to]0,\infty]$ be its Laplace transformate. Let us denote by Θ the set of "parameters" for which L_μ is finite, i.e.,

$$\Theta:=\left\{\theta\in\mathbb{R}^n\colon \int e^{\langle\theta,t\rangle}\,d\mu(t)<+\infty\right\}.$$

By the canonical exponential family generated by μ we mean the family $\{\mu_\theta\colon \theta\in\Theta\}$ of probability measures on $\mathbb{R}^n$ which are absolutely continuous with respect to μ and such that their densities are given by

$$\frac{d\mu_\theta}{d\mu}=a(\theta)\,e^{\langle\theta,\cdot\rangle}.$$

The normalizing coefficient $a(\theta)$ verifies, of course,

$$a(\theta)=\frac{1}{L_\mu(\theta)}\quad \forall\theta\in\Theta.$$

A more general definition for this type of family can be found in the literature. Usually, $\{\mu_\theta\colon \theta\in\Theta\}$ is defined not only in terms of μ, but also in terms of a measurable vector-valued function T defined over $\mathbb{R}^n$. We have chosen, once and for all, $T:\mathbb{R}^n\to\mathbb{R}^n$ given by

$$T(t)=t\quad \forall t\in\mathbb{R}^n. \tag{3.11}$$

Let us denote the cumulant transformate of μ simply by K, i.e.,

$$K(\theta):=\log\int e^{\langle\theta,t\rangle}\,d\mu(t)\quad \forall\theta\in\mathbb{R}^n.$$

The log-likelihood function of the family $\{\mu_\theta\colon \theta\in\Theta\}$ corresponding to the observation t is defined as

$$\theta\in\mathbb{R}^n\mapsto l(\theta;t):=\begin{cases}\log[d\mu_\theta/d\mu](t) & \text{if } \theta\in\Theta,\\ -\infty & \text{otherwise.}\end{cases}$$

A straightforward calculus shows that

$$l(\theta;\, t) = \langle \theta, t\rangle - K(\theta)$$

and therefore the sup-log-likelihood function

$$t \in \mathbb{R}^n \mapsto \sup_{\theta\in\mathbb{R}^n} l(\theta;\, t)$$

coincides with the conjugate K^* of K (cf. [B-N, Thm. 6.1]).

In the next proposition the mean value vector $E_\theta T$ and the covariance matrix $V_\theta T$ of T, with respect to μ_θ, play an important role. Recall that T is given by (3.11) and therefore

$$(E_\theta T)_i = \int t_i \, d\mu_\theta(t) \quad \forall i = 1, \cdots, n$$

and

$$(V_\theta T)_{ij} = \int t_i t_j \, d\mu_\theta(t) - \int t_i \, d\mu_\theta(t) \int t_j \, d\mu_\theta(t) \quad \forall i, j = 1, \cdots, n.$$

The next result, due to Barndorff-Nielsen [B-N, Thm. 6.3], can be proved in a straightforward manner by using Crouzeix's theorem.

PROPOSITION 3.13. *Let $\bar\theta$ be in the interior of Θ and assume that the covariance matrix $V_{\bar\theta} T$ is nonsingular. Then the* sup-log-*likelihood function*

$$t \in \mathbb{R}^n \mapsto K^*(t) = \sup_{\theta\in\mathbb{R}^n} l(\theta;\, t)$$

is twice continuously differentiable on a neighborhood of $\bar\tau = E_{\bar\theta} T$ and we have

$$\nabla^2 K^*(\bar\tau) = [V_{\bar\theta} T]^{-1}. \tag{3.12}$$

Proof. A direct calculus shows that

$$E_{\bar\theta} T = \int t a(\bar\theta)\, e^{\langle \bar\theta, t\rangle} \, d\mu(t) = \frac{1}{L_\mu(\bar\theta)} \int t\, e^{\langle \bar\theta, t\rangle} \, d\mu(t) = \nabla K(\bar\theta).$$

The last equality was obtained by using Lemmata 3.9 and 3.10 with the obvious change $D \leftrightarrow \Theta$, $\bar x \leftrightarrow \bar\theta$, and $K_\mu \leftrightarrow K$ in the notation. In a similar way we can get

$$V_{\bar\theta} T = \nabla^2 K(\bar\theta).$$

Formula (3.12) is then a consequence of Proposition 3.11, which in turn was obtained by using Crouzeix's theorem. □

Usually the value of $\bar\tau$ is known first and then the parameter $\bar\theta$ is estimated by solving the so-called log-likelihood equation

$$\nabla_\theta l(\theta;\, \bar\tau) = 0, \tag{3.13}$$

where $\nabla_\theta l$ denotes the gradient of the log-likelihood function l with respect to θ. Under the hypotheses of the above proposition it is clear that $\bar{\tau}=E_{\bar{\theta}}T$ if and only if one of the following equivalent conditions holds:

(a) $\bar{\theta}$ solves the equation (3.13), i.e., $\nabla_\theta l(\bar{\theta},\bar{\tau})=0$;

(b) $\bar{\tau}=\nabla K(\bar{\theta})$;

(c) $\bar{\theta}=\nabla K^*(\bar{\tau})$.

In connection with the log-likelihood equation (3.13), the mapping $\tau\mapsto\nabla^2K^*(\tau)$ plays an important role from a theoretical and from an algorithmic viewpoint.

4. Second-order differentiability of the conjugate: The nonsmooth case. In a first attempt to generalize Crouzeix's theorem, we could try to remove the nonsingularity assumption made on $\nabla^2 f(\bar{x})$ and obtain a similar version of the formula (3.1) with the generalized inverse $[\nabla^2 f(\bar{x})]^+$ used in the place of $[\nabla^2 f(\bar{x})]^{-1}$. Nevertheless, this approach does not lead us too far because, at any rate, the use of the Hessian matrix $\nabla^2 f(\bar{x})$ is involved. In this section we are concerned with convex functions that are not necessarily differentiable and therefore we need to consider a quite different approach. It is based on the so-called second-order subdifferential of a convex function, which has been introduced by Hiriart-Urruty [H3] and further developed by Seeger [Se]. We now recall this notion and its main properties, and we point out the references [H3], [Se], [HS1], and [HS2] for a more complete discussion.

For $f\in\Gamma_0(\mathbb{R}^n)$ and $\bar{x}\in\operatorname{dom}f$ we write, as customary,

$$f'(\bar{x};h)=\lim_{t\to0^+}\frac{f(\bar{x}+th)-f(\bar{x})}{t}\qquad\forall h\in\mathbb{R}^n.$$

The above limit always exists in $\mathbb{R}\cup\{\pm\infty\}$ and for those directions $h\in\mathbb{R}^n$ in

$$\operatorname{dom}f'(\bar{x};\cdot):=\{h\in\mathbb{R}^n: f'(\bar{x};h)<+\infty\}$$

we set

$$\bar{f}''(\bar{x};h):=\limsup_{t\to0^+}\frac{2}{t}\left[\frac{f(\bar{x}+th)-f(\bar{x})}{t}-f'(\bar{x};h)\right].$$

If the above upper limit is actually a limit, then $\bar{f}''(\bar{x};h)$ is denoted simply by $f''(\bar{x};h)$. If this occurs for all $h\in\mathbb{R}^n$ and $f''(\bar{x};\cdot)$ is finite everywhere, then f is said to be twice directionally differentiable at $\bar{x}$.

Given $\bar{x}$ and $\bar{y}\in\partial f(\bar{x})$, the (upper) second-order directional derivative of $f\in\Gamma_0(\mathbb{R}^n)$ at $\bar{x}$ relative to $\bar{y}$ is defined as

$$h\in\mathbb{R}^n\mapsto\bar{f}''(\bar{x},\bar{y};h)=\limsup_{t\to0^+}\frac{2}{t}\left[\frac{f(\bar{x}+th)-f(\bar{x})}{t}-\langle\bar{y},h\rangle\right].$$

The main properties of the function $\bar{f}''(\bar{x},\bar{y};\cdot)$ are:

$$\bar{f}''(\bar{x},\bar{y};0)=0,\quad \bar{f}''(\bar{x},\bar{y};h)\in[0,\infty]\quad\forall h\in\mathbb{R}^n.$$

$$\bar{f}''(\bar{x},\bar{y};\cdot)\text{ is convex and positively homogeneous of degree 2.}$$

It follows then that the square root of the lower-semicontinuous hull of the function $\bar{f}''(\bar{x},\bar{y};\cdot)$ is the support function of a unique closed convex set in $\mathbb{R}^n$ containing 0. This unique set $\partial^2 f(\bar{x},\bar{y})$, called the second-order subdifferential of f at $\bar{x}$ relative to $\bar{y}$, is given by

$$\partial^2 f(\bar{x},\bar{y})=\{z\in\mathbb{R}^n:\langle z,h\rangle\leqq\sqrt{\bar{f}''(\bar{x},\bar{y};h)},\ \forall h\in\mathbb{R}^n\}.$$

As said before, we have the equality

$$\sqrt{\operatorname{cl}\bar{f}''(\bar{x},\bar{y};\cdot)}=\psi^*[\cdot;\partial^2 f(\bar{x},\bar{y})],$$

where

$$h \mapsto \psi^*[h; C] := \sup_{z \in C} \langle z, h \rangle$$

denotes the support function of a set $C \subset \mathbb{R}^n$ and cl g denotes the lower-semicontinuous hull of g.

An intrinsic concept associated only with the point $\bar{x}$ is the second-order subdifferential $\partial^2 f(\bar{x})$ of f at $\bar{x}$, which is merely defined as

$$\partial^2 f(\bar{x}) = \bigcap \{\partial^2 f(\bar{x}, \bar{y}) \colon \bar{y} \in \partial f(\bar{x})\}.$$

The main properties of this set are summarized in the next proposition.

PROPOSITION 4.1. *Let $f \in \Gamma_0(\mathbb{R}^n)$ and $\bar{x} \in \operatorname{dom} f$. Then*

(a) *The second-order subdifferential $\partial^2 f(\bar{x})$ of f at $\bar{x}$ is a closed convex set in $\mathbb{R}^n$ containing* 0.

(b) *If $\bar{x} \in \operatorname{int}(\operatorname{dom} f)$, then the support function of $\underline{\partial^2 f(\bar{x})}$ is equal to the largest convex lower-semicontinuous function, which minorizes $\sqrt{\bar{f}''(\bar{x}; \cdot)}$. In particular, the set $\partial^2 f(\bar{x})$ is compact if $\bar{f}''(\bar{x}; \cdot)$ is finite everywhere.*

If f is twice differentiable at $\bar{x}$, then

$$f''(\bar{x}; h) = \bar{f}''(\bar{x}, \nabla f(\bar{x}); h) = \langle h, \nabla^2 f(\bar{x}) h \rangle \quad \forall h \in \mathbb{R}^n,$$

and the set $\partial^2 f(\bar{x}) = \partial^2 f(\bar{x}, \nabla f(\bar{x}))$ reduces to the ellipsoid associated to the Hessian matrix $\nabla^2 f(\bar{x})$, i.e.,

$$\partial^2 f(\bar{x}) = \{z \in \mathbb{R}^n \colon \langle z, h \rangle \leqq \sqrt{\langle h, \nabla^2 f(\bar{x}) h \rangle}, \forall h \in \mathbb{R}^n\}. \tag{4.1}$$

This set does not necessarily contain 0 in its interior, since the matrix $\nabla^2 f(\bar{x})$ could be singular. The possibility of a "degenerate" ellipsoid is therefore not excluded. If $\nabla^2 f(\bar{x})$ turns out to be nonsingular, then the ellipsoid (4.1) also admits the characterization

$$\partial^2 f(\bar{x}) = \{u \in \mathbb{R}^n \colon \langle u, [\nabla^2 f(\bar{x})]^{-1} u \rangle \leqq 1\}.$$

In this case Crouzeix's formula (3.1) amounts to saying that the sets $\partial^2 f^*(\bar{y})$ and $\partial^2 f(\bar{x})$ are polar to each other, i.e.,

$$\partial^2 f^*(\bar{y}) = [\partial^2 f(\bar{x})]^0 \tag{4.2}$$

where

$$C^0 := \{v \in \mathbb{R}^n \colon \langle v, u \rangle \leqq 1, \forall u \in C\}$$

denotes the polar set of $C \subset \mathbb{R}^n$. We claim that the polarity relationship (4.2) holds not only under the hypotheses of Crouzeix's theorem, but also in a much more general setting. To begin, let us consider the following example in which the Hessian matrix $\nabla^2 f(\bar{x})$ exists but is singular.

Example 4.2. Let $f \colon \mathbb{R}^2 \to \mathbb{R}$ be defined by

$$f(x_1, x_2) = \tfrac{1}{2}(x_1)^2 + \tfrac{1}{4}(x_2)^4$$

and $\bar{x} = (0, 0)^T$. Then

$$\nabla^2 f(\bar{x}) = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \quad \text{and} \quad \partial^2 f(\bar{x}) = [-1, 1] \times \{0\}.$$

In this case

$$f^*(y_1, y_2) = \tfrac{1}{2}(y_1)^2 + \tfrac{3}{4}(y_2)^{4/3}$$

and for $\bar{y}=\nabla f(\bar{x})=(0,0)^T$ we have

$$(f^*)''(\bar{y};h)=\begin{cases}(h_1)^2 & \text{if } h_2=0,\\ +\infty & \text{otherwise,}\end{cases}$$

and

$$\partial^2 f^*(\bar{y})=[-1,1]\times\mathbb{R}=[\partial^2 f(\bar{x})]^0.$$

The purpose of this section is to explore the type of assumptions we should consider to ensure the validity of the polarity relationship (4.2). More generally, it is intended to obtain some estimates for the sets $\partial^2 f^*(\bar{y})$ and $\partial^2 f^*(\bar{y},\bar{x})$ in terms of $\partial^2 f(\bar{x})$ and $\partial^2 f(\bar{x},\bar{y})$, respectively. For this purpose it is convenient to introduce beforehand some definitions and to establish some useful lemmata.

Recall that $\bar{f}''(\bar{x},\bar{y};\cdot)$ is by definition the (upper) pointwise limit as $t\to 0^+$ of the sequence of functions $\{\varphi_t\}_{t>0}$ given by

$$\varphi_t(h):=\frac{2}{t}\left[\frac{f(\bar{x}+th)-f(\bar{x})}{t}-\langle\bar{y},h\rangle\right]\quad\forall h\in\mathbb{R}^n. \tag{4.3}$$

Some variants of this second-order directional derivative are provided by the expressions

$$\underline{f}''_e(\bar{x},\bar{y};\cdot):=\operatorname{epi}-\liminf_{t\to 0^+}\varphi_t,\qquad \bar{f}''_e(\bar{x},\bar{y};\cdot):=\operatorname{epi}-\limsup_{t\to 0^+}\varphi_t,$$

where the symbol e refers to the fact that the above lower and upper limits are taken in the "epigraphical" sense. These epilimits of the second-order difference quotient (4.3) were introduced by Rockafellar [R2] and used extensively by him [R3], [R4] and his students Poliquin [Po] and Do [D]. For the purpose of this section it suffices to mention that $\underline{f}''_e(\bar{x},\bar{y};\cdot):\mathbb{R}^n\to[0,\infty]$ is the lower-semicontinuous function given simply by

$$\underline{f}''_e(\bar{x},\bar{y};h)=\liminf_{\substack{t\to 0^+\\ h'\to h}}\varphi_t(h')\quad\forall h\in\mathbb{R}^n$$

and that we can write, in general, the inequalities

$$\underline{f}''_e(\bar{x},\bar{y};\cdot)\leqq\bar{f}''_e(\bar{x},\bar{y};\cdot)\leqq\operatorname{cl}\bar{f}''(\bar{x},\bar{y};\cdot).$$

Ndoutoume [Nd, Prop. 4.1] noticed that, in the same way as $\sqrt{\operatorname{cl}\bar{f}''(\bar{x},\bar{y};\cdot)}$, the function $\sqrt{\bar{f}''_e(\bar{x},\bar{y};\cdot)}$ is also a support function. Following Hiriart-Urruty [H3] and Seeger [Se], he introduced the following variation of the set $\partial^2 f(\bar{x},\bar{y})$:

$$\partial^2_e f(\bar{x},\bar{y})=\{z\in\mathbb{R}^n:\langle z,h\rangle\leqq\sqrt{\bar{f}''_e(\bar{x},\bar{y};h)},\ \forall h\in\mathbb{R}^n\}.$$

The later second-order subdifferential is a suitable tool for handling the problem we are considering in this section. However, the computation of the upper epilimit $\bar{f}''_e(\bar{x},\bar{y};\cdot)$ is usually a heavy task and this makes the set $\partial^2_e f(\bar{x},\bar{y})$ less useful than $\partial^2 f(\bar{x},\bar{y})$ for practical purposes. To eliminate the gap existing between these two sets and to put ourselves in a convenient framework, it is natural to introduce here the following definition.

DEFINITION 4.3. The function $f\in\Gamma_0(\mathbb{R}^n)$ is second-order regular at $\bar{x}$ relative to $\bar{y}\in\partial f(\bar{x})$ if

$$\underline{f}''_e(\bar{x},\bar{y};\cdot)=\operatorname{cl}\bar{f}''(\bar{x},\bar{y};\cdot). \tag{4.4}$$

If the above equality holds true for each $\bar{y}$ in $\partial f(\bar{x})$, then we say simply that f is second-order regular at $\bar{x}$.

Note that if the condition (4.4) holds, then the epilimit

$$f''_e(\bar{x}, \bar{y}; \cdot) = \text{epi} - \lim_{t \to 0^+} \varphi_t \tag{4.5}$$

exists and we can write

$$\psi^*[\cdot\,; \partial_e^2 f(\bar{x}, \bar{y})] = \sqrt{f''_e(\bar{x}, \bar{y}, \cdot)} = \sqrt{\text{cl}\, \bar{f}''(\bar{x}, \bar{y}, \cdot)} = \psi^*[\cdot\,; \partial^2 f(\bar{x}, \bar{y})]. \tag{4.6}$$

The mere existence of the epilimit (4.5) is a condition referred to by saying that f is twice epidifferentiable at $\bar{x}$ relative to $\bar{y}$ (cf. [R3, Def. 2.2]).

For a geometric interpretation of (4.4) and for examples of functions that are second-order regular at a given point $\bar{x}$, the reader can consult Hiriart-Urruty and Seeger [HS2, § 3.2]. Besides the case of a function f that is twice continuously differentiable on a neighborhood of $\bar{x}$, let us mention here two nontrivial examples.

Example 4.4. If the function $f \in \Gamma_0(\mathbb{R}^n)$ is differentiable and twice directionally differentiable at the point $\bar{x} \in \mathbb{R}^n$, then f is second-order regular at $\bar{x}$ relative to $\bar{y} = \nabla f(\bar{x})$. The proof of this fact is given in the Appendix. It involves the use of many technical tools that are of minor importance in this paper.

Example 4.5. Let $f \in \Gamma_0(\mathbb{R}^n)$ be a piecewise linear-quadratic function, i.e., its effective domain dom f can be expressed as the union of finitely many convex polyhedral sets and the restriction of f to each one of these sets is a quadratic (or affine) function. According to Rockafellar [R3, Thm. 3.1], for all $\bar{x} \in \text{dom} f$ and all $\bar{y} \in \partial f(\bar{x})$, we can write in this case

$$f''_e(\bar{x}, \bar{y}; \cdot) := \text{epi} - \lim_{t \to 0^+} \varphi_t = \lim_{t \to 0^+} \varphi_t =: f''(\bar{x}, \bar{y}; \cdot).$$

Thus, f is second-order regular at each point $\bar{x} \in \text{dom} f$.

Let us now establish the next result.

LEMMA 4.6. *Let $f, f^* \in \Gamma_0(\mathbb{R}^n)$ be a couple of conjugate functions and let $(\bar{x}, \bar{y}) \in G$ (cf. Proposition* 2.1). *Then*

(a) *f is twice epidifferentiable at $\bar{x}$ relative to $\bar{y}$ if and only if f^* is twice epidifferentiable at $\bar{y}$ relative to $\bar{x}$.*

(b) *Under the equivalent conditions stated in* (a), *we can write*

$$\tfrac{1}{2}(f^*)''_e(\bar{y}, \bar{x}; d) = [\tfrac{1}{2} f''_e(\bar{x}, \bar{y}; \cdot)]^*(d) \quad \forall d \in \mathbb{R}^n \tag{4.7}$$

and

$$\partial_e^2 f^*(\bar{y}, \bar{x}) = [\partial_e^2 f(\bar{x}, \bar{y})]^0. \tag{4.8}$$

Proof. The proof of Lemma 4.6(a) and the conjugacy formula (4.7) can be found in Do [D, Thm. 1.2.7]. Let us prove, then, the polarity formula (4.8). The nonnegative function

$$l = \tfrac{1}{2} f''_e(\bar{x}, \bar{y}; \cdot)$$

is lower-semicontinuous, convex, positively homogeneous of degree 2, and vanishes at 0. According to [R1, Cor. 15.3.2] the sets

$$C = \{h \in \mathbb{R}^n : \sqrt{2l(h)} \leqq 1\}$$

and

$$D = \{d \in \mathbb{R}^n : \sqrt{2l^*(d)} \leqq 1\}$$

are polar to each other. Taking into account the definition of l and the formula (4.7), it is clear that

$$\sqrt{2l(h)}=\sqrt{f''_e(\bar{x},\bar{y};h)}=\psi^*[h;\partial^2_e f(\bar{x},\bar{y})] \quad \forall h\in\mathbb{R}^n$$

and

$$\sqrt{2l^*(d)}=\sqrt{(f^*)''_e(\bar{y},\bar{x};d)}=\psi^*[d;\partial^2_e f^*(\bar{y},\bar{x})] \quad \forall d\in\mathbb{R}^n.$$

Therefore

$$C=[\partial^2_e f(\bar{x},\bar{y})]^0$$

coincides with

$$D^0=\{[\partial^2_e f^*(\bar{y},\bar{x})]^0\}^0=\partial^2_e f^*(\bar{y},\bar{x})$$

as we wanted to prove. □

Taking into account the previous lemma it is easy to now prove the following generalization of Crouzeix's theorem.

THEOREM 4.7. *Let $f,f^*\in\Gamma_0(\mathbb{R}^n)$ be a couple of conjugate functions and let $(\bar{x},\bar{y})\in G$. Assume that either*

(a) *f is second-order regular at $\bar{x}$ relative to $\bar{y}$*

or

(b) *f^* is second-order regular at $\bar{y}$ relative to $\bar{x}$.*

Then the inclusion

$$\partial^2 f^*(\bar{y},\bar{x})\supset[\partial^2 f(\bar{x},\bar{y})]^0 \tag{4.9}$$

holds true. Furthermore, we have the equality

$$\partial^2 f^*(\bar{y},\bar{x})=[\partial^2 f(\bar{x},\bar{y})]^0 \tag{4.10}$$

if both conditions (a) *and* (b) *are satisfied.*

Proof. It is clear that we always have the inclusions

$$\partial^2_e f(\bar{x},\bar{y})\subset\partial^2 f(\bar{x},\bar{y})$$

and

$$\partial^2_e f^*(\bar{y},\bar{x})\subset\partial^2 f^*(\bar{y},\bar{x}),$$

which follow immediately from the very definition of the above sets. If either (a) or (b) of Theorem 4.7 is true, then the first or, respectively, the second inclusion becomes an equality. Moreover, the equivalent conditions stated in Lemma 4.6(a) are satisfied. The formula (4.8) can be applied in such a case and yields the inclusion (4.9). If (a) and (b) of Theorem 4.7 are true, then (4.8) reduces, of course, to the desired equality (4.10). □

Theorem 4.7(a) and (b) are not equivalent and therefore the set $\partial^2 f^*(\bar{y},\bar{x})$ could include $[\partial^2 f(\bar{x},\bar{y})]^0$ strictly. Let us illustrate this fact with an example.

Example 4.8. Let f and f^* be, respectively, the indicator and the support function of the (Euclidean) closed unit ball B of $\mathbb{R}^n$. We intentionally choose $\bar{x}$ in the boundary of dom f, i.e., satisfying $\|\bar{x}\|=1$. We then have $\partial f(\bar{x})=\{\lambda\bar{x}:\lambda\geqq 0\}$. Let us now choose $\bar{y}=\lambda\bar{x}$ with $\lambda>0$. Then f^* is twice continuously differentiable on a neighborhood of $\bar{y}$ (therefore Theorem 4.7(b) holds), but f is not second-order regular at $\bar{x}$ relative to $\bar{y}$. As shown by Seeger [Se, Ex. B.2.70], in this case the set

$$\partial^2 f(\bar{x},\bar{y})^0=(\mathbb{R}^n)^0=\{0\}$$

is strictly included in

$$\partial^2 f^*(\bar{y}, \bar{x}) = \partial^2 f^*(\bar{y}) = \text{ellipsoid associated to } \nabla^2 f^*(\bar{y}).$$

The theorem just established is not only a slight generalization of Crouzeix's result, but also a considerable one. Not only is the nonsingularity assumption made on $\nabla^2 f(\bar{x})$ removed, but we are also allowed to consider a function $f \in \Gamma_0(\mathbb{R}^n)$ that is not even differentiable at $\bar{x}$. We do not need to illustrate this point with further examples. Instead, we will establish some important corollaries.

COROLLARY 4.9. *Suppose that the function f^* is differentiable and twice directionally differentiable at the point $\bar{y} \in \mathbb{R}^n$. Assume also that $f \in \Gamma_0(\mathbb{R}^n)$ is second-order regular at $\bar{x} = \nabla f^*(\bar{y})$ relative to $\bar{y}$. Then*

$$\partial^2 f^*(\bar{y}) = [\partial^2 f(\bar{x}, \bar{y})]^0.$$

Proof. The proof is immediate. The fact that f^* is second-order regular at $\bar{y}$ relative to $\bar{x} = \nabla f^*(\bar{y})$ is proved in the Appendix. □

For the sake of completeness let us write the analogous version of Corollary 4.9, which is obtained by exchanging the roles of f and f^*.

COROLLARY 4.10. *Suppose that the function $f \in \Gamma_0(\mathbb{R}^n)$ is differentiable and twice directionally differentiable at the point $\bar{x} \in \mathbb{R}^n$. Assume also that f^* is second-order regular at $\bar{y} = \nabla f(\bar{x})$ relative to $\bar{x}$. Then*

$$\partial^2 f^*(\bar{y}, \bar{x}) = [\partial^2 f(\bar{x})]^0.$$

In the next corollary no differentiability assumption is made either on f or on f^*. In his recent dissertation Sun [Su] demonstrated that a function $f \in \Gamma_0(\mathbb{R}^n)$ is piecewise linear quadratic if and only if its conjugate $f^* \in \Gamma_0(\mathbb{R}^n)$ is piecewise linear quadratic.

COROLLARY 4.11. *Let $f, f^* \in \Gamma_0(\mathbb{R}^n)$ be a couple of conjugate functions that are piecewise linear quadratic and let $(\bar{x}, \bar{y}) \in G$. Then*

$$\partial^2 f^*(\bar{y}, \bar{x}) = [\partial^2 f(\bar{x}, \bar{y})]^0.$$

Proof. The proof is immediate. See Example 4.5. □

Let us end this section by giving an estimate of the second-order subdifferential of the conjugate f^* at a given point $\bar{y}$. Again, no differentiability assumption is made either on f or on f^*.

COROLLARY 4.12. *Assume that f^* is second-order regular at $\bar{y} \in \mathbb{R}^n$ and that $f \in \Gamma_0(\mathbb{R}^n)$ is second-order regular at each point in the nonempty set*

$$\partial f^*(\bar{y}) = \{\bar{x} \in \mathbb{R}^n \colon \bar{y} \in \partial f(\bar{x})\}.$$

Then

$$\partial^2 f^*(\bar{y}) = \left[\bigsqcup_{\bar{x} \in \partial f^*(\bar{y})} \partial^2 f(\bar{x}, \bar{y}) \right]^0. \tag{4.11}$$

Proof. Under the hypotheses of this corollary, the set

$$\partial^2 f^*(\bar{y}) := \bigcap \{\partial^2 f^*(\bar{y}, \bar{x}) \colon \bar{x} \in \partial f^*(\bar{y})\}$$

can be written as

$$\partial^2 f^*(\bar{y}) = \bigcap \{[\partial^2 f(\bar{x}, \bar{y})]^0 \colon \bar{x} \in \partial f^*(\bar{y})\}.$$

The equality (4.11) is then obtained by using the general calculus rule

$$\left[\bigcup_{\alpha} C_\alpha \right]^0 = \bigcap_{\alpha} C_\alpha^0$$

(cf. [Ma, p. 84]). □

Appendix. The demonstration of the next result is rather technical. For this reason it has been postponed until the Appendix.

LEMMA A. *Suppose that the function $f \in \Gamma_0(\mathbb{R}^n)$ is differentiable at $\bar{x}$ and that, for all $h \in \mathbb{R}^n$, the limit*

$$f''(\bar{x};\, h) = \lim_{t \to 0^+} \frac{2}{t}\left[\frac{f(\bar{x}+th)-f(\bar{x})}{t} - \langle \nabla f(\bar{x}),\, h\rangle\right]$$

exists and is finite. Then, for all $h \in \mathbb{R}^n$, we can write

$$f''(\bar{x};\, h) = \lim_{\substack{t \to 0^+ \\ h' \to h}} \frac{2}{t}\left[\frac{f(\bar{x}+th')-f(\bar{x})}{t} - \langle \nabla f(\bar{x}),\, h'\rangle\right]. \tag{A.1}$$

Proof. Under the hypotheses of this lemma the second-order directional derivative $f''(\bar{x};\, \cdot)$ admits also the characterizations

$$f''(\bar{x};\, h) = (\psi^*[h;\, \partial^2 f(\bar{x})])^2 \quad \forall h \in \mathbb{R}^n \tag{A.2}$$

and

$$f''(\bar{x};\, \cdot) = \lim_{\varepsilon \to 0^+} \frac{[f'_\varepsilon(\bar{x};\, h) - \langle \nabla f(\bar{x}),\, h\rangle]^2}{2\varepsilon} \quad \forall h \in \mathbb{R}^n. \tag{A.3}$$

The first equality is immediate from Proposition 4.1. The second equality was established first by Hiriart-Urruty [H2], at least for those h for which $f''(\bar{x};\, h) > 0$. A proof of the case remaining can be found in [Se, Prop. B.4.3] or [HS2, Thm. 4.1]. Equality (A.3) asks for an explanation. For us it suffices to know that the ε-directional derivative $f'_\varepsilon(\bar{x};\, \cdot)$ of f at $\bar{x}$ coincides with the support function of a set denoted by $\partial_\varepsilon f(\bar{x})$ and called the ε-subdifferential of f at $\bar{x}$. Here $\partial_\varepsilon f(\bar{x})$ is for all $\varepsilon \geqq 0$ a nonempty convex compact set in $\mathbb{R}^n$. What (A.3) says, then, is that $\sqrt{f''(\bar{x};\, \cdot)}$ is a pointwise limit of a sequence of support functions. We have, more precisely,

$$\psi^*[h;\, \partial^2 f(\bar{x})] = \sqrt{f''(\bar{x};\, h)} = \lim_{\varepsilon \to 0^+} \psi^*[h;\, A_\varepsilon] \quad \forall h \in \mathbb{R}^n,$$

where

$$A_\varepsilon := \frac{\partial_\varepsilon f(\bar{x}) - \nabla f(\bar{x})}{\sqrt{2\varepsilon}}.$$

Since, in this case, the sets $\partial^2 f(\bar{x})$ and A_ε are compact, the convergence of $\psi^*[\cdot\,;\, A_\varepsilon]$ towards $\psi^*[\cdot\,;\, \partial^2 f(\bar{x})]$ as $\varepsilon \to 0^+$ is not only pointwise, but it is also uniform in the sense that

$$\psi^*[h;\, \partial^2 f(\bar{x})] = \lim_{\substack{\varepsilon \to 0^+ \\ h' \to h}} \psi^*[h';\, A_\varepsilon] \quad \forall h \in \mathbb{R}^n.$$

Hence we can write

$$f''(\bar{x};\, h) = \lim_{\substack{\varepsilon \to 0^+ \\ h' \to h}} \frac{[f'_\varepsilon(\bar{x};\, h') - \langle \nabla f(\bar{x}),\, h'\rangle]^2}{2\varepsilon} \quad \forall h \in \mathbb{R}^n. \tag{A.4}$$

Now, following step-by-step the proof given by Lemarechal and Zowe [LZ, Thm. 2.1] for equality (A.3), we demonstrate that the existence of the limit on the right-hand side of (A.4) implies the existence of the limit on the right-hand side of (A.1). This proves, of course, the result announced in the present lemma. □

REFERENCES

[At] H. Attouch, *Variational convergence for functions and operators*, Pitman Research Notes in Mathematics, Pitman, London, 1984.

[Az] R. Azencott, *Grandes deviations et applications*, in Lecture Notes in Mathematics 774, Ecole d'Été de Probabilités de Saint-Flour, P. L. Hennequin, ed., Springer-Verlag, Berlin, New York, 1980.

[B-N] O. Barndorff-Nielsen, *Exponential families: Exact theory*, Various Publication Series 19, Institute of Mathematics, University of Aarhus, Aarhus, Denmark, 1970.

[CY] S. Y. Cheng and S. T. Yau, *On the regularity of the Monge-Ampère equation* det $(\partial^2 u/\partial x_i\, \partial x_j) = F(x, y)$, Comm. Pure Appl. Math., 30 (1977), pp. 41-68.

[Cr] J. P. Crouzeix, *A relationship between the second derivative of a convex function and of its conjugate*, Math. Programming, 13 (1977), pp. 364-365.

[D] C. H. Do, *Second-order nonsmooth analysis and sensitivity in optimization problems involving convex integral functionals*, Ph.D. thesis, Department of Mathematics, University of Washington, Seattle, WA, 1989.

[E1] R. Ellis, *Entropy, large deviations and statistical mechanics*, in Grundlehren der Mathematischen Wissenschaften 271, Springer-Verlag, Berlin, New York, 1985.

[E2] ———, *Large deviations and statistical mechanics*, in Contemporary Mathematics Series 41, American Mathematical Society, Providence, RI, 1985, pp. 101-123.

[Fu] P. Funk, *Variationsrechnung und ihre Anwendung in Physik und Technik*, Grundlehren der Mathematischen Wissenschaften 94, Springer-Verlag, Berlin, Göttingen, Heidelberg, 1962.

[GF] J. H. Gelfand and S. V. Fomin, *Calculus of Variations*, Prentice-Hall, Englewood Cliffs, NJ, 1963.

[HR] R. Hill and J. R. Rice, *Elastic potentials and the structure of inelastic constitutive laws*, SIAM J. Appl. Math., 25 (1973), pp. 448-461.

[H1] J. B. Hiriart-Urruty, *Contributions à la programmation mathématique: cas déterministe et stochastique*, Ph.D. thesis, Department of Mathematics, University of Clermont-Ferrand II, France, 1977.

[H2] ———, *Limiting behaviour of the approximate first order and second order directional derivatives for a convex function*, Nonlinear Anal. Theory Math. Appl., 6 (1982), pp. 1309-1326.

[H3] ———, *A new set-valued second-order derivative for convex functions*, in Mathematics for Optimization, Mathematical Studies Series 129, North-Holland, Amsterdam, 1986.

[HS1] J. B. Hiriart-Urruty and A. Seeger, *Règles de calcul sur le sous-différential second d'une fonction convexe*, C. R. Acad. Sci. Paris, 304 (1987), pp. 259-262.

[HS2] ———, *The second-order subdifferential and the Dupin indicatrices of a nondifferentiable convex function*, Proc. London Math. Soc., 58 (1989), pp. 351-365.

[Ho] A. S. Householder, *The theory of matrices in numerical analysis*, Blaisdell, New York, 1964.

[Ir] M. Iri, *Network Flows, Transportation and Scheduling: Theory and Algorithms*, Academic Press, New York, 1969.

[Ki] C. O. Kiselman, *Sur la définition de l'opérateur de Monge-Ampère complexe*, in Proc. Journées Fermat on Complex Analysis, Toulouse, 1983, Lecture Notes in Mathematics 1094, Springer-Verlag, Berlin, 1984, pp. 139-150.

[Ku] S. S. Kutateladze, *Convex operators*, Russian Math. Surveys, 34 (1979), pp. 181-224.

[La] J. P. Laurent, *Approximation et optimisation*, Hermann, Paris, 1972.

[LZ] C. Lemarechal and J. Zowe, *Some remarks on the construction of higher order algorithms in convex optimization*, Appl. Math. Optim., 10 (1983), pp. 51-68.

[Li] P. Linz, *Theoretical numerical analysis*, in Pure and Applied Mathematics Series, John Wiley, New York, 1979.

[Ma] G. Marinescu, *Tratat de analiză functională* II, Editura Academiei, Bucharest, 1972.

[Mo] J. J. Moreau, *Fonctionnelles convexes*, Séminaire sur les équations aux dérivées partielles, Collège de France, Paris, 1967.

[Nd] J. L. Ndoutoume, *Calcul differentiel du second ordre*, in Publications AVAMAC, Université de Perpignan, Perpignan, France, 1987.

[P] A. V. Pogorelov, *The Minkowski multidimensional problem*, Scripta Series in Mathematics, V. H. Winston, Washington, D.C., 1978.

[Po] R. Poliquin, *Proto-differentiation and integration of proximal subgradients*, Ph.D. thesis, Department of Mathematics, University of Washington, Seattle, WA, 1988.

[Ra] N. A. Rahman, *A Course in Theoretical Statistics*, Hafner, New York, 1968.

[RT] J. Rauch and B. A. Taylor, *The Dirichlet problem for the multidimensional Monge-Ampère equation*, Rocky Mountain J. Math., 7 (1977), pp. 345-364.

[R1] R. T. Rockafellar, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.

[R2] R. T. Rockafellar, *Maximal monotone relations and the second derivatives of nonsmooth functions*, Ann. Inst. H. Poincaré, 2 (1985), pp. 167–184.

[R3] ———, *First- and second-order epi-differentiability in nonlinear programming*, Trans. Amer. Math. Soc., 307 (1988), pp. 75–108.

[R4] ———, *Second-order optimality conditions in nonlinear programming obtained by way of epi-derivatives*, Math. Oper. Res., 14 (1989), pp. 462–484.

[Ru] H. Rund, *The Hamilton-Jacobi Theory in the Calculus of Variations*, Van Nostrand, London, 1966.

[Se] A. Seeger, *Analyse du second-ordre de problèmes non différentiables*, Ph.D. thesis, Université Paul Sabatier, Toulouse, France, 1986.

[S1] M. J. Sewell, *On Legendre transformations and elementary catastrophes*, Math. Proc. Camb. Philos. Soc., 82 (1977), pp. 147–163.

[S2] ———, *On Legendre transformations and umbilic catastrophes*, Math. Proc. Camb. Philos. Soc., 83 (1978), pp. 273–288.

[Su] J. Sun, *On monotropic piecewise quadratic programming*, Ph.D. thesis, Department of Applied Mathematics, University of Washington, Seattle, WA, 1986.

SIAM J. OPTIMIZATION
Vol. 2, No. 3, pp. 425–434, August 1992

005

ON THE CONVERGENCE OF THE PRODUCTS OF FIRMLY NONEXPANSIVE MAPPINGS*

PAUL TSENG†

Abstract. Consider a finite collection of firmly nonexpansive self-mappings on a Hilbert space whose fixed-point sets intersect. It is shown that, in the finite-dimensional case, any iteration of mappings drawn from this collection converges. This resolves, for the finite-dimensional case at least, a popular conjecture concerning the convergence of the successive projection method. In the infinite-dimensional case, it is shown that if the mappings are drawn according to a certain order, called the quasi-cyclic order, then the iteration converges weakly in a sense. The quasi-cyclic order may be viewed as an extension of the well-known cyclic order in which the lengths of the cycles are permitted to grow without bound.

Key words. firmly nonexpansive mapping, successive projection, convex set, quasi-cyclic order

AMS(MOS) subject classifications. 49, 90

1. Introduction.

Let $\mathcal{H}$ be some real Hilbert space endowed with an inner product $\langle\cdot,\cdot\rangle$, and let $T_i : \mathcal{H} \to \mathcal{H}$, $i = 1, \cdots, m$ $(m \geq 1)$ be a collection of *firmly nonexpansive* mappings, i.e., for each $x, y \in \mathcal{H}$,

$$\langle T_i(x) - T_i(y), x - y\rangle \geq ||T_i(x) - T_i(y)||^2, \tag{1}$$

whose fixed-point sets, denoted by $F_1, \cdots, F_m$, respectively, make a nonempty intersection. Here $||\cdot||$ is the norm induced by the inner product $\langle\cdot,\cdot\rangle$ (i.e., $||x|| = \sqrt{\langle x, x\rangle}$). Our problem is to find a common fixed point of the T_i's, that is,

$$\text{find a point in } F = \cap_{i=1}^m F_i. \tag{P}$$

Firmly nonexpansive mappings are nonexpansive and are closely related to the notion of maximal monotone operators. In particular, we can take each T_i to be the *resolvent* of some maximal monotone operator A_i in $\mathcal{H}$ (see [Roc76]); i.e.,

$$T_i = (I + A_i)^{-1}. \tag{2}$$

In this case, the problem (P) reduces to finding a common zero for the A_i's. (See [Bré73] for general discussions of maximal monotone operators. See [Eck89, Chap. 3] for a detailed discussion of firmly nonexpansive mappings and their relation to maximal monotone operators.)

Another way to get a firmly nonexpansive mapping T_i is to take

$$T_i = \tfrac{1}{2}(I + S_i),$$

where $S_i : \mathcal{H} \to \mathcal{H}$ is *any* nonexpansive mapping. (The converse of this in fact also holds. See, for example, [Roc76].) In this case (P) reduces to finding a common fixed

*Received by the editors January 10, 1990; accepted for publication (in revised form) September 9, 1991. This work was supported by the U.S. Army Research Office contract DAAL03-86-K-0171 (Center for Intelligent Control Systems).

†Department of Mathematics, GN-50, University of Washington, Seattle, Washington 98195.

point of the S_i's. (See [Bro76] and [Sin83] for surveys of results on nonexpansive mappings.)

Consider the following iterative method for solving (P). We begin with an arbitrary $x(0) \in \mathcal{H}$. At the tth iteration ($t \geq 0$), we are given an $x(t) \in \mathcal{H}$; we choose an index $\sigma(t) \in \{1, 2, \cdots, m\}$ and apply $T_{\sigma(t)}$ to $x(t)$ to obtain a point $y(t)$; i.e.,

$$y(t) = T_{\sigma(t)}(x(t)). \tag{3a}$$

Then we move $x(t)$ in the direction $y(t) - x(t)$ with a step size of $\omega(t)$ to obtain the new iterate $x(t+1)$; i.e.,

$$x(t+1) = \omega(t)y(t) + (1 - \omega(t))x(t). \tag{3b}$$

To ensure convergence of the iterates, we impose the standard restriction that the step sizes are bounded inside $(0, 2)$; that is,

$$\epsilon \leq \omega(t) \leq 2 - \epsilon \quad \forall t, \tag{3c}$$

where ϵ is any fixed scalar in $(0, 1]$.

In the case where $m = 1$, it readily follows from a result of Opial [Opi67] on *asymptotically regular* mappings (of which firmly nonexpansive mappings are special cases) that the sequence $\{x(t)\}$ generated by (3a)–(3c) converges weakly to an element of F (also see [Roc76, §1]). If $m = 2$, it is but a slight modification of Opial's proof to show that, again, $\{x(t)\}$ converges weakly (although not necessarily to an element of F). But what about $m \geq 3$? Does $\{x(t)\}$ still converge weakly then?

The above question is motivated by a long-standing question about the convergence of the *successive projection* method. Let $C_1, C_2, \cdots, C_m$ be closed convex sets in $\mathcal{H}$ with nonempty intersection. Consider the following classical problem in optimization:

$$\text{find a point in } \cap_{i=1}^{m} C_i. \tag{I}$$

This problem is a special case of (P) in which each T_i is taken to be the projection (or proximity) mapping onto C_i; i.e.,

$$T_i(x) = \arg\min_{y \in C_i} ||x - y||. \tag{4}$$

That T_i given by (4) is firmly nonexpansive readily follows from the properties of the projection mapping (see [Bre65] and [GPR67]). In fact, T_i is of the form (2) with A_i being the subdifferential of the indicator function for C_i (see [Roc76]). With T_i given by (4), the method (3a)–(3c) becomes what is commonly known as the successive projection (SP) method for solving (I).

The SP method was first proposed by Kaczmarz [Kac37] for the special case in which the sets $C_1, \cdots, C_m$ are linear varieties (i.e., translates of subspaces) and was rediscovered by von Neumann [voN50], Agmon [Agm54], and Motzkin and Schoenberg [MoS54]. (Also see [AmA65], [Deu85], [Gof80], [Gof82], [Hal62], [Man84], [Mer62], [Pra60], [SSW77], and [Tan71] for more detailed treatments of the linear case.) Extensions of the SP method to problems with general convex sets were made by Bregman [Bre65] and Eremin [Ere65] (also see [Bru82], [ChG59], [Ere66], [GeS66], [GPR67], [Ott88], [Pol69], [You87], and [You89]). The SP method can also be applied to problems in a product space to obtain a highly parallelizable method of barycenters [Pie84].

(There has also been some work analyzing the case in which the number of sets is *infinite* [Bre65], [Bru83], [GPR67] or where the C_i's do not intersect [GPR67].) The notion of a step size (also called *relaxation parameter*) $\omega(t)$ [cf. (3c)] was introduced in [Agm54] and [MoS54]. It has been observed that, in certain cases, a value of $\omega(t)$ different from 1 (i.e., under- or over-relaxation) can significantly improve the convergence (see [Gof80], [Her80], [LeS83], and [Man84]).

An important question associated with the SP method concerns the convergence of the iterates generated by it. In most of the analyses it is assumed that the sets are chosen either in an essentially cyclic order (i.e., every set is chosen at least once every B iterations, for some fixed $B \geq m$) or according to a maximal distance rule (i.e., choose a set that is in some sense farthest away from the current iterate). Under such assumptions, one can prove weak convergence of the iterates [Bre65]; if, in addition, either the C_i's are linear varieties or a certain regularity condition holds, then one can also prove strong convergence [GPR67], [Hal62], [Ott88], [VoN50]. (A related question concerns the *rate* of convergence of the iterates. However, rate of convergence analysis requires fairly restrictive assumptions, such as that C_i's are halfspaces [Agm54], [Gof80], [Gof82], [GPR67], [Man84], [Mer62], [SSW77] or that a certain regularity condition holds [GPR67]. It also requires the order of projections to be restricted to either essentially cyclic or one given by a maximal distance rule. *Finite* convergence of the iterates can also be proven under more restrictive assumptions [MoS54], [Gof80], [Gof82], [GPR67].)

Can the above assumptions on the order of projections be weakened? In particular, if no assumption on the order of projections is made, would the iterates converge weakly? The answer was shown by Prager [Pra60] to be "yes" if $\mathcal{H}$ is finite-dimensional and the C_i's are linear subspaces of $\mathcal{H}$. Prager's result was later extended by Amemiya and Ando [AmA65] to the case where the C_i's are closed linear subspaces. Bruck [Bru82] proved a similar result assuming that $m = 3$ and $C_i = -C_i$ for all i. (Strong convergence can also be shown if one of the C_i's is compact [Bru82], [You87, p. 74] or if a certain regularity condition holds [Ott88].) The most general result in this direction is perhaps that given by Youla [You89] (also see [You87, §2.6]), which proves weak convergence assuming only that the C_i's share an "inner" point. (An $x \in C_i$ is said to be inner if it is the *only* point in the linear manifold spanned by C_i whose projection onto C_i is x.) In a finite-dimensional space, this amounts to the regularity condition that the relative interior of the sets $C_1, \cdots, C_m$ make a nonempty intersection. Although such a regularity condition is fairly mild, nonetheless there are problem instances for which it fails to hold. For example, when a primal-dual pair of linear programs is formulated as a feasibility problem of the form (P) (see [Bre65] and [Sch86, p. 125]), the relative interior of the corresponding C_i's typically do not intersect.

The contribution of this article is threefold: First, we prove that when $\mathcal{H}$ is finite-dimensional, no assumption is needed to guarantee convergence of the SP method (see Corollary 1). This resolves, in the finite-dimensional case at least, a conjecture of Bruck [Bru83, p. 37]. Second, we prove a weak convergence result under a new order of projections, called *quasi-cyclic order*. The quasi-cyclic order may be viewed as an extension of the essentially cyclic order in which the lengths of the cycles, namely the B given previously, are allowed to increase without bound, but not too fast. Although this order is clearly more restrictive than having no restriction on the order of projections at all, it has the advantage that no additional assumption on the problem is needed to obtain weak convergence. Third, we prove the above results in the more general context of the products of firmly nonexpansive mappings, not just projection mappings, and obtain the results for the SP method as simple consequences of these results.

2. Convergence analysis. Before we prove our main results, we need the following known technical lemma.

LEMMA 1. *Let* $\{x(t)\}$ *be a sequence generated by* (3a)–(3c). *Then,*

$$\sum_{t=0}^{\infty} ||x(t+1)-x(t)||^2 < \infty, \tag{5}$$

$$\{||y(t)-x(t)||\} \to 0. \tag{6}$$

Moreover, for any $\bar{x} \in F$, *there holds*

$$||\bar{x}-x(t)|| \geq ||\bar{x}-x(t+1)|| \quad \forall t \geq 0, \tag{7}$$

$$\theta(\bar{x}) = \lim_{t\to\infty} ||\bar{x}-y(t)||^2, \tag{8}$$

where we let

$$\theta(\bar{x}) = \lim_{t\to\infty} ||\bar{x}-x(t)||^2. \tag{9}$$

Proof. Fix any $\bar{x} \in F$. For any integer $t \geq 0$, we have $y(t) = T_{\sigma(t)}(x(t))$ [cf. (3a)] and $\bar{x} = T_{\sigma(t)}(\bar{x})$ [cf. $\bar{x} \in F$], so the firmly nonexpansive property of $T_{\sigma(t)}$ [cf. (1)] yields

$$\langle \bar{x}-y(t), \bar{x}-x(t)\rangle \geq ||\bar{x}-y(t)||^2$$

or, equivalently,

$$\langle \bar{x}-y(t), y(t)-x(t)\rangle \geq 0.$$

Since [cf. (3b)]

$$\begin{aligned}||\bar{x}-x(t)||^2 =&||\bar{x}-x(t+1)||^2 + 2\omega(t)\langle \bar{x}-y(t), y(t)-x(t)\rangle \\ &+ \omega(t)(2-\omega(t))||y(t)-x(t)||^2,\end{aligned}$$

this, together with (3c), implies

$$||\bar{x}-x(t)||^2 \geq ||\bar{x}-x(t+1)||^2 + \epsilon^2||x(t+1)-x(t)||^2 \quad \forall t \geq 0,$$

so (5) and (7) follow. Equation (7) implies that $\theta(\bar{x})$ given by (9) is well defined, and (5) implies that

$$\{||x(t+1)-x(t)||\} \to 0.$$

Since $y(t)-x(t) = (x(t+1)-x(t))/\omega(t)$ [cf. (3b)] and $\omega(t) \geq \epsilon$ [cf. (3c)] for all t, then the latter proves (6), which together with (9) proves (8). □

By using Lemma 1, we can show our first main result.

THEOREM 1. *Suppose that* $\mathcal{H}$ *is finite-dimensional. Let* $\{x(t)\}$ *be a sequence generated by* (3a)–(3c) *under the assumption that each element of* $\{1, 2, \cdots, m\}$ *appears in the sequence* $\{\sigma(0), \sigma(1), \cdots\}$ *an infinite number of times. Then* $\{x(t)\}$ *converges to a point in* F.

Proof. Equation (8) shows that $\{y(t)\}$ is bounded, so it has a cluster point in $\mathcal{H}$. (Recall that in a finite-dimensional space, weak convergence is equivalent to ordinary convergence.) Let Y^∞ denote the set of cluster points of $\{y(t)\}$.

We claim that $Y^\infty \cap F \neq \emptyset$. We argue this by contradiction. Suppose that $Y^\infty \cap F = \emptyset$. Fix any $\bar{x} \in F$ (so $\bar{x} \notin Y^\infty$) and any $y^\infty \in Y^\infty$. Let N be any subsequence of $\{0, 1, \cdots\}$ such that

$$\{y(t)\}_{t\in N} \text{ converges to } y^\infty. \tag{10}$$

By passing into a subsequence if necessary, we can assume that, for some i, $\sigma(t) = i$ for all $t \in N$. Then $y(t) = T_i(x(t))$ for all $t \in N$ [cf. (3a)], so (6), (10), and the continuity property of T_i imply $y^\infty = T_i(y^\infty)$ or, equivalently, $y^\infty \in F_i$. Since $y^\infty \notin F$, then, by reindexing the F_i's if necessary, we can assume that for some $\bar{i} \in \{1, \cdots, m-1\}$, there holds

$$y^\infty \in F_1 \cap \cdots \cap F_{\bar{i}}, \qquad y^\infty \notin F_i \quad \forall i > \bar{i}. \tag{11}$$

For each $t \in N$, let $\Delta(t)$ be the smallest integer $\tau \geq t$ such that $\sigma(\tau) > \bar{i}$. $\Delta(t)$ is well defined for all $t \in N$ because, by hypothesis, each element of $\{\bar{i}+1, \bar{i}+2, \cdots, m\}$ appears in $\{\sigma(0), \sigma(1), \cdots\}$ an infinite number of times. Notice that $\Delta(t)$ is monotonically increasing with t and tends to ∞ as $t \to \infty$. More importantly, we have, by the construction of $\Delta(t)$, that $\sigma(\tau) \leq \bar{i}$ for all $\tau \in \{t, t+1, \cdots, \Delta(t) - 1\}$. Since $y^\infty \in F_1 \cap \cdots \cap F_{\bar{i}}$ (cf. (11)), then an argument analogous to the proof of (7) (with t, $\bar{x}$, and F therein replaced by, respectively, τ, y^∞, and $F_1 \cap \cdots \cap F_{\bar{i}}$) yields

$$||y^\infty - x(\tau)|| \geq ||y^\infty - x(\tau + 1)||, \qquad \tau = t, t+1, \cdots, \Delta(t) - 1,$$

so that

$$||y^\infty - x(t)|| \geq ||y^\infty - x(\Delta(t))||. \tag{12}$$

Since our choice of $t \in N$ was arbitrary, then (12) holds for all $t \in N$. Also, since $\{y(\Delta(t))\}_{t\in N}$ is bounded (cf. (8)), it has some cluster point $\hat{y}^\infty$ (so $\hat{y}^\infty \in Y^\infty$). By reindexing the F_i's and further passing into a subsequence if necessary, we can assume that

$$\sigma(\Delta(t)) = \bar{i} + 1 \quad \forall t \in N \tag{13}$$

and

$$\{y(\Delta(t))\}_{t\in N} \text{ converges to } \hat{y}^\infty. \tag{14}$$

Equation (13), together with (3a), implies $y(\Delta(t)) = T_{\bar{i}+1}(x(\Delta(t)))$ for all $t \in N$. Since $T_{\bar{i}+1}$ is continuous, this, together with (6) and (14), implies $\hat{y}^\infty = T_{\bar{i}+1}(\hat{y}^\infty)$ or, equivalently, $\hat{y}^\infty \in F_{\bar{i}+1}$. Hence $\hat{y}^\infty \neq y^\infty$ (cf. (11)). Also, we have

$$\begin{aligned} ||y^\infty - y(\Delta(t))||^2 =& ||y^\infty - \hat{y}^\infty||^2 + 2\langle y^\infty - \hat{y}^\infty, \hat{y}^\infty - y(\Delta(t))\rangle \\ &+ ||\hat{y}^\infty - y(\Delta(t))||^2, \end{aligned}$$

for all $t \in N$. Upon passing into the limit as $t \to \infty$, $t \in N$ and by using (14), we obtain

$$\lim_{t\to\infty, t\in N} \sup ||y^\infty - y(\Delta(t))||^2 \geq ||y^\infty - \hat{y}^\infty||^2.$$

By using (6) and (12) to upper bound the left-hand limit in the above relation, we obtain

$$\lim_{t\to\infty, t\in N} ||y^\infty - y(t)||^2 \geq ||y^\infty - \hat{y}^\infty||^2.$$

By (10) the left-hand limit vanishes, so we are left with

$$0 \geq ||y^\infty - \hat{y}^\infty||^2,$$

a contradiction of the fact $y^\infty \neq \hat{y}^\infty$ shown earlier.

We now show that $\{x(t)\}$ has a unique cluster point in F. Since $Y^\infty \cap F \neq \emptyset$, then we can find an $\bar{x} \in F$ and a subsequence N of $\{0, 1, \cdots\}$ such that $\{y(t)\}_{t\in N}$ converges to $\bar{x}$. This implies $\{||\bar{x} - y(t)||\}_{t\in N} \to 0$, so (8) yields $\theta(\bar{x}) = 0$. By (9), the entire sequence $\{x(t)\}$ converges to $\bar{x}$. □

COROLLARY 1. *Suppose that $\mathcal{H}$ is finite-dimensional. Let $\{x(t)\}$ be a sequence generated by* (3a)–(3c). *Then $\{x(t)\}$ converges.*

It is an open question whether Theorem 1 can be extended to the infinite-dimensional case (and thus completely resolve the conjecture of Bruck noted earlier).

Below we consider an order of iterations more restrictive than the one in Theorem 1, under which a weak convergence result is obtainable in the infinite-dimensional setting as well. This order, introduced in [TsB87], is as follows.

Quasi-cyclic order. There exists a sequence of integers $\{\tau_1, \tau_2, \cdots\}$ satisfying

$$\text{(15a)} \qquad \tau_1 = 1, \quad \tau_{k+1} - \tau_k \geq m \quad \forall k \geq 1, \quad \sum_{k=1}^{\infty} \frac{1}{\tau_{k+1} - \tau_k} = \infty,$$

such that

$$\text{(15b)} \qquad \{1, 2, \cdots, m\} \subseteq \{\sigma(\tau_k), \sigma(\tau_k + 1), \cdots, \sigma(\tau_{k+1} - 1)\} \quad \forall k \geq 1.$$

Roughly speaking, the quasi-cyclic order of iterations means that every T_i is applied at least once between the τ_kth and the $(\tau_{k+1} - 1)$th iteration (called the kth quasi cycle) for all k (cf. (15b)) and that the length of the kth quasi cycle, namely $\tau_{k+1} - \tau_k$, cannot grow too fast with k (cf. (15a)). One particular choice of the τ_k's, namely $\tau_k = m(k-1)$ for all k, gives rise to the well known cyclic order for which $\sigma(t) = t(\text{mod } m) + 1$ for all t (and the length of each quasi cycle is exactly m). A more interesting choice of the τ_k's is given by

$$\tau_{k+1} = \tau_k + km \quad \forall k \geq 1,$$

for which the length of the kth quasi cycle increases linearly with k.

By using Lemma 1 we can show the second main result of this investigation. The proof of this is based on an interesting application of the Cauchy–Schwartz inequality.

THEOREM 2. *Let $\{x(t)\}$ be a sequence generated by* (3a)–(3c) *under the quasi-cyclic order* (15a)–(15b). *Then $\{x(t)\}$ has a unique weak cluster point in F.*

Proof. First, we claim that there exists a subsequence K of $\{1, 2, \cdots\}$ for which

$$\text{(16)} \qquad \sum_{t=\tau_k+1}^{\tau_{k+1}} ||x(t+1) - x(t)|| \to 0 \quad \text{as } k \to \infty, \qquad k \in K.$$

To see this, suppose that such a subsequence does not exist. Then there would exist a positive scalar δ and an integer $\bar{k}$ such that

$$\delta \leq \sum_{t=\tau_k+1}^{\tau_{k+1}} ||x(t+1) - x(t)|| \quad \forall k \geq \bar{k}.$$

Since by the Cauchy–Schwartz inequality there holds

$$\sum_{t=\tau_k+1}^{\tau_{k+1}} \|x(t+1)-x(t)\| \le \sqrt{\sum_{t=\tau_k+1}^{\tau_{k+1}} \|x(t+1)-x(t)\|^2} \cdot \sqrt{\tau_{k+1}-\tau_k},$$

this implies

$$\delta^2 \le \sum_{t=\tau_k+1}^{\tau_{k+1}} \|x(t+1)-x(t)\|^2(\tau_{k+1}-\tau_k) \quad \forall k \ge \bar{k},$$

so that

$$\delta^2 \sum_{k=\bar{k}}^{\infty} \frac{1}{\tau_{k+1}-\tau_k} \le \sum_{k=\bar{k}}^{\infty} \left[\sum_{t=\tau_k+1}^{\tau_{k+1}} \|x(t+1)-x(t)\|^2 \right]$$

(17)
$$= \sum_{t=\tau_{\bar{k}}+1}^{\infty} \|x(t+1)-x(t)\|^2.$$

By (15a) the left-hand side of (17) has the extended value of ∞, whereas the right-hand side of (17), according to (5), has finite value, thereby reaching a contradiction. Hence, (16) holds for some $K \subseteq \{1,2,\cdots\}$.

Let K be any subsequence of $\{1,2,\cdots\}$ satisfying (16). Since $\{x(t)\}$ is bounded [cf. (7)], there exist some $x^\infty \in \mathcal{H}$ and some subsequence K' of K such that

(18) $$\{x(\tau_k+1)\}_{k\in K'} \text{ converges weakly to } x^\infty.$$

We claim that $x^\infty \in F$. To see this, fix any $i \in \{1,2,\cdots,m\}$. Since the mappings are applied in the quasi-cyclic order, then for each integer $k \ge 1$ there exists some $\rho_k \in \{\tau_k, \tau_k+1, \cdots, \tau_{k+1}-1\}$ satisfying $\sigma(\rho_k)=i$ (cf. (15b)). By using the triangle inequality, together with the fact that $\|x(\rho_k+1)-y(\rho_k)\| \le (1/\epsilon-1)\|x(\rho_k+1)-x(\rho_k)\|$ for all $k \ge 1$ (cf. (3b), (3c)), we have

(19)

$$\|x(\tau_k+1)-y(\rho_k)\| \le \sum_{t=\tau_k+1}^{\tau_{k+1}} \|x(t+1)-x(t)\| + \|x(\rho_k+1)-y(\rho_k)\|$$

$$\le \sum_{t=\tau_k+1}^{\tau_{k+1}} \|x(t+1)-x(t)\| + \left(\frac{1}{\epsilon}-1\right)\|x(\rho_k+1)-x(\rho_k)\| \quad \forall k.$$

Equations (16) and (19), together with $\{\|x(t+1)-x(t)\|\} \to 0$ [cf. (5)], imply

$$\lim_{k\to\infty, k\in K'} \|x(\tau_k+1)-y(\rho_k)\| = 0,$$

which, combined with (18), yields

$$\lim_{k\to\infty, k\in K'} \langle u, y(\rho_k)\rangle = \lim_{k\to\infty, k\in K'} \langle u, x(\tau_k+1)\rangle = \langle u, x^\infty\rangle \quad \forall u \in \mathcal{H},$$

so that $\{y(\rho_k)\}_{k\in K'}$ converges weakly to x^∞. Since $y(\rho_k) = T_i(x(\rho_k))$ (cf. $\sigma(\rho_k)=i$ and (3a)) for all k, this, together with $\{\|y(\rho_k)-x(\rho_k)\|\} \to 0$ (cf. (6)), implies

$\{||(I - T_i)(x(\rho_k))||\}_{k\in K'} \to 0$ and $\{x(\rho_k)\}_{k\in K'}$ converges weakly to x^∞. Since T_i is nonexpansive so that the mapping $I - T_i$ is demiclosed (see [Opi67, Lemma 2]), we have $(I - T_i)(x^\infty) = 0$ or, equivalently, $x^\infty \in F_i$. Since the choice of i was arbitrary, we obtain $x^\infty \in F_i$ for all i, and therefore $x^\infty \in F$.

We now show that $\{x(t)\}$ has a unique weak cluster point in F. Our argument follows that given in [Bre65] (also see [Roc76, p. 885]) and is presented here for completeness. Suppose that $\{x(t)\}$ does not have a unique weak cluster point in F. Then there would exist $x_1^\infty \in F$ and $x_2^\infty \in F$ with $x_1^\infty \neq x_2^\infty$ and subsequences $\{x(t)\}_{t\in N_1}$, $\{x(t)\}_{t\in N_2}$ converging weakly to, respectively, x_1^∞ and x_2^∞. By replacing $\bar{x}$ in (7) by x_1^∞, we find that $||x_1^\infty - x(t)||$ is nonincreasing with t, so there exists a scalar α_1 such that

$$\{||x_1^\infty - x(t)||^2\} \to \alpha_1. \tag{20a}$$

Similarly, by replacing $\bar{x}$ in (7) by x_2^∞, we find that there exists a scalar α_2 such that

$$\{||x_2^\infty - x(t)||^2\} \to \alpha_2. \tag{20b}$$

Now, for any $t \in N_1$ we have

$$||x_2^\infty - x(t)||^2 = ||x_2^\infty - x_1^\infty||^2 + 2\langle x_2^\infty - x_1^\infty, x_1^\infty - x(t)\rangle + ||x_1^\infty - x(t)||^2,$$

so that, by letting $t \to \infty$, $t \in N_1$, we obtain from (20a) and (20b) and the weak convergence of $\{x(t)\}_{t\in N_1}$ to x_1^∞ that $\alpha_2 = ||x_2^\infty - x_1^\infty||^2 + \alpha_1$. By an analogous argument with the role of x_1^∞ and x_2^∞ reversed, we also obtain $\alpha_1 = ||x_1^\infty - x_2^\infty||^2 + \alpha_2$. Adding these two relations yields $0 = ||x_1^\infty - x_2^\infty||^2$, and hence $x_1^\infty = x_2^\infty$, a contradiction. □

REFERENCES

[Agm54] S. Agmon, *The relaxation method for linear inequalities*, Canad. J. Math., 6 (1954), pp. 382–392.

[AmA65] I. Amemiya and T. Ando,*Convergence of random products of contractions in Hilbert space*, Acta Sci. Math. (Szeged), 26 (1965), pp. 239–244.

[Bre65] L. M. Bregman, *The method of successive projection for finding a common point of convex sets*, Akad. Nauk SSSR Dokl., 162 (1965), pp. 487–490; English translation in Soviet Math. Dokl., 162 (1965), pp. 688–692.

[Bré73] H. Brézis, *Opérateurs Maximaux Monotones et Semi-Groupes de Contractions dans les Espaces de Hilbert*, North-Holland, Amsterdam, 1973.

[Bro76] F. E. Browder, *Nonlinear operators and nonlinear equations of evolution in Banach spaces*, in Proc. Symposia in Pure Mathematics, Vol. 18 (Pt. 2), American Mathematical Society, Providence, RI, 1976.

[Bru82] R. E. Bruck, *Random products of contractions in metric and Banach spaces*, J. Math. Anal. Appl., 88 (1982), pp. 319–332.

[Bru83] ———, *Asymptotic behavior of nonexpansive mappings*, in Fixed Points and Nonexpansive Mappings, R. C. Sine, ed., American Mathematical Society, Providence, RI, 1983, pp. 1–47.

[ChG59] W. Cheney and A. A. Goldstein, *Proximity maps for convex sets*, Proc. Amer. Math. Soc., 10 (1959), pp. 448–450.

[Deu85] F. Deutsch, *Rate of convergence of the method of alternating projections*, in Parametric Optimization and Approximation, B. Brosowski and F. Deutsch, eds., Birkhäuser, Boston, MA, 1985.

[Eck89] J. Eckstein, *Splitting methods for monotone operators with applications to parallel op-*

timization, Ph.D. thesis, Department of Civil Engineering, Massachusetts Institute of Technology, Cambridge, MA, 1989.

[Ere65] I. I. EREMIN, *A generalization of the Motzkin–Agmon relaxational method*, Uspekhi Mat. Nauk, 20 (1965), pp. 183–187. (In Russian.)

[Ere66] ———*On systems of inequalities with convex functions in the left sides*, Izv. Akad. Nauk SSSR Ser. Mat., 30 (1966), pp. 265–278; English translation in Amer. Math. Soc. Transl., 88 (1966), pp. 67–83.

[GeS66] M. A. GERMANOV AND V. S. SPIRIDONOV, *On a method of solving systems of non-linear inequalities*, Zh. Vychisl. Mat. i Mat. Fiz., 6 (1966), pp. 335–336; English translation in U.S.S.R. Comput. Math. and Math. Phys., 2 (1966), pp. 194–196.

[Gof80] J. L. GOFFIN, *The relaxation method for solving systems of linear inequalities*, Math. Oper. Res., 5 (1980), pp. 388–414.

[Gof82] ———, *On the nonpolynomiality of the relaxation method for systems of linear inequalities*, Math. Programming, 22 (1982), pp. 93-103.

[GPR67] L. G. GUBIN, B. T. POLYAK, AND E. V. RAIK, *The method of projections for finding the common point of convex sets*, Zh. Vychisl. Mat. i Mat. Fiz., 7 (1967), pp. 1211–1228; English translation in U.S.S.R. Comput. Math. and Math. Phys., 6 (1967), pp. 1–24.

[Hal62] I. HALPERIN, *The product of projection operators*, Acta Sci. Math. (Szeged), 23 (1962), pp. 96–99.

[Her80] G. T. HERMAN, *Image Reconstructions from Projections, the Fundamentals of Computerized Tomography*, Academic Press, New York, 1980.

[Kac37] S. KACZMARZ, *Angenherte Auflosung von Systemn Linearer Gleichungen*, Bull. Internat. l'Acad. Polonaise Sci. A, (1937), pp. 355–357.

[LeS83] A. LEVI AND H. STARK, *Signal restoration from phase by projections onto convex sets*, J. Opt. Soc. Amer., 73 (1983), pp. 810–822.

[Man84] J. MANDEL, *Convergence of the cyclical relaxation method for linear inequalities*, Math. Programming, 30 (1984), pp. 218–228.

[Mer62] YU. I. MERZLYAKOV, *On a relaxation method of solving systems of linear inequalities*, Zh. Vychisl. Mat. i Mat. Fiz., 2 (1962), pp. 482–487; English translation in U.S.S.R. Comput. Math. and Math. Phys., 2 (1963), pp. 504–510.

[MoS54] T. S. MOTZKIN AND I. J. SCHOENBERG, *The relaxation method for linear inequalities*, Canad. J. Math., 6 (1954), pp. 393–404.

[Opi67] Z. OPIAL, *Weak convergence of the sequence of successive approximations for nonexpansive mappings*, Bull. Amer. Math. Soc., 73 (1967), pp. 591–597.

[Ott88] N. OTTAVY, *Strong convergence of projection-like methods in Hilbert spaces*, J. Optim. Theory Appl., 56 (1988), pp. 433–461.

[Pie84] G. PIERRA, *Decomposition through formalization in a product space*, Math. Programming, 28 (1984), pp. 96–115.

[Pol69] B. T. POLYAK, *Minimization of unsmooth functionals*, Zh. Vychisl. Mat. i Mat. Fiz., 9 (1969), pp. 509–521; English translation in U.S.S.R. Comput. Math. and Math. Phys., 9 (1969), pp. 14–29.

[Pra60] M. PRAGER, *On a principle of convergence in a Hilbert space*, Czechoslovak Math. J., 10 (1960), pp. 271–282.

[Roc76] R. T. ROCKAFELLAR, *Monotone operators and the proximal point algorithm*, SIAM J. Control Optim., 14 (1976), pp. 877–898.

[Sch86] A. SCHRIJVER, *Theory of Linear and Integer Programming*, Wiley-Interscience, New York, 1986.

[Sin83] R. SINE, ED., *Fixed Points and Nonexpansive Mappings*, Contemporary Mathematics, Vol. 18, American Mathematical Society, Providence, RI, 1983.

[SSW77] K. T. SMITH, D. C. SOLMON, AND S. L. WAGNER, *Practical and mathematical aspects of the problem of reconstructing objects from radiographs*, Bull. Amer. Math. Soc., 83 (1977), pp. 1227–1270.

[Tan71] K. TANABE, *Projection method for solving a singular system of linear equations and its applications*, Numer. Math., 17 (1971), pp. 203–214.

[TsB87] P. TSENG AND D. P. BERTSEKAS, *Relaxation methods for problems with strictly convex separable costs and linear constraints*, Math. Programming, 38 (1987), pp. 303–321.

[voN50] J. VON NEUMANN, *Functional Operators, Vol.* II: *The Geometry of Orthogonal Spaces*, Annals of Mathematics Studies, No. 22, Princeton University Press, Princeton, NJ, 1950. (This is a reprint of mimeographed lecture notes first distributed in 1933.)

[You87] D. C. YOULA, *Mathematical theory of image restoration by the method of convex projections*, in Image Recovery: Theory and Application, H. Stark, ed., Academic Press, New York, 1987.

[You89] ———, *On deterministic convergence of iterations of relaxed projection operators*, Tech. Report, Department of Electrical Engineering and Computer Science, Polytechnic University, Farmingdale, NY, 1989.

SIAM J. OPTIMIZATION
Vol. 2, No. 3, pp. 435–449, August 1992

006

ON IMPLEMENTING MEHROTRA'S PREDICTOR–CORRECTOR INTERIOR-POINT METHOD FOR LINEAR PROGRAMMING*

IRVIN J. LUSTIG[†], ROY E. MARSTEN[‡], AND DAVID F. SHANNO[§]

Abstract. Mehrotra [*Tech. Report* 90-03, Department of Industrial Engineering and Management Sciences, Northwestern University, Evanston, IL, 1990] recently described a predictor–corrector variant of the primal–dual interior-point algorithm for linear programming. This paper describes a full implementation of this algorithm, with extensions for solving problems with free variables and problems with bounds on primal variables. Computational results on the NETLIB test set are given to show that this new method almost always improves the performance of the primal–dual algorithm and that the improvement increases dramatically as the size and complexity of the problem increases. A numerical instability in using Schur complements to remove dense columns is identified, and a numerical remedy is given.

Key words. linear programming, interior-point methods, predictor–corrector algorithms

AMS(MOS) subject classifications. 90C05, 90C06

1. Introduction. Mehrotra [10] has recently introduced a remarkable higher-order primal–dual logarithmic barrier method for linear programming. He motivates this method as a power series method, but in a nonstandard way. He also introduces a potential function that can be linearly searched to ensure a constant reduction at each step. A numerical algorithm using this method is described, and computational comparisons are given to MINOS 5.3 and OB1, the primal–dual interior-point code of Lustig, Marsten, and Shanno [8]. These new results indicate that the method represents a significant computational advance for linear programming.

The test set that Mehrotra used to evaluate his method is a small subset of the NETLIB [6] test set, excluding all problems with upper-bounded variables, free variables, and ranges. Since most of the large and difficult problems in the NETLIB test set were omitted from Mehrotra's tests, the initial goal of this paper is to describe an implementation extending Mehrotra's algorithm to solve all standard form linear programs. Mehrotra's initial results tend to sell his method short, because its performance, when compared with that of the pure primal–dual algorithm, improves even more as the problems get larger and more complex.

A second goal of this paper is to isolate the effects of various new components of Mehrotra's framework. Mehrotra discusses the results of his code compared only with OB1. However, because OB1 has many safeguards (such as iterative refinement) needed to solve large ill-conditioned problems, and because OB1 incurs overhead to

*Received by the editors May 25, 1990; accepted for publication (in revised form) August 20, 1991.

[†]Department of Civil Engineering and Operations Research, Princeton University, Princeton, New Jersey 08544.

[‡]School of Industrial Engineering and Operations Research, Georgia Institute of Technology, Atlanta, Georgia 30332. The research of this author is partially supported by Office of Naval Research grant N00014-88-K-0104.

[§]Rutgers Center for Operations Research, Rutgers University, New Brunswick, New Jersey 08903. The author is currently a Visiting Fellow at the Department of Civil Engineering and Operations Research, Princeton University, Princeton, New Jersey 08544. The research of this author is sponsored by the U.S. Air Force Office of Scientific Research, Air Force Systems Command, under grant AFOSR-87-0215.

handle upper bounds (which are always assumed to be present), direct comparisons of the two codes and evaluations of the relative effects of components of the algorithm require a common implementation. Once this was accomplished within the framework of OB1, it became easy to compare Mehrotra's algorithms more accurately with the primal–dual algorithms of OB1.

Another purpose of this paper is to develop Mehrotra's algorithm, including bounds, ranges, and free variables, within the framework of a predictor–corrector method. Although Mehrotra's paper [10] uses a power series motivation, Mehrotra himself gave a different motivation for the method in a presentation at a conference in Asilomar. In this presentation, he described a method drawn from predictor–corrector methods for ordinary differential equations. This interpretation is particularly seminal in motivating effective higher-order methods and is extremely clear geometrically in showing how the methods incorporate higher-order information. Therefore, our paper uses the predictor–corrector motivation.

Finally, since we are concerned with solving large, difficult problems, we again confront the enigma of removing dense columns. This is essential for solving two new problems, `fit1p` and `fit2p`, of the NETLIB test set within our memory constraints. We identify an instability that arises from using Schur complements to accomplish this task and suggest a remedy that has worked well in practice.

Section 2 briefly introduces the primal–dual algorithm OB1, with some minor improvements. Section 3 derives the predictor–corrector method for bounded problems and discusses some of the difficulties that had to be overcome. Section 4 deals with dense column removal, and §5 gives computational results indicating that the predictor–corrector method is indeed a major advance in the state of the computational art.

2. The primal–dual method for linear programming. Lustig, Marsten, and Shanno [8] describe in detail a path-following primal–dual interior-point method for linear programming derived from a logarithmic barrier method. Here we briefly recapitulate the algorithm both for the convenience of the reader and because the first-order conditions will be required to motivate subsequent work on the new primal–dual predictor–corrector method. The primal problem we are concerned with here is

$$\min c^T x \qquad \text{subject to} \qquad Ax = b, \qquad 0 \le x \le u, \tag{1}$$

where some or all of the upper bounds may be infinite. We assume that $A \in \Re^{m \times n}$, $b \in \Re^m$, $c \in \Re^n$, $u \in \Re^n$, and $x \in \Re^n$. Adding slack variables s to transform the upper-bound inequalities to equalities and eliminating the inequality constraints by incorporating them in a logarithmic barrier term appended to the objective function, the Lagrangian for (1) becomes

$$L(x,s,y,w,\mu) = c^T x - \mu \sum_{j=1}^{n} \ln x_j - \mu \sum_{j=1}^{n} \ln s_j - y^T(Ax-b) - w^T(u-x-s). \tag{2}$$

The first-order necessary conditions for a stationary point of (2) are

$$\begin{aligned} Ax &= b, \\ x + s &= u, \\ A^T y - w + z &= c, \\ XZe &= \mu e, \\ SWe &= \mu e, \end{aligned} \tag{3}$$

where X, Z, S, and W are diagonal matrices with the elements x_j, z_j, s_j, and w_j, respectively, and $z \in \Re^n$ is the vector of the dual slack variables.

The primal–dual algorithm is derived from the first-order conditions (3) by applying one iteration of Newton's method to find an approximate solution to (3) for a fixed value of μ and continuing until the complementarity $x^T z + s^T w$ is reduced to a predetermined tolerance. In [8] this algorithm was implemented specifically by assuming that a current estimate x, y, z, w, s was available satisfying $x > 0, z > 0, w > 0$, and $s > 0$, with the further restriction that $x + s = u$, i.e., that the upper bounds were always explicitly satisfied. Fixing μ and applying one step of Newton's method to (3) yields the set of equations

$$\begin{aligned} A\Delta x &= b - Ax, \\ \Delta x + \Delta s &= 0, \\ A^T\Delta y + \Delta z - \Delta w &= c - A^T y - z + w, \\ Z\Delta x + X\Delta z &= -XZe + \mu e, \\ W\Delta s + S\Delta w &= -SWe + \mu e. \end{aligned} \tag{4}$$

Defining

$$\Theta = (X^{-1}Z + S^{-1}W)^{-1} \tag{5}$$

and

$$\rho(\mu) = \mu(S^{-1} - X^{-1})e - (W - Z)e, \tag{6}$$

the solution to (4) is

$$\begin{aligned} \Delta y &= (A\Theta A^T)^{-1}[(b - Ax) + A\Theta((c - A^T y - z + w) + \rho(\mu))], \\ \Delta x &= \Theta[A^T\Delta y - \rho(\mu) - (c - A^T y - z + w)], \\ \Delta z &= \mu X^{-1}e - Ze - X^{-1}Z\Delta x, \\ \Delta w &= \mu S^{-1}e - We + S^{-1}W\Delta x, \\ \Delta s &= -\Delta x. \end{aligned} \tag{7}$$

A new point x^*, s^*, y^*, z^*, w^* is then defined by

$$\begin{aligned} x^* &= x + \alpha_P\Delta x, \\ s^* &= s + \alpha_P\Delta s, \\ y^* &= y + \alpha_D\Delta y, \\ z^* &= z + \alpha_D\Delta z, \\ w^* &= w + \alpha_D\Delta w, \end{aligned} \tag{8}$$

where α_P and α_D are respective step lengths in the primal and dual spaces chosen to assure the nonnegativity of the variables x, s, z, and w. At each step the barrier parameter μ is reduced by a method discussed below, and the algorithm continues until the relative duality gap satisfies

$$\frac{c^T x - b^T y + u^T w}{1 + |b^T y - u^T w|} < \epsilon \tag{9}$$

for a user-predetermined ϵ.

In [8] μ is chosen at each step to be

$$\mu = \frac{c^Tx - b^Ty + u^Tw + M\delta_1 + M\delta_2}{\phi(n)}, \tag{10}$$

where

$$\delta_1 = \|b - Ax\|/\|b - Ax^0\|, \tag{11}$$

$$\delta_2 = \|c - A^Ty - z + w\|/\|c - A^Ty^0 - z^0 + w^0\|, \tag{12}$$

$$\phi(n) = \begin{cases} n^2, & \text{if } n \leq 5000, \\ n^{3/2}, & \text{if } n > 5000, \end{cases} \tag{13}$$

and M is an appropriately chosen large constant, which in [8] was given by

$$M = \xi\phi(n)\max\{\max_{1\leq j\leq n}\{|c_j|\}, \max_{1\leq i\leq n}\{|b_i|\}\}. \tag{14}$$

A somewhat complex algorithm for choosing x^0, y^0, z^0, and ξ is documented in [8]. In the numerical results given in [8], it was noted that for these choices of x^0, y^0, z^0, and ξ, 69 of the 71 problems of the NETLIB test set tested in [8] converged to eight digits of accuracy, whereas `pilot4` and `capri` required ξ to be reduced in order to obtain that degree of accuracy.

The algorithm corresponding to (7) and (8) that we use to compare with the new predictor–corrector algorithm modifies the ξ of [8] to achieve eight digits of accuracy on all problems with the default settings. Here we choose ξ initially, as in [8], and compute the initial search vector. Denoting this vector by

$$\begin{aligned} \Delta x &= \Delta x_1 + \mu\Delta x_2, \\ \Delta y &= \Delta y_1 + \mu\Delta y_2, \\ \Delta z &= \Delta z_1 + \mu\Delta z_2, \\ \Delta w &= \Delta w_1 + \mu\Delta w_2, \end{aligned} \tag{15}$$

we compute the vector norms

$$\Delta_1 = \|\Delta x_1 + \Delta y_1 + \Delta z_1 + \Delta w_1\|$$

and

$$\Delta_2 = \|\Delta x_2 + \Delta y_2 + \Delta z_2 + \Delta w_2\|. \tag{16}$$

Then, if $\mu\Delta_2 < 0.7\Delta_1$, we increase ξ by a factor of 10, and if $10\Delta_1 < \mu\Delta_2$, we decrease ξ by a factor of 10. Thus, ξ is dynamically adjusted to attempt to bring the lengths of the optimality–feasibility vector Δ_1 and the centering vector Δ_2 to approximately equal magnitudes, where in (16) $\|\cdot\|$ is the l_1 norm.

This readjustment of ξ allowed all previously tested problems, plus 14 of the 15 additional problems recently added to the NETLIB test set, to be solved to 8 digits of accuracy using the default options. Moreover, a comparison between the iteration counts given in §4 for this method and the counts of [8] shows that overall the method is not only more stable, but more efficient. Thus, if a pure primal–dual method is desired, it seems important to choose the initial μ to roughly equate the initial norms of the vectors Δ_1 and Δ_2. However, because of the clear numerical dominance of the method that is documented in the next section, the importance of the pure primal–dual algorithm is questionable at best.

3. Mehrotra's predictor–corrector method. Mehrotra [10] introduces a power series variant of the primal–dual algorithm without considering explicit bounds. This algorithm was initially described by Mehrotra [9] without any given motivation. Here we derive a version of the method described in [9] including bounded variables but, as previously noted, using the predictor–corrector motivation. This algorithm can also be viewed as an extension of one of the algorithms presented by Mehrotra in [10]. The method again uses the logarithmic barrier Lagrangian (2) to derive the first-order conditions (3). Rather than applying Newton's method to (3) to generate correction terms to the current estimate, we substitute the new point into (3) directly, yielding

$$\begin{aligned} A(x+\Delta x) &= b,\\ (x+\Delta x)+(s+\Delta s) &= u,\\ A^T(y+\Delta y)-(w+\Delta w)+(z+\Delta z) &= c,\\ (X+\Delta X)(Z+\Delta Z)e &= \mu e,\\ (S+\Delta S)(W+\Delta W)e &= \mu e, \end{aligned} \tag{17}$$

where ΔX, ΔZ, ΔS, and ΔW are diagonal matrices having elements Δx, Δz, Δs, and Δw, respectively. Simple algebra reduces (17) to the equivalent system

$$A\Delta x = b - Ax, \tag{18a}$$

$$\Delta x + \Delta s = u - x - s, \tag{18b}$$

$$A^T\Delta y - \Delta w + \Delta z = c - A^Ty + w - z, \tag{18c}$$

$$X\Delta z + Z\Delta x = \mu e - XZe - \Delta X\Delta Ze, \tag{18d}$$

$$S\Delta w + W\Delta s = \mu e - SWe - \Delta S\Delta We. \tag{18e}$$

The left-hand side of (18) is identical to (4), while the right-hand side has two distinct differences. The first deals with the equation defining the upper bounds. In the algorithm of §2 we always choose x^0 and s^0 so that $x^0 > 0$, $s^0 > 0$, and $x^0+s^0 = u$. Thus the right-hand side of equation (18b) is always zero. For reasons to be discussed later, this proves to be computationally unstable for the predictor–corrector method on problems with small upper bounds. Thus we assume only that $x^0 > 0$ and $s^0 > 0$ but not that the upper-bound constraint $x^0 + s^0 = u$ is satisfied initially. Rather, we allow the method to iterate to bound feasibility in precisely the same manner it iterates to primal and dual feasibility.

The major difference between (4) and (18) is the presence of the nonlinear terms $\Delta X\Delta Ze$ and $\Delta S\Delta We$ in the right-hand side of (18d) and (18e). Thus (18) implicitly defines the step Δx, Δy, Δz, Δs, Δw. To determine a step approximately satisfying (18), Mehrotra suggests first solving the defining equations for the primal–dual affine direction:

$$\begin{aligned} A\Delta\hat{x} &= b - Ax,\\ \Delta\hat{x} + \Delta\hat{s} &= u - x - s,\\ A^T\Delta\hat{y} - \Delta\hat{w} + \Delta\hat{z} &= c - A^Ty + w - z,\\ X\Delta\hat{z} + Z\Delta\hat{x} &= -XZe,\\ S\Delta\hat{w} + W\Delta\hat{s} &= -SWe. \end{aligned} \tag{19}$$

These directions are then used in two distinct ways: to approximate the nonlinear terms in the right-hand side of (18) and to dynamically estimate μ.

To estimate μ, Mehrotra performs the standard ratio test on both the primal and dual variables to determine the step that would actually be taken if the primal–dual affine direction defined by (19) were used. Thus, we define

$$\begin{aligned} \hat{\delta}_P &= \min\left\{\min_j\left\{\frac{x_j}{-\Delta\hat{x}_j}, \Delta\hat{x}_j < 0\right\}, \min_j\left\{\frac{s_j}{-\Delta\hat{s}_j}, \Delta\hat{s}_j < 0\right\}\right\}, \\ \hat{\delta}_D &= \min\left\{\min_j\left\{\frac{z_j}{-\Delta\hat{z}_j}, \Delta\hat{z}_j < 0\right\}, \min_j\left\{\frac{w_j}{-\Delta\hat{w}_j}, \Delta\hat{w}_j < 0\right\}\right\}, \end{aligned} \tag{20}$$

and let

$$\begin{aligned} \delta_P &= 0.99995\hat{\delta}_P, \\ \delta_D &= 0.99995\hat{\delta}_D. \end{aligned}$$

Then the new complementarity gap that would result from a step in the affine direction is

$$\hat{g} = (x + \delta_P\Delta\hat{x})^T(z + \delta_D\Delta\hat{z}) + (s + \delta_P\Delta\hat{s})^T(w + \delta_D\Delta\hat{w}). \tag{21}$$

Mehrotra's estimate in [9] for μ, generalized to include lower bounds, is then

$$\mu = \left(\frac{\hat{g}}{x^Tz + s^Tw}\right)^2\left(\frac{\hat{g}}{n}\right), \tag{22}$$

which chooses a small μ when good progress can be made in the affine direction and a large μ when the affine direction produces little improvement. This is appealing, since poor progress in the affine direction generally indicates the need for more centering and hence a larger value of μ. In [10] Mehrotra uses a cubic rather than a quadratic multiplier, but comparison of the iteration counts of §5 with those of [10] indicates that this appears to make little difference.

In the implementation tested and documented in §5, we have essentially adopted this same algorithm for choosing μ with one minor difference. We found that choosing μ by (22) can result in numerically unstable systems as the optimum is approached on poorly conditioned problems, such as the `pilot` models. Thus when the absolute complementarity $x^Tz + s^Tw \geq 1$, we define μ by (22), but when $x^Tz + s^Tw < 1$, we define

$$\mu = (x^Tz + s^Tw)/\phi(n), \tag{23}$$

where $\phi(n)$ is defined by (13). In practice this proved totally satisfactory and far more stable than always choosing μ by (22).

The actual new step $\Delta x, \Delta y, \Delta z, \Delta s, \Delta w$ is then chosen as the solution to

$$\begin{aligned} A\Delta x &= b - Ax, \\ \Delta x + \Delta s &= u - x - s, \\ A^T\Delta y - \Delta w + \Delta z &= c - A^Ty + w - z, \\ X\Delta z + Z\Delta x &= \mu e - XZe - \Delta\hat{X}\Delta\hat{Z}e, \\ S\Delta w + W\Delta s &= \mu e - SWe - \Delta\hat{S}\Delta\hat{W}e. \end{aligned} \tag{24}$$

Clearly, all that has changed from (4) is the right-hand side, so the matrix algebra remains the same as in the solution (7). Ratio tests identical to (20) are now done using $\Delta x, \Delta y, \Delta z, \Delta s$, and Δw to determine actual step sizes α_P and α_D, and the actual new point x^*, y^*, z^*, s^*, w^* is defined by (8).

After the coefficient matrix $A\Theta A^T$ has been factored, the additional work of the predictor–corrector method is in the extra backsolve to compute the affine direction and the extra ratio test used to compute μ. What is gained from this extra work is approximate second-order information concerning the trajectory from the current estimate to the optimal point as μ is varied continuously.

To see this, note that the full primal–dual affine correction terms, $\Delta\hat{X}\Delta\hat{Z}e$ and $\Delta\hat{S}\Delta\hat{W}e$, are added to the right-hand side of (24). The step lengths δ_P and δ_D defined by (20) are used to compute μ but not to modify the correction. Thus the correction added is one that would result from taking a full step of length 1 in the affine direction. Whenever a primal and dual full step of length 1 can be taken, primal feasibility and dual feasibility are achieved exactly within the numerical accuracy of the computations. Now the solution to (24) can be written as

$$\begin{aligned} \Delta y &= \Delta\hat{y} + c_y, \\ \Delta x &= \Delta\hat{x} + c_x, \\ \Delta z &= \Delta\hat{z} + c_z, \\ \Delta w &= \Delta\hat{w} + c_w, \\ \Delta s &= \Delta\hat{s} + c_s, \end{aligned} \tag{25}$$

where $\Delta\hat{x}, \Delta\hat{y}, \Delta\hat{z}, \Delta\hat{s}, \Delta\hat{w}$ are the solution to (19) and the correction terms c_x, c_y, c_z, c_s, and c_w satisfy

$$\begin{aligned} Ac_x &= 0, \\ c_x + c_s &= 0, \\ A^Tc_y + c_z - c_w &= 0, \\ Xc_z + Zc_x &= \mu e - \Delta\hat{X}\Delta\hat{Z}e, \\ Sc_w + Wc_s &= \mu e - \Delta\hat{S}\Delta\hat{W}e. \end{aligned} \tag{26}$$

Now, if the full step of 1 were achieved on this affine step, the new complementarity would be precisely

$$\begin{aligned} &(x + \Delta\hat{x})^T(z + \Delta\hat{z}) + (s + \Delta\hat{s})^T(w + \Delta\hat{w}) \\ &\quad = x^Tz + \Delta\hat{x}^Tz + \Delta\hat{z}^Tx + \Delta\hat{x}^T\Delta\hat{z} + s^Tw + \Delta\hat{s}^Tw + \Delta\hat{w}^Ts + \Delta\hat{s}^T\Delta\hat{w} \\ &\quad = \Delta\hat{x}^T\Delta\hat{z} + \Delta\hat{s}^T\Delta\hat{w}, \end{aligned}$$

as by the definition of $\Delta\hat{x}, \Delta\hat{s}, \Delta\hat{z}, \Delta\hat{w}$ in (19),

$$x^Tz + \Delta\hat{x}^Tz + \Delta\hat{z}^Tx = s^Tw + \Delta\hat{z}^Tw + \Delta\hat{w}^Ts = 0.$$

Thus (26) could describe a centered Newton step from the point $x + \Delta\hat{x}$, $y + \Delta\hat{y}$, $z + \Delta\hat{z}$, $s + \Delta\hat{s}$, $w + \Delta\hat{w}$ except that the corrections $\Delta\hat{x}$, $\Delta\hat{y}$, $\Delta\hat{z}$, $\Delta\hat{s}$, $\Delta\hat{w}$ have not been added to the diagonal matrices on the left-hand side of (26). Thus the correction terms are a Newton step from the point achieved by a full affine step, but using the

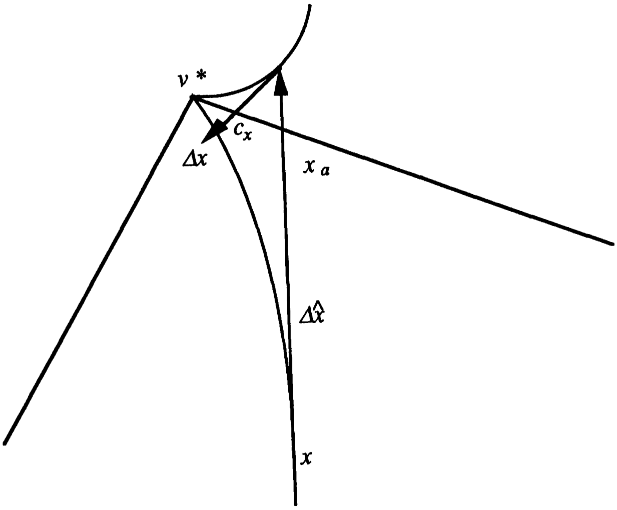


FIG. 1. *Diagram of direction vectors.*

second-derivative matrix at the current point x, y, z, w, s as an approximation to the Hessian at the point corresponding to a full affine step.

The effect of this is demonstrated in Fig. 1, shown in terms of the primal variables x. Here the vertex v^* is the desired optimum. The affine direction $\Delta\hat{x}$ from the current estimate x is a vector tangent to the continuous trajectory as $\mu \to 0$, as is shown in the figure. The point x_a is the new point predicted by the affine variant with a step of length 1, which by the previous analysis must lie outside the feasible region. The correction term c_x that is added at x_a is not tangent to the trajectory through x_a for two reasons. First, a centering term corresponding to $\mu > 0$ is added, rotating c_x toward the center of the polytope. Second, the second-derivative matrix used to calculate c_x is computed at x and not at x_a. Hence, even without a centering term, the vector c_x would not be exactly tangent at x_a. However, the curvature of the trajectories defined through each point by continuously varying μ is clearly estimated by the method. The results of §5 show this to be extremely effective computationally.

Higher-order methods for interior-point algorithms have been proposed by Bayer and Lagarias [2] and implemented by Adler, Karmarkar, Resende, and Veiga [1] and Domich, Boggs, Rogers, and Witzgall [5]. Each of these implementations uses power series estimates to the trajectory at x and shows that higher-order methods reduce the iteration count, with occasional significant reductions. However, Adler et al. report that the increased costs of computing the higher-order terms generally eliminate any effective advantage gained from the lower iteration counts, whereas Domich et al. do not provide comparative timings. As previously noted, Mehrotra [10] uses a different power series derivation for the predictor–corrector method.

The results of §5 show that the predictor–corrector method almost always reduces the iteration count and usually reduces computation time. Furthermore, as problem size and complexity increase, the improvements in both iteration count and execution time become greater. Thus the predictor–corrector method is a higher-order method that is generally very computationally efficient.

Another small variant in the algorithm involves the choice of step size. Zhang, Tapia, and Dennis [11] have recently shown that for nondegenerate problems the primal–dual method will have an asymptotic quadratic rate of convergence if a step of length 1 is taken as the optimum is approached. Thus, if $0.99995\hat{\delta}_P > 1$ or $0.99995\hat{\delta}_D > 1$, we restrict the step length to 1. Further, if δ_P and δ_D are both equal to 1, a correction is not added on this iteration, but the affine direction is accepted with a step length of 1, thus attaining primal and dual feasibility in one step in the affine direction.

As a final note on this section, we have stated that this implementation of the predictor–corrector algorithm allows bound infeasibility. This proved to be necessary for the predictor–corrector method as it is more sensitive to the starting point than the pure primal–dual method. Although it is theoretically always easy to maintain bounds exactly, the predictor–corrector method performs best with relatively large initial estimates to the primal variables x. Some models of the NETLIB test set, notably the `pilot` models, have upper bounds of 10^{-5} for some variables, and hence maintaining bound feasibility requires very small initial estimates to x. For these estimates, the method has proved quite unstable and often fails to converge because of numerical problems. Allowing the initial estimates for x and s to violate the bounds resolves these problems and is quite efficient in practice.

4. Yet again, dense columns. In Choi, Monma, and Shanno [4] the use of Schur complements to eliminate dense columns from A to assure a sufficiently sparse factorization of $A\Theta A^T$ is discussed. In [8] the authors discuss the effects of this in greater detail and document numerical stability problems with the algorithm. It is the purpose of this section to identify the instability, suggest a possible remedy, and advise that the remedy be used with extreme caution.

The nature of the problem can be seen easily from the following simple linear programming problem:

$$\begin{array}{lll} \min & -x_1 & \\ \text{subject to} & & x_1 + x_2 = 2, \\ & & x_1 + x_3 = 1, \\ & & x_1,\ x_2,\ x_3 \geq 0. \end{array} \tag{27}$$

This has the solution $x_1 = 1$ and $x_2 = 1$. If we consider the first column to be dense, the resulting sparse matrix factored during the Schur complement procedure is

$$\begin{aligned} A_s\Theta_s A_s^T &= \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} x_2/z_2 & 0 \\ 0 & x_3/z_3 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \\ &= \begin{bmatrix} x_2/z_2 & 0 \\ 0 & x_3/z_3 \end{bmatrix}. \end{aligned} \tag{28}$$

As the solution is approached, $x_3 \to 0$ and this matrix becomes rank deficient. This introduces sufficient instability into the Schur complement procedures to make it impossible at times to achieve the desired accuracy. Further, this is not a contrived example, for often the removal of dense columns leaves all remaining elements of one or more rows identically zero. Thus an artificial variable must be added to each of these rows to assure that $A_s\Theta_s A_s^T$ is not rank deficient. Since these artificial variables must be driven to zero, rank deficiency always results as the optimum is approached, and the desired accuracy is almost never attained.

To attempt to alleviate this problem when Schur complements are being used, the algorithm monitors the spread between the largest and the smallest diagonal elements of the factorization

$$L_s D_s L_s^T = A_s\Theta_s A_s^T. \tag{29}$$

When this spread becomes large, all smaller diagonal elements are set to 1 and the remaining elements of these columns are set to 0. These Cholesky factors are used as preconditioners for a preconditioned conjugate gradient method [1] with the dense column added to the matrix A. This eliminates rank deficiency and removes much of the instability. Although, in general, Schur complements are more efficient than preconditioned conjugate gradients, in this case preconditioned conjugate gradients allowed for accurate solutions while Schur complements did not.

For example, the linear program (27) is easily solved in 10 iterations by the primal–dual method of §2 without dense columns removed and by the same method combined with the preconditioned conjugate gradient method with the single dense column removed. However, failure occurs because of numerical instability when (27) is solved by using pure Schur complements. Interestingly, (27) is easily solved with the single dense column removed by using the predictor–corrector algorithm of §3 and Schur complements, indicating that the new method may be much more stable than pure primal-dual algorithms. On problem `seba`, when columns of length 50 or more are removed and the minimum local fill-in ordering for the matrix $A_s\Theta_s A_s^T$ is used, the predictor–corrector method can achieve only seven digits of accuracy before failing at the 19th iteration. However, the pure primal–dual method easily achieves eight digits of accuracy in 28 iterations.

It is indeed instructive to examine how the predictor–corrector method performs on `seba` under different options. Five runs were executed and are identified in Table 1 as `minfil` for Schur complements with the minimum local fill-in ordering, `mmind` for Schur complements with the multiple minimum degree ordering, `nodense` for a pure Cholesky factorization with no dense columns removed, `switch` for Schur complements with the minimum local fill-in ordering switching to preconditioned conjugate gradients when the spread of the diagonal elements was greater than 10^{14}, and `conjgrad` for removing dense columns by the preconditioned conjugate gradient method at each iteration. The times reported are simply solution times.

TABLE 1
Timings on SEBA.

Method	CPU seconds	Accuracy
`minfil`	14.60	10^{-7}
`mmind`	14.70	10^{-8}
`nodense`	388.70	10^{-8}
`switch`	32.44	10^{-8}
`conjgrad`	86.85	10^{-8}

Clearly, removal of dense columns on `seba` is valuable. Each of the four methods that removed dense columns took approximately 0.15 seconds to do the ordering, independent of whether it was minimum local fill-in or multiple minimum degree. However, minimum local fill-in without removal of the dense columns required 91.05 seconds to do the ordering. The fact that the same predictor–corrector algorithm achieved different accuracies with different orderings indicates the sensitivity of the algorithm to ill-conditioned matrices. Finally, the results clearly show that even for problems with a few dense columns (**seba** has 14 dense columns), preconditioned conjugate gradient methods are much slower than Cholesky factorizations and should be used sparingly. In view of this, the **switch** option appears sensible when accuracy is being lost using Schur complements, but further experimentation is needed to determine the optimal time to switch from Schur complements to preconditioned conjugate gradients.

For the NETLIB test set we needed to remove dense columns for only four problems: **israel, seba, fit1p**, and **fit2p**. Problems **israel** and **seba** can both be solved without removing dense columns. However, **fit1p** and especially **fit2p** must have dense columns removed in order to be solved, due to their sufficient size and density. Problems **fit1p** and **fit2p** demonstrate that the capability of removing dense columns is not just more efficient, but at times is actually required. In view of this need, the computational results of the next section give the results for **israel, seba, fit1p**, and **fit2p** with pure Schur complements. However, only seven digits of accuracy were achieved for **seba**. Eight were achieved by using the switch option.

5. Computational results. OB1, a modularized FORTRAN-77 code that was developed to implement the primal–dual barrier method, is documented by Lustig, Marsten, and Shanno [8]. This code for the pure barrier method was modified to alter the initial μ, as described in §2, but otherwise remains as in [8] with the single further exception that at each step we move to a point corresponding to 0.99995 of the distance to the boundary, rather than 0.995.

The predictor–corrector method described in §3 was implemented within the same software, for which the method is selected by a user-specified parameter. As noted in §3, the predictor–corrector method is quite sensitive to the initial guess to the optimal solution. Following Mehrotra [9], an initial estimate to the primal variables x was chosen to be

$$\tilde{x} = A^T(AA^T)^{-1}b. \tag{30}$$

We then define

$$\xi_1 = \max(-\min_{1\leq j\leq n} \tilde{x}_j, 100, \|b\|/100) \quad \text{and} \quad \xi_2 = 1 + \|c\|, \tag{31}$$

where $\|\cdot\|$ is the l_1 norm.

Then, for each $j = 1, \cdots, n$,

$$x_j^0 = \max(\tilde{x}_j, \xi_1) \quad \text{and} \quad s_j^0 = \max(\xi_1, u_j - x_j^0). \tag{32}$$

We set $y^0 = 0$ and the pair z^0, w^0 to satisfy

$$\begin{aligned} &z_j = c_j + \xi_2, \quad w_j = \xi_2, \quad \text{if } c_j > \xi_2,\\ &z_j = -c_j, \quad w_j = -2c_j, \quad \text{if } c_j < -\xi_2,\\ &z_j = c_j + \xi_2, \quad w_j = \xi_2, \quad \text{if } 0 \leq c_j < \xi_2,\\ &z_j = \xi_2, \quad w_j = -c_j + \xi_2, \quad \text{if } -\xi_2 \leq c_j \leq 0. \end{aligned} \tag{33}$$

In addition, if we have upper bounds satisfying $u_j < 0.001$ for any j, the components of x^0 are all set to 100. We believe that further experimentation along the lines of Mehrotra [9] would yield a more stable algorithm to determine the starting point.

Besides adding the predictor–corrector method to OB1, we implemented the option of ordering the matrix $A\Theta A^T$ by either multiple minimum degree or minimum local fill-in. These are discussed in detail by de Carvalho [3]. His tests show that overall performance on large problems indicates a definite preference for minimum local fill-in. We tested both orderings as well, and the results of our experiments on a large test set totally reinforce his conclusion, although no strong inference can be drawn from testing on small problems.

The numerical experiments conducted here were done on 86 problems of the expanded NETLIB [6] test set. We did not solve problems `stocfor3` or `truss` because they require substantial time to generate the MPS file from FORTRAN code. The problem `stocfor3` is large enough to be difficult to solve in our computing environment. All experiments were conducted on a Silicon Graphics 4D/70 workstation with 16 megabytes of memory and one processor. The code was compiled with the MIPS `f77` compiler using options `-O2` and `-Olimit 800`.

The main results compare the relative efficiency of the primal–dual and predictor–corrector algorithms. The results are documented in Tables 2 and 3. Solution times do not include preprocessing time or the time to compute the minimum local fill-in ordering. Each method solved 85 of the 86 test problems with the default settings. As previously noted for `seba`, once dense columns of more than 50 elements were eliminated, the predictor–corrector method attained only seven digits of accuracy. The primal–dual method achieved only seven digits of accuracy on `fit2p` with the standard default and dense columns again removed, but easily achieved eight digits of accuracy with the value of ξ in (14) raised from the default of 0.1 to 1. Both methods in this comparison used the minimum local fill-in ordering.

The predictor–corrector method outperformed the primal–dual method on iteration count for 85 of the 86 problems. More importantly, it outperformed the primal–dual in execution time for 71 of the 86 problems. As expected, the percentage decrease in iterations is always greater than the percentage decrease in run time because of the extra back solve and ratio test required. However, the percentage decrease in run time is still quite impressive, especially on large, difficult problems. Yet the run time percentages can still be improved. Both methods test to assure that $A\Delta x$ is sufficiently close to $b - Ax$, which entails calculating $A\Delta x$ at each iteration. When this test fails, iterative refinement is invoked. Here, we use the specific form of iterative refinement devised for least-squares problems [7]. We have found this to be an important safety feature for large, difficult problems, but on most problems of the NETLIB test set it represents unnecessary overhead and could be removed. It should be noted that on many problems the number of nonzeros in A and L are of the same order of magnitude; hence, this test can be as expensive as a forward and backward pass through L. OB1 has been designed to solve difficult problems with little intervention from the user. Therefore, we prefer to present results on an algorithm designed for maximum stability. For the user who never requires iterative refinement, the computational superiority of the predictor–corrector method may be even more pronounced. The one unmistakable conclusion, with or without iterative refinement, is that the predictor–corrector algorithm is substantially superior to the pure primal–dual algorithm and that this superiority grows with problem size and complexity.

Iteration counts of the algorithms documented here are comparable to Mehrotra's algorithm [10], even though small differences occur on individual problems. Thus Mehrotra's somewhat greater computational advantage in terms of computation time over the primal–dual algorithms of OB1 is due almost entirely to the extra overhead imposed by the check for iterative refinement, inclusion of bounds, and other safeguards we found necessary for larger, more complex problems. The one check that we have not included is one to assure that his potential function is reduced sufficiently at

TABLE 2
Computational results for OB1 (A–N).

Problem	No. of	No. of	No. of	Primal/Dual Meth.		Pred./Corr. Meth.		Dynamic μ	
Name	Rows	Cols.	Nonzeros	Its.	Sol. Time	Its.	Sol. Time	Its.	Sol. Time
25fv47	821	1571	10400	47	148.48	25	92.41	46	167.55
80bau3b	2262	9799	21002	70	363.72	38	253.98	64	425.97
adlittle	56	97	383	16	0.70	12	0.73	17	1.03
afiro	27	32	83	13	0.26	9	0.28	11	0.32
agg	488	163	2410	28	19.24	24	19.52	29	23.65
agg2	516	302	4284	27	46.34	18	36.22	27	53.62
agg3	516	302	4300	26	44.82	17	34.50	25	49.99
bandm	305	472	2494	28	7.53	17	5.87	28	9.41
beaconfd	173	262	3375	18	3.39	10	2.45	14	3.31
blend	74	83	491	15	1.07	14	1.26	18	1.61
bnl1	643	1175	5121	56	54.05	27	32.77	49	58.49
bnl2	2324	3489	13999	59	1037.78	33	632.11	61	1159.18
boeing1	351	384	3485	38	15.72	24	12.84	40	21.08
boeing2	166	143	1196	27	3.39	14	2.41	23	3.75
bore3d	233	315	1429	39	4.69	18	2.87	27	4.28
brandy	220	249	2148	28	6.41	19	5.35	24	6.77
capri	271	353	1767	44	16.07	18	8.43	29	13.48
cycle	1903	2857	20720	51	196.09	30	137.74	48	218.43
czprob	929	3523	10669	59	48.58	35	40.09	53	60.14
d2q06c	2171	5167	32417	53	804.24	31	525.11	51	856.96
degen2	444	534	3978	24	38.44	14	26.31	23	42.02
degen3	1503	1818	24646	31	766.81	20	551.83	40	1082.83
e226	223	282	2578	27	7.01	22	7.19	34	11.07
etamacro	400	688	2409	45	37.61	29	28.80	50	49.30
fffff800	524	854	6227	60	79.15	28	42.99	44	67.21
finnis	497	614	2310	41	13.96	26	11.78	42	18.88
fit1d	24	1026	13404	22	16.19	18	17.16	28	26.39
fit1p	627	1677	9868	22	34.01	16	29.91	23	42.89
fit2d	25	10500	129018	47	320.21	24	219.82	40	361.11
fit2p	3000	13525	50284	32	302.19	18	206.91	31	366.76
forplan	161	421	4563	40	16.71	21	10.71	30	15.20
ganges	1309	1681	6912	33	43.49	16	27.10	30	50.13
gfrdpnc	616	1092	2377	27	8.29	18	8.07	27	11.97
greenbea	2392	5405	30877	62	290.71	41	218.21	61	340.34
greenbeb	2392	5405	30877	70	287.16	33	167.31	60	301.41
grow15	300	645	5620	26	14.33	16	11.61	20	14.42
grow22	440	946	8252	29	23.73	16	17.44	21	22.40
grow7	140	301	2612	22	5.49	14	4.54	19	6.12
israel	174	142	2269	30	13.61	23	12.19	33	17.44
kb2	43	41	286	23	0.91	15	0.81	24	1.24
lotfi	153	308	1078	34	4.46	16	2.94	24	4.30
nesm	662	2923	13288	66	148.57	30	86.46	57	161.93

each step. Since our only difficulties with convergence occur through failures of linear algebra, we have not found this test to be necessary. However, this test may be required later if we have unexplained computational difficulties. Also, work is underway to create a stripped-down version of OB1 that allows the user to choose speed rather than safety if experience indicates this will be satisfactory for a given problem. Thus our results and Mehrotra's are totally compatible for an efficient algorithm restricted

TABLE 3
Computational results for OB1 (P–W).

Problem	No. of	No. of	No. of	Primal/Dual Meth.		Pred./Corr. Meth.		Dynamic μ	
Name	Rows	Cols.	Nonzeros	Its.	Sol. Time	Its.	Sol. Time	Its.	Sol. Time
perold	625	1376	6018	62	178.54	33	104.66	55	173.56
pilot4	410	1000	5141	56	71.72	36	54.63	50	78.63
pilot.ja	940	1988	14698	66	537.75	46	408.00	56	500.32
pilotnov	975	2172	13057	36	369.98	20	230.03	29	330.19
pilot	1441	3652	43167	67	3813.52	29	1784.35	63	3798.77
pilot.we	722	2789	9126	57	94.67	46	100.60	56	120.59
recipe	91	180	663	17	1.17	10	0.96	13	1.22
sc105	105	103	280	16	0.78	10	0.66	14	0.89
sc205	205	203	551	21	1.83	11	1.31	15	1.83
sc50a	50	48	130	14	0.42	10	0.37	13	0.47
sc50b	50	48	118	12	0.36	8	0.30	11	0.43
scagr25	471	500	1554	24	4.86	16	4.51	24	6.65
scagr7	129	140	420	20	1.16	12	0.87	18	1.40
scfxm1	330	457	2589	29	9.11	17	7.03	26	10.48
scfxm2	660	914	5183	36	23.30	19	16.07	32	26.73
scfxm3	990	1371	7777	37	36.27	20	25.70	33	41.74
scorpion	388	358	1426	21	3.25	14	2.94	21	4.32
scrs8	490	1169	3182	36	15.73	27	15.84	48	27.82
scsd1	77	760	2388	12	2.52	11	3.01	13	3.63
scsd6	147	1350	4316	15	5.67	12	6.10	17	8.46
scsd8	397	2750	8584	14	11.18	10	10.98	15	16.08
sctap1	300	480	1692	21	4.28	15	4.12	25	6.77
sctap2	1090	1880	6714	23	23.85	20	26.73	27	36.41
sctap3	1480	2480	8874	24	34.13	17	32.24	30	55.52
seba	515	1028	4352	28	15.38	19	36.88	29	21.51
share1b	117	225	1151	36	3.55	20	2.68	35	4.71
share2b	96	79	694	18	1.37	12	1.21	18	1.72
shell	536	1775	3556	31	13.67	21	12.96	32	19.58
ship04l	402	2118	6332	25	13.20	15	11.02	23	16.57
ship04s	402	1458	4352	24	8.62	15	7.47	22	10.72
ship08l	778	4283	12802	26	23.92	16	20.65	26	32.62
ship08s	778	2387	7114	25	12.02	14	9.50	24	15.82
ship12l	1151	5427	16170	29	35.35	18	30.58	28	46.53
ship12s	1151	2763	8178	30	17.52	18	14.71	28	22.34
sierra	1227	2036	7302	31	40.14	18	31.41	30	51.25
stair	356	467	3856	25	25.87	16	19.75	23	27.92
standata	359	1075	3031	18	6.13	15	7.04	24	11.05
standmps	467	1075	3679	28	12.19	24	13.97	34	19.60
stocfor1	117	111	447	18	1.09	19	1.46	21	1.63
stocfor2	2157	2031	8343	34	51.78	22	44.14	36	71.28
tuff	333	587	4520	45	30.13	19	15.69	31	25.05
vtp.base	198	203	908	24	1.17	13	0.91	15	1.09
wood1p	244	2594	70215	22	108.18	14	80.66	20	119.66
woodw	1098	8405	37474	35	156.05	20	108.83	34	181.96

to his simple test set.

The predictor–corrector algorithm introduces two new concepts, namely, the correction term and the dynamic choice of μ by (22). It is interesting to study the effects of each concept by eliminating the corrective term and comparing the primal–dual of §2 with a pure primal–dual using the μ given by (22). The results in Tables 2 and

3 (columns labeled "Dynamic μ") clearly demonstrate that the improvement in the algorithm is attributed largely to the correction term rather than the choice of μ. Nevertheless, we found that the predictor–corrector worked best with the algorithm for μ documented in §3, rather than the simpler μ of §2. Therefore, whereas the dynamic μ given by (22) has a largely negative effect on the primal–dual algorithm in terms of execution time, it significantly helps the predictor–corrector algorithm. For the version labeled "Dynamic μ," numerical difficulties prevented convergence to eight digits of accuracy on problems `greenbea` and `pilot.we`.

On the Silicon Graphics 4D/70, our tests strongly substantiated de Carvalho's conclusion [3] that minimum local fill-in generally outperforms multiple minimum degree. However, this result is very architecture dependent and is certainly invalid for vector architectures such as the CRAY Y-MP. Careful testing must be done on any specific architecture to determine the correct default algorithm.

In conclusion, the predictor–corrector algorithm of §3 implemented with the minimum local fill-in ordering is a substantial improvement over the primal–dual algorithm of [8]. Compared with other interior-point methods, the predictor–corrector algorithm is to date the most computationally efficient method for solving large-scale linear programs.

REFERENCES

[1] I. Adler, N. Karmarkar, M. G. C. Resende, and G. Veiga, *An implementation of Karmarkar's algorithm for linear programming*, Math. Programming, 44 (1989), pp. 297–325.

[2] D. Bayer and J. Lagarias, *The nonlinear geometry of linear programming* I. *Affine and projective scaling trajectories*, Trans. Amer. Math. Soc., 314 (1989), pp. 499–526.

[3] M. de Carvalho, *On the minimization of work needed to factor a symmetric positive definite matrix*, Tech. Report ORC 87-14, Department of Industrial Engineering and Operations Research, University of California, Berkeley, CA, 1987.

[4] I. C. Choi, C. L. Monma, and D. F. Shanno, *Further development of a primal–dual interior point method*, ORSA J. Comput., 2 (1990), pp. 304–311.

[5] P. D. Domich, P. T. Boggs, J. E. Rogers, and C. Witzgall, *Optimizing over* 3-*dimensional subspaces in an interior point method for linear programming*, Linear Algebra Appl., 152 (1989), pp. 315–342.

[6] D. M. Gay, *Electronic mail distribution of linear programming test problems*, Mathematical Programming Society COAL Newsletter, No. 13 (December 1985), pp. 10–12.

[7] G. H. Golub and C. F. Van Loan, *Matrix Computations*, The Johns Hopkins University Press, Baltimore, MD, 1983, pp. 182–183.

[8] I. J. Lustig, R. E. Marsten, and D. F. Shanno, *Computational experience with a primal-dual interior point method for linear programming*, Linear Algebra Appl., 152 (1989), pp. 191–222.

[9] S. Mehrotra, *On finding a vertex solution using interior point methods*, Linear Algebra Appl., 152 (1990), pp. 233–253.

[10] ———, *On the implementation of a (primal–dual) interior point method*, Tech. Report 90-03, Department of Industrial Engineering and Management Sciences, Northwestern University, Evanston, IL 1990.

[11] Y. Zhang, R. A. Tapia, and J. E. Dennis, *On the superlinear and quadratic convergence of primal-dual interior point linear programming algorithms*, Tech. Report 90-6, Department of Mathematical Sciences, Rice University, Houston, TX, 1990.

SIAM J. OPTIMIZATION
Vol. 2, No. 3, pp. 450–473, August 1992

AN INTERIOR-POINT ALGORITHM FOR LINEARLY CONSTRAINED OPTIMIZATION*

STEPHEN J. WRIGHT†

Abstract. This paper describes an algorithm for optimization of a smooth function subject to general linear constraints. An algorithm of the gradient projection class is used, with the important feature that the "projection" at each iteration is performed by using a primal–dual interior-point method for convex quadratic programming. Convergence properties can be maintained even if the projection is done inexactly in a well-defined way. Higher-order derivative information on the manifold defined by the apparently active constraints can be used to increase the rate of local convergence.

Key words. potential reduction algorithm, gradient projection algorithm, linearly constrained optimization

AMS(MOS) subject classifications. 65K10, 90C30

1. Introduction.

We address the problem

$$\min_x f(x) \quad \text{s.t. } A^T x \le b, \tag{1}$$

where $x \in R^n$ and $b \in R^m$ and f is assumed throughout to be twice continuously differentiable on the level set

$$\mathcal{L} = \{x \mid A^T x \le b,\ f(x) \le f(x^0)\},$$

where x^0 is some given initial choice for x. Recent literature on this problem can for the most part be divided into two main classes. On the one hand, there are the "active set" approaches, such as sequential quadratic programming, which are most suitable when the constraints $A^T x \le b$ lack any special structure, such as separability. In these algorithms a model of f (for example, the quadratic approximation $f(x) + \nabla f(x)^T d + (\frac{1}{2}) d^T \nabla^2 f(x) d$) is formed at each "outer" iteration and minimized over some subset of the feasible region. The algorithm tends to move along edges and faces of the boundary of the feasible set, changing its set of currently active constraints by at most one element on each "inner" iteration. A second class of methods, known as "gradient projection" methods, allow more substantial changes to the active set at each iteration by choosing a direction g (for example, $\nabla f(x)$ or some scaled version of it) and searching along the piecewise linear path $P(x - \alpha g)$, where $\alpha > 0$ and P is the projection onto the feasible set. Gradient projection methods are best suited to the case in which the projection $P(.)$ is easy to perform, for example, when the feasible region is a box whose sides are parallel to the principal coordinate axes.

In this paper, our aim is to describe an algorithm of the gradient projection class, in which we allow the projections to be performed *inexactly*. We focus on the case of Euclidean norm projections, which can be solved by using interior-point methods for convex quadratic programming problems. In this way, general polyhedral feasible regions can be handled. We thus hope to combine the much-vaunted advantages of interior-point methods with the desirable properties of gradient projection

[*] Received by the editors August 8, 1990; accepted for publication (in revised form) July 29, 1991. This research was partly performed at the University of Queensland, Australia, whose support is gratefully acknowledged. The work was also supported by the Applied Mathematical Sciences subprogram of the Office of Energy Research, U.S. Department of Energy, under contract W-31-109-Eng-38.

[†] Argonne National Laboratory, 9700 South Cass Avenue, Argonne, Illinois 60439.

algorithms—most notably, rapid identification of the final active set. In addition, we allow second-derivative information to be used in the definition of g (as is also done by Dunn [3], [4] and Gafni and Bertsekas [5]) to speed up the asymptotic convergence rate after the correct active set has been identified.

The "inexactness" in the projection is quantified by a duality gap, which is updated at each iteration of the projection subproblem. The global convergence analysis in §4 is not tied to the use of an interior-point method for the projection; any algorithm (including an active set method) that allows a duality gap to be calculated for each iterate may be used.

The point x^* is a critical point for (1) if there are scalars $y_i \geq 0$ such that

$$-\nabla f(x^*) = \sum_{i \in \mathcal{A}} y_i a_i,$$

where a_i are columns of A and

$$\mathcal{A} = \{i = 1, \cdots, m \mid a_i^T x^* = b_i\}.$$

Equivalently,

$$-\nabla f(x^*) \in N(x^*; X), \tag{2}$$

where X is the feasible set $\{x \mid A^T x \leq b\}$ and $N(x; X)$ is the normal cone to X at x defined by

$$N(x; X) = \{v \mid v^T(u - x) \leq 0, \text{ for all } u \in X\}.$$

In the next section, we specify the algorithm. The interior-point method that may be used to perform the projection is discussed in §3. The global and local convergence properties of the algorithm are analyzed in §§4 and 5, respectively.

In the remainder of the paper, the following notational conventions will be used:

- $\|x\| = (x^T x)^{\frac{1}{2}}$ (the Euclidean norm), unless otherwise specified.
- $P_Y(x)$ denotes the Euclidean projection of the vector x onto the convex set $Y \subset R^n$; that is,

  $$P_Y(x) = \arg\min_{z \in Y} \|z - x\|.$$

  If the subscript is omitted from P, projection onto X is assumed.
- int$Y$ denotes the interior of Y, and ∂Y denotes its boundary.
- When x is a vector, relations such as $x > 0$ are meant to apply componentwise.
- Subscripts on vectors and matrices denote components, while superscripts are used to distinguish different iterates. Subscripts on scalars denote iteration numbers.
- When $\{\xi_k\}$ and $\{\bar{\xi}_k\}$ are nonnegative sequences, the notation $\xi_k = O(\bar{\xi}_k)$ means that there is a constant s such that $\xi_k \leq s\bar{\xi}_k$ for all k sufficiently large. $\xi_k = o(\bar{\xi}_k)$ means that there is a nonnegative sequence $\{s_k\}$ converging to zero such that $\xi_k \leq s_k \bar{\xi}_k$ for all k sufficiently large.
- The sequence $\{v^k\}$ is said to converge Q-quadratically to v^* if $\|v^{k+1} - v^*\| = O(\|v^k - v^*\|^2)$. It is said to converge R-quadratically if there is a sequence $\{\xi_k\}$ that converges Q-quadratically to zero such that $\|v^k - v^*\| \leq \xi_k$ for all k.

- If $\{v^k\}$ and $\{\bar{v}^k\}$ are two sequences of vectors, the notation "$v^k \to \bar{v}^k$" means that $\lim_{k\to\infty} \|v^k - \bar{v}^k\| = 0$.
- In §§4 and 5, we introduce constants denoted by C and $\bar{C}$ with a subscript. In all cases these represent *strictly positive* constants, even where not stated explicitly.

2. The algorithm. We start this section by giving an outline of the major operations at each iteration of the basic algorithm. Then we state a formal outline and conclude by mentioning possible variations.

The algorithm first defines an "almost active" set of constraints at each iterate x^k. It partitions the gradient into two orthogonal components (which are orthogonal to and tangent to the manifold defined by the almost active set, respectively) and then scales the tangent component by a matrix with suitable positive-definiteness properties (possibly an inverse reduced Hessian or a quasi-Newton approximation to it). A projected Armijo-like line search is then performed along the resulting direction.

The activity tolerance at the point x^k is ϵ_k, where for the moment we require only that $\epsilon_k \geq 0$. The almost active set $\mathcal{I}^k$ is defined by

$$\mathcal{I}^k = \{i = 1, \cdots, m \mid a_i^T x^k \geq b_i - \epsilon_k \|a_i\|\}. \tag{3}$$

We use T^k to denote the tangent manifold corresponding to this set:

$$T^k = \{z \mid a_i^T z = 0, \text{ all } i \in \mathcal{I}^k\}. \tag{4}$$

The negative gradient is then decomposed by using T^k and setting

$$d^k = P_{T^k}(-\nabla f(x^k)), \qquad d^{k+} = -[\nabla f(x^k) + d^k]. \tag{5}$$

The tangent component d^k is modified by setting

$$\tilde{d}^k = D^k d^k, \tag{6}$$

where D^k is a matrix such that $P_{T^k} \circ D^k \circ P_{T^k} = D^k$ and

$$\lambda_1 z^T z \leq z^T D^k z \leq \lambda_2 z^T z, \quad \text{all } z \in T^k, \tag{7}$$

where λ_1 and λ_2 are positive constants. The search direction is assembled as

$$g^k = -(\tilde{d}^k + d^{k+}). \tag{8}$$

A projected Armijo search is carried out along the path

$$x^k(\alpha) = P(x^k - \alpha g^k),$$

where the values $\alpha = 1, \beta, \beta^2, \beta^3, \cdots$ ($\beta \in (0,1)$ is some constant) are tried. For each such value of α, the projection is calculated with the algorithm described in the next section. This algorithm generates a sequence of *feasible* approximations to $x^k(\alpha)$, which we denote by $x^{kj}(\alpha)$. For each such estimate, the algorithm produces a duality gap $\gamma_{kj}(\alpha)$. Defining the more convenient quantity

$$\delta_{kj}(\alpha) = \sqrt{2\gamma_{kj}(\alpha)},$$

we can obtain upper and lower bounds on the distance from $x^k - \alpha g^k$ to X; that is,

$$\|x^{kj}(\alpha) - (x^k - \alpha g^k)\|^2 - \delta_{kj}(\alpha)^2 \leq \|x^k(\alpha) - (x^k - \alpha g^k)\|^2 \leq \|x^{kj}(\alpha) - (x^k - \alpha g^k)\|^2.$$

These "inner iterations" are stopped at a value of j for which $\delta_{kj}(\alpha)$ becomes sufficiently small according to the following criteria:

$$\delta_{kj}(\alpha) \le \eta\alpha^{\tau/2} \max\left(\frac{\|x^{kj}(\alpha) - (x^k + \alpha\tilde{d}^k)\|}{\alpha}, \|d^k\|\right)^2 \tag{9}$$

and

$$\delta_{kj}(\alpha) \le C_1\alpha^{\tau/2}. \tag{10}$$

Here τ, C_1, and η are constants that satisfy the conditions

$$\tau > 2, \qquad \eta C_1 < 1.$$

We denote the final computed $\delta_{kj}(\alpha)$ by $\delta_k(\alpha)$, and we denote the corresponding $x^{kj}(\alpha)$ by $x^k(\alpha;\delta_k(\alpha))$. The step α is then accepted if the following "sufficient decrease" test is satisfied:

$$f(x^k) - f(x^k(\alpha;\delta_k(\alpha))) \ge \sigma\left\{\alpha d^{kT}D^kd^k + \frac{\|x^k(\alpha;\delta_k(\alpha)) - (x^k + \alpha\tilde{d}^k)\|^2}{\alpha}\right\}, \tag{11}$$

where $\sigma \in (0,1)$ is a constant.

The algorithm can be summarized as follows:

Step 1: Choose ϵ_k. Compute $\mathcal{I}^k$ from (3), and compute g^k according to (5)–(8).

Step 2: For $\alpha = \beta^p$, $p = 0, 1, 2, \cdots$ (in sequence) approximately calculate $x^k(\alpha) = P(x^k - \alpha g^k)$, terminating when $x^k(\alpha;\delta_k(\alpha)) = x^{kj}(\alpha)$ and its associated $\delta_k(\alpha) = \delta_{kj}(\alpha)$ are found that satisfy (9), (10). If the test (11) is passed for this value of α, set $\alpha_k = \alpha = \beta^p$, $x^{k+1} = x^k(\alpha;\delta_k(\alpha))$, $k \leftarrow k+1$, and go to the next iteration. Otherwise, increase p by 1, and try the next $\alpha = \beta^p$.

In its "exact" form (i.e., $\delta_k(\alpha) \equiv 0$), and when D^k is defined as the reduced Hessian or a quasi-Newton approximation to it, the step g^k is the same as that obtained by specializing the algorithm of Dunn [4] to the linearly constrained case. The calculation of g^k is somewhat different in Gafni and Bertsekas [5]. They define an "almost tangent cone" at x^k by

$$C^k = \{z \mid a_i^T z \le 0, \text{ all } i \in \mathcal{I}^k\}$$

and then define d^k as the projection of $-\nabla f(x^k)$ onto this cone. Additionally, the conditions on D^k are slightly different, and $\tilde{d}^k$ is the projection of D^kd^k onto C^k. Our reason for following Dunn [4] and using the simpler decomposition relative to T^k is our assumption that projection onto the subspace T^k can be done exactly and cheaply. This is not unreasonable—the cost would normally be comparable to one iteration of the interior-point algorithm used for the projection onto X. Projection onto C^k may, on the other hand, be as expensive as projection onto X. Still, there are intuitive reasons for preferring C^k to T^k, and it would be of interest to see whether the extra cost per iteration (and the extra algorithmic complexity of doing the projection onto C^k inexactly) is justified.

The step-length rule (11) reduces to the one proposed by Gafni and Bertsekas [5] (and also used by Dunn [3]) when $\delta_k(\alpha) \equiv 0$. Another obvious possibility, to which we will return briefly in §5, is

$$(12)\qquad \begin{aligned} &f(x^k) - f(x^k(\alpha;\delta_k(\alpha))) \\ &\quad \geq \sigma\left\{\alpha d^{kT} D^k d^k + \nabla f(x^k)^T[x^k + \alpha\tilde{d}^k - x^k(\alpha;\delta_k(\alpha))]\right\}. \end{aligned}$$

3. Projection onto X. Projection onto the polyhedral set X can be achieved by solving a convex quadratic program or, equivalently, a linear complementarity problem (LCP). In this section we formulate the problem and outline a primal–dual potential reduction algorithm for solving it. The discussion will be brief, since other papers, such as [6], [7], [10], and [11], can be consulted for details about motivation, analysis, and implementation issues for this class of interior-point algorithms.

Throughout the remainder of the paper, we use the following assumptions:

(A) The feasible set X has an interior in R^n.

(B) At the solution $z^* = P(t)$ of the projection subproblem, the set of vectors

$$\{a_i \mid a_i^T z^* = b_i\}$$

is linearly independent.

The (unique) vector $P(t)$ is obtained by solving

$$\min \tfrac{1}{2}\|z - t\|^2 \quad \text{s.t. } A^T z \leq b,$$

or, equivalently,

$$(13)\qquad \min \tfrac{1}{2}\|z - t\|^2 \quad \text{s.t. } A^T z + \nu = b, \quad \nu \geq 0.$$

Introducing Lagrange multipliers y for the constraints, we find that (13) is equivalent to the (mixed) LCP

$$(14)\qquad \begin{bmatrix} 0 \\ \nu \end{bmatrix} = \begin{bmatrix} I & A \\ -A^T & 0 \end{bmatrix} \begin{bmatrix} z \\ y \end{bmatrix} + \begin{bmatrix} -t \\ b \end{bmatrix}, \quad \nu \geq 0, \ \ y \geq 0, \ \ \nu^T y = 0.$$

The coefficient matrix in (14) is clearly positive semidefinite.

The progress of the interior-point algorithm can be gauged by using the potential function defined by

$$(15)\qquad \psi(\nu, y) = \rho_P \log(\nu^T y) - \sum_{i=1}^{m} \log(\nu_i y_i),$$

where $\rho_P \geq m + \sqrt{m}$. In Kojima, Mizuno, and Yoshise [7], the step from iterate j to iterate $j+1$ is obtained by solving the linear system

$$(16)\qquad \begin{bmatrix} 0 \\ \Delta\nu \end{bmatrix} = \begin{bmatrix} I & A \\ -A^T & 0 \end{bmatrix} \begin{bmatrix} \Delta z \\ \Delta y \end{bmatrix},$$

together with

$$(17)\qquad \nu_i^j \Delta y_i^j + y_i^j \Delta\nu_i = \left(\frac{-\rho_j}{1-\rho_j}\right)\left[-\nu_i^j y_i^j + \frac{\gamma_j}{\rho_j}\right], \qquad i = 1, \cdots, m,$$

where $\gamma_j = \sum_{i=1}^m \nu_i^j y_i^j$ and the value of ρ_j is as discussed below. A step length θ_j is chosen such that

$$(18)\qquad \theta_j \left|\frac{\Delta\nu_i}{\nu_i^j}\right| \leq \tau, \quad \theta_j \left|\frac{\Delta\nu_i}{\nu_i^j}\right| \leq \tau, \quad i = 1, \cdots, m$$

for some $\tau \in (0,1)$. Trivial modifications of the results of Kojima, Mizuno, and Yoshise [7] indicate that for the choices $\rho_j \equiv \rho_P = m + \sqrt{m}$ and $\tau = 0.4$, we have that

$$\psi(\nu^j + \theta_j \Delta\nu, y^j + \theta_j \Delta y) \leq \psi(\nu^j, y^j) - 0.2. \tag{19}$$

When some iterate (z^j, ν^j, y^j) satisfies $\psi(\nu^j, y^j) \leq -O(\sqrt{m}L)$, it can easily be shown that $(\nu^j)^T y^j \leq 2^{-O(L)}$. This suggests that, provided the initial point (z^0, ν^0, y^0) satisfies $\psi(\nu^0, y^0) = O(\sqrt{m}L)$, convergence to a point with duality gap less than $2^{-O(L)}$ can be achieved in $O(\sqrt{m}L)$ iterations. (For purposes of the complexity analysis, L is taken to be the "size" of the problem.) Although this choice of ρ_j yields the best complexity result to date, it has been observed that, in practice, larger values of ρ_j lead to fewer iterations. In Han, Pardalos, and Ye [6], the choice $\rho_j \equiv m^2$ is made for convex quadratic programs. In the context of linear programs, Zhang, Dennis, and Tapia [11] observe that it is even desirable to let ρ_j grow unboundedly large as the solution is approached. (The steps produced by (16), (17) are then very close to being Newton steps for the nonlinear equations formed by the equalities in (14).) Ye, Kortanek, Kaliski, and Huang [10] have shown that such "large" choices of ρ_j are not incompatible with obtaining reductions in the potential function. In practical implementations, the line-search parameter θ_j is also chosen differently. In Han, Pardalos, and Ye [6], the following choice appeared to give good experimental results:

$$\theta_j = 0.99 \min \left(\min_{i=1,\cdots,m,\ \Delta\nu_i<0} -\frac{\nu_i^j}{\Delta\nu_i}, \min_{i=1,\cdots,m,\ \Delta y_i<0} -\frac{y_i^j}{\Delta y_i} \right).$$

An issue of particular concern in this context is the choice of a feasible initial point at which to start the interior-point iteration. Such a point can be found by augmenting the problem in a simple way. We can reasonably assume that a vector z^0 that satisfies $A^T z^0 < b$ is available from some previous iteration. If y^0 is also chosen from a previous iteration, we usually have, from the first equation in (14), that $z^0 + Ay^0$ is similar in magnitude to the primal quantities z and t. We can thus define a (reasonably scaled) vector q by

$$q = -(z^0 - t + Ay^0)$$

and obtain the following augmented version of (14):

$$\begin{bmatrix} 0 \\ \nu \\ \nu_{m+1} \end{bmatrix} = \begin{bmatrix} I & A & q \\ -A^T & 0 & 0 \\ -q^T & 0 & 0 \end{bmatrix} \begin{bmatrix} z \\ y \\ y_{m+1} \end{bmatrix} + \begin{bmatrix} -t \\ b \\ b_{m+1} \end{bmatrix}, \tag{20}$$
$$\nu \geq 0, \quad \nu_{m+1} \geq 0, \quad y \geq 0, \quad y_{m+1} \geq 0, \quad \nu^T y + \nu_{m+1} y_{m+1} = 0.$$

The corresponding projection problem is

$$\min \tfrac{1}{2}\|z - t\|^2 \quad \text{s.t. } A^T z \leq b, \quad q^T z \leq b_{m+1}. \tag{21}$$

If we choose b_{m+1} to satisfy

$$b_{m+1} > \max(q^T z^0, q^T P(t)),$$

then we find that a feasible initial point for (20) is

$$(z, \nu, \nu_{m+1}, y, y_{m+1}) = (z^0, b - A^T z^0, b_{m+1} - q^T z^0, y^0, 1).$$

At the optimal solution, $\nu_{m+1}^* = b_{m+1} - q^T P(t)$ and $y_{m+1}^* = 0$. A practical choice for b_{m+1} can be made as follows: When $t = x - \alpha g$ with x feasible, note that

$$\begin{aligned} q^T P(t) &= q^T[P(x-\alpha g) - (x - \alpha g)] + q^T t \\ &\le \|q\| \|P(x-\alpha g) - (x-\alpha g)\| + q^T t \le \alpha \|q\| \|g\| + q^T t. \end{aligned}$$

Hence b_{m+1} can be chosen as any number greater than

$$\max(q^T z^0, \alpha\|q\|\|g\| + q^T t).$$

We tacitly assume throughout the remainder of the paper that b_{m+1} is chosen large enough that the extra constraint in (21) does not come into play during the projection process (that is, ν_{m+1} stays reasonably large).

Two more points about the computational aspects of the projection should be made since, for many variants of the algorithm described in this paper, it will be the most time-consuming step, apart from the function evaluations. First, note that the cost per interior-point iteration, which is dominated by the cost of solving augmented versions of the linear system (16), (17), is similar to the cost of decomposing the gradient, as in (5). (The latter operation may be performed by solving a system whose coefficient matrix is a submatrix of the matrix in (16).) Second, the number of interior-point iterates that will be necessary for a given α should not be too large. A rule of thumb seems to be that 20–30 iterates are required for an accurate solution when no a priori information about the solution is known. In our case, the situation is better: Good starting points will usually be available from previous iterates and from approximate projections for larger values of α. A priori information has been observed to significantly decrease the number of interior-point iterations (see, for example, [9]).

In §5, we assume that the points $(z^j, \nu^j, \nu_{m+1}^j, y^j, y_{m+1}^j)$ generated by the interior-point algorithm do not stray too far from the central path defined by

$$\left\{ (z, \nu, \nu_{m+1}, y, y_{m+1}) \text{ feasible in (20)} \mid \nu_i y_i = \sum_{l=1}^{m+1} \nu_l y_l/(m+1),\ i = 1, \cdots, m+1 \right\}.$$

The following assumption is used to prove that unit steps $\alpha_k = 1$ are always eventually used by the method.

(C) There is a constant $\mu > 1$ such that the final iterate $(z, \nu, \nu_{m+1}, y, y_{m+1})$ generated by the projection algorithm, each time it is called, satisfies

$$\nu_i y_i \ge \frac{\sum_{l=1}^{m+1} \nu_l y_l/(m+1)}{\mu}.$$

Although this assumption conflicts to some extent with the desire for fast asymptotic convergence of the interior-point method, Zhang, Dennis, and Tapia [11, Thm. 3.1] observed that, at least in the case of linear programming that they considered, it appears to hold in practice.

4. Global convergence.

In this section we prove that all accumulation points of the algorithm of §2 are critical. The result depends crucially on the following lemma, which bounds the distance between $x^k(\alpha; 0)$ and $x^k(\alpha; \delta_k(\alpha))$ in terms of $\delta_k(\alpha)$.

LEMMA 4.1. *Suppose that* (A) *holds and that* (B) *holds at* $z^* = x^k(\alpha, 0)$. *Then*

$$\|x^k(\alpha; 0) - x^k(\alpha; \delta_k(\alpha))\| \le \delta_k(\alpha).$$

Proof. Setting $t = x^k - \alpha g^k$, we obtain

$$\begin{aligned}\|t - x^k(\alpha;0)\|^2 &\geq \|t - x^k(\alpha;\delta_k(\alpha))\|^2 - \delta_k(\alpha)^2\\ \Rightarrow \delta_k(\alpha)^2 &\geq 2[t - x^k(\alpha;0)]^T[x^k(\alpha;0) - x^k(\alpha;\delta_k(\alpha))] + \|x^k(\alpha;0) - x^k(\alpha;\delta_k(\alpha))\|^2.\end{aligned}$$

Now, since $t - x^k(\alpha;0) \in N(x^k(\alpha;0)\ ;\ X)$ and $x^k(\alpha;\delta_k(\alpha)) \in X$, the first term on the right-hand side above is nonnegative and can be omitted from the inequality. The result follows. □

Under appropriate nondegeneracy assumptions, application of the implicit function theorem to a subset of the equalities in (13) (or (20)) would suggest that, locally, a stronger bound of $O(\gamma_k(\alpha)) = O(\delta_k(\alpha)^2)$ might be obtained. In fact, some of the local convergence analysis in §5 relies on just this observation. In general, however, given a point x^k and a search direction g^k, there are usually values of α such that the solution of (13) (or (21)) for $t = x^k - \alpha g^k$ is degenerate. Our result in Lemma 4.1 is similar to, but more specific than, the bound that would be obtained by applying the analysis of Mangasarian and Shiau [8] to (13).

We state without proof the following well-known result, which actually applies for any closed convex $X \subset R^n$.

LEMMA 4.2. *For any $x \in X$ and $z \in R^n$,*

(a) $\|P(x + \alpha z) - x\|/\alpha$ *is a nonincreasing function of* $\alpha > 0$,

(b) $\|P(x + \alpha z) - x\|/\alpha \leq \|z\|$.

Before proving the main result (Theorem 4.5), we show that the conditions (9), (10) ensure that the projection is computed exactly when x^k is critical (Lemma 4.3) and, in a technical result, show that the algorithm produces descent at a noncritical point (Lemma 4.4).

LEMMA 4.3. *Suppose that* (A) *holds and that* (B) *holds at* $z^* = x^k$. *When* x^k *is critical, then* $\delta_k(\alpha) = 0$ *for all* $\alpha \in [0,1]$ *and* $x^k(\alpha;\delta_k(\alpha)) = x^k$ *for all* $\alpha \in [0,1]$.

Proof. Clearly the result is true for $\alpha = 0$. For the remainder of the proof, we assume that $\alpha \in (0,1]$.

All vectors in the subspace T^k are orthogonal to $N(x^k;X)$. Hence by (2) and (5), $d^k = \tilde{d}^k = 0$ and $d^{k+} = -\nabla f(x^k)$. Also, by (2),

$$x^k(\alpha,0) = P(x^k - \alpha\nabla f(x^k)) = x^k,$$

and so

$$\|x^k(\alpha;\delta_k(\alpha)) - (x^k + \alpha\tilde{d}^k)\| \leq \|x^k(\alpha;0) - x^k\| + \|x^k(\alpha;\delta_k(\alpha)) - x^k(\alpha;0)\| \leq \delta_k(\alpha).$$

Substituting this expression in (9), we have

$$\delta_k(\alpha) \leq \eta\alpha^{\tau/2}\frac{\delta_k(\alpha)^2}{\alpha^2},$$

and hence

$$\delta_k(\alpha)[1 - \eta\alpha^{\tau/2-2}\delta_k(\alpha)] \leq 0. \tag{22}$$

From (10) and the fact that $\eta C_1 < 1$,

$$1 - \eta\alpha^{\tau/2-2}\delta_k(\alpha) \geq 1 - \eta C_1\alpha^{\tau-2} > 1 - \alpha^{\tau-2} \geq 0,$$

since $\alpha \in [0,1]$ and $\tau > 2$. Since $\delta_k(\alpha) \geq 0$, the inequality (22) can hold only if $\delta_k(\alpha) = 0$. Thus, the first statement is proved. Proof of the second statement follows immediately. □

LEMMA 4.4. *Suppose $\mathcal{I}^k$ is defined by* (3), *where ϵ_k is any positive number. Suppose that* (A) *holds and that* (B) *holds for all $z^* = x^k(\alpha, 0)$ for $\alpha \in [0,1]$. Then, given any $\bar{\sigma} \in (0,1)$, there exists an $\bar{\alpha}(\bar{\sigma}) \in (0, \epsilon_k/\|g^k\|)$ such that*

$$\begin{aligned}(23)\qquad &\nabla f(x^k)^T[x^k - x^k(\alpha;\delta_k(\alpha))]\\ &\geq \bar{\sigma}\left[\alpha d^{kT}D^k d^k + \frac{1}{\alpha}\|x^k(\alpha;\delta_k(\alpha)) - (x^k + \alpha\tilde{d}^k)\|^2\right].\end{aligned}$$

Hence, provided x^k is not critical, there is an $\hat{\alpha}(\bar{\sigma}) \in (0, \bar{\alpha}]$ such that $f(x^k) > f(x^k(\alpha, \delta_k(\alpha)))$ for all $\alpha \in (0, \hat{\alpha}]$.

Proof.

$$\begin{aligned}(24)\qquad &\nabla f(x^k)^T[x^k - x^k(\alpha;\delta_k(\alpha))]\\ &= \nabla f(x^k)^T[x^k - x^k(\alpha;0)] + \nabla f(x^k)^T[x^k(\alpha;0) - x^k(\alpha;\delta_k(\alpha))],\end{aligned}$$

and for $\alpha \in (0, \epsilon_k/\|g^k\|)$ it can be proved by using an argument similar to that in [5, Prop. 1(b)]:

$$\nabla f(x^k)^T[x^k - x^k(\alpha,0)] \geq \alpha d^{kT}D^k d^k + \frac{1}{\alpha}\|x^k(\alpha,0) - (x^k + \alpha\tilde{d}^k)\|^2.$$

By the smoothness assumptions on f, there is a constant B such that

$$\|\nabla f(x)\| \leq B \quad \text{for all } x \in \mathcal{L}.$$

Since all $x^k \in \mathcal{L}$, we have, using Lemma 4.1, that

$$(25)\qquad \left|\nabla f(x^k)^T[x^k(\alpha;0) - x^k(\alpha;\delta_k(\alpha))]\right| \leq B\delta_k(\alpha).$$

Now

$$\begin{aligned}(26)\ &\|x^k(\alpha;0) - (x^k + \alpha\tilde{d}^k)\|^2\\ &= \|x^k(\alpha;\delta_k(\alpha)) - (x^k + \alpha\tilde{d}^k)\|^2\\ &\quad+2[x^k(\alpha;0) - x^k(\alpha;\delta_k(\alpha))]^T[x^k(\alpha;\delta_k(\alpha)) - (x^k + \alpha\tilde{d}^k)]\\ &\quad+\|x^k(\alpha;0) - x^k(\alpha;\delta_k(\alpha))\|^2\\ &\geq \|x^k(\alpha;\delta_k(\alpha)) - (x^k + \alpha\tilde{d}^k)\|^2 - 2\delta_k(\alpha)\|x^k(\alpha;\delta_k(\alpha)) - (x^k + \alpha\tilde{d}^k)\|,\end{aligned}$$

and so from (24)–(26)

$$\begin{aligned}(27)\qquad &\\ \nabla f(x^k)^T[x^k - x^k(\alpha;\delta^k(\alpha))] &\geq \alpha d^{kT}D^k d^k + \frac{1}{\alpha}\|x^k(\alpha;\delta_k(\alpha)) - (x^k + \alpha\tilde{d}^k)\|^2\\ &\quad-\frac{2}{\alpha}\delta_k(\alpha)\|x^k(\alpha;\delta_k(\alpha)) - (x^k + \alpha\tilde{d}^k)\| - B\delta_k(\alpha).\end{aligned}$$

Now,

$$\begin{aligned}(28)\qquad &\frac{1}{\alpha}\|x^k(\alpha;\delta_k(\alpha)) - (x^k + \alpha\tilde{d}^k)\|\\ &\leq \frac{1}{\alpha}\|x^k(\alpha;\delta_k(\alpha)) - x^k(\alpha;0)\| + \frac{1}{\alpha}\|x^k(\alpha;0) - (x^k + \alpha\tilde{d}^k)\|\\ &\leq \frac{1}{\alpha}\delta_k(\alpha) + \frac{1}{\alpha}\|P(x^k + \alpha d^{k+} + \alpha\tilde{d}^k) - (x^k + \alpha\tilde{d}^k)\|.\end{aligned}$$

The following simple argument shows that $x^k + \alpha\tilde{d}^k \in X$ for $\alpha \in (0, \epsilon_k/\|g^k\|)$:

$$\begin{aligned} i \notin \mathcal{I}^k &\Rightarrow a_i^T[x^k + \alpha\tilde{d}^k] \leq b_i - \epsilon_k\|a_i\| + \alpha\|g^k\|\|a_i\| < b_i, \\ i \in \mathcal{I}^k &\Rightarrow a_i^T[x^k + \alpha\tilde{d}^k] = a_i^T x^k \leq b_i. \end{aligned}$$

Hence, by Lemma 4.2(b), (28) becomes

$$\text{(29)} \qquad \frac{1}{\alpha}\|x^k(\alpha;\delta_k(\alpha)) - (x^k + \alpha\tilde{d}^k)\| \leq \frac{1}{\alpha}\delta_k(\alpha) + \|d^{k+}\| \leq \frac{1}{\alpha}\delta_k(\alpha) + B.$$

Hence, (27) becomes

$$\text{(30)} \qquad \begin{aligned} &\nabla f(x^k)^T[x^k - x^k(\alpha;\delta_k(\alpha))] \\ &\geq \alpha d^{kT}D^k d^k + \frac{1}{\alpha}\|x^k(\alpha;\delta_k(\alpha)) - (x^k + \alpha\tilde{d}^k)\|^2 - \frac{2}{\alpha}\delta_k(\alpha)^2 - 3B\delta_k(\alpha). \end{aligned}$$

We now consider two cases. First, suppose that

$$\frac{1}{\alpha}\|x^k(\alpha;\delta_k(\alpha)) - (x^k + \alpha\tilde{d}^k)\| \geq \|d^k\|.$$

Then from (9) it follows that

$$\delta_k(\alpha) \leq \eta\alpha^{\tau/2}\frac{\|x^k(\alpha;\delta_k(\alpha)) - (x^k + \alpha\tilde{d}^k)\|^2}{\alpha^2}.$$

Using this, together with (10) and the fact that $\eta C_1 < 1$, we have from (30) that

$$\text{(31)} \qquad \begin{aligned} &\nabla f(x^k)^T[x^k - x^k(\alpha;\delta_k(\alpha))] \\ &\geq \alpha d^{kT}D^k d^k + \frac{1}{\alpha}\|x^k(\alpha;\delta_k(\alpha)) - (x^k + \alpha\tilde{d}^k)\|^2 \\ &\quad - \left[\frac{2}{\alpha}(C_1\alpha^{\tau/2})\eta\alpha^{\tau/2} + 3B\eta\alpha^{\tau/2}\right]\left(\frac{1}{\alpha^2}\right)\|x^k(\alpha;\delta_k(\alpha)) - (x^k + \alpha\tilde{d}^k)\|^2 \\ &\geq \alpha d^{kT}D^k d^k + \frac{1}{\alpha}\left[1 - 2\alpha^{\tau-2} - 3B\eta\alpha^{(\tau-2)/2}\right]\|x^k(\alpha;\delta_k(\alpha)) - (x^k + \alpha\tilde{d}^k)\|^2. \end{aligned}$$

The inequality (23) will be satisfied provided

$$\text{(32)} \qquad 1 - 2\alpha^{\tau-2} - 3B\eta\alpha^{(\tau-2)/2} \geq \bar{\sigma}.$$

Setting $\beta = \alpha^{(\tau-2)/2}$, we find that the quadratic $2\beta^2 + (3B\eta)\beta + (\bar{\sigma} - 1)$ has one positive root. Hence we can find an $\bar{\alpha}_1 > 0$ such that the required inequality will be satisfied for all $\alpha \in (0, \bar{\alpha}_1]$.

For the second case, assume that

$$\frac{1}{\alpha}\|x^k(\alpha;\delta_k(\alpha)) - (x^k + \alpha\tilde{d}^k)\| \leq \|d^k\|.$$

Then from (9),

$$\delta_k(\alpha) \leq \eta\alpha^{\tau/2}\|d^k\|^2,$$

and so from (30),

$$\text{(33)} \qquad \begin{aligned} &\nabla f(x^k)^T[x^k - x^k(\alpha;\delta_k(\alpha))] \\ &\geq \alpha d^{kT}D^k d^k + \frac{1}{\alpha}\|x^k(\alpha;\delta_k(\alpha)) - (x^k + \alpha\tilde{d}^k)\|^2 \\ &\quad -\frac{2}{\alpha}\eta C_1\alpha^{\tau}\|d^k\|^2 - 3B\eta\alpha^{\tau/2}\|d^k\|^2. \end{aligned}$$

From (7) it follows that

$$\|d^k\|^2 \le \frac{1}{\lambda_1} d^{kT} D^k d^k,$$

and so, using $\eta C_1 < 1$, we have

$$\begin{aligned} &\nabla f(x^k)^T [x^k - x^k(\alpha; \delta_k(\alpha))] \\ &\quad \ge \alpha \left[1 - \frac{2}{\lambda_1}\alpha^{\tau-2} - \frac{3B\eta}{\lambda_1}\alpha^{(\tau-2)/2} \right] d^{kT} D^k d^k \\ &\quad + \frac{1}{\alpha}\|x^k(\alpha; \delta_k(\alpha)) - (x^k + \alpha\tilde{d}^k)\|^2. \end{aligned} \tag{34}$$

For (23), it is sufficient that

$$1 - \frac{2}{\lambda_1}\alpha^{\tau-2} - \frac{3B\eta}{\lambda_1}\alpha^{(\tau-2)/2} \ge \bar{\sigma}.$$

An argument similar to that above shows that a positive value $\bar{\alpha}_2$ can be found so that this inequality is satisfied for $\alpha \in (0, \bar{\alpha}_2]$. Hence, the first part of the result follows by setting

$$\bar{\alpha}(\bar{\sigma}) = \min(1, \epsilon_k/(2\|g^k\|), \bar{\alpha}_1, \bar{\alpha}_2).$$

The second part of the result, i.e., that $f(x^k) > f(x^k(\alpha, \delta_k(\alpha)))$ for sufficiently small α, is obtained by modifying the argument of Gafni and Bertsekas [5, Prop. 1(b)]. By the mean value theorem, we can find a point $\zeta^k(\alpha)$ on the line joining x^k to $x^k(\alpha, \delta_k(\alpha))$ such that

$$f(x^k) - f(x^k(\alpha; \delta_k(\alpha))) = \nabla f(\zeta^k(\alpha))^T [x^k - x^k(\alpha; \delta_k(\alpha))].$$

Hence, from (23), for $\alpha \in (0, \bar{\alpha})$,

$$\begin{aligned} &\frac{1}{\alpha}[f(x^k) - f(x^k(\alpha, \delta_k(\alpha)))] \\ &\quad \ge \bar{\sigma}\left[d^{kT} D^k d^k + \frac{1}{\alpha^2}\|x^k(\alpha; \delta_k(\alpha)) - (x^k + \alpha\tilde{d}^k)\|^2 \right] \\ &\quad + \frac{1}{\alpha}[\nabla f(\zeta^k(\alpha)) - \nabla f(x^k)]^T [x^k - x^k(\alpha; \delta_k(\alpha))]. \end{aligned} \tag{35}$$

Again, writing

$$x^k - x^k(\alpha; \delta_k(\alpha)) = x^k - x^k(\alpha; 0) + x^k(\alpha; 0) - x^k(\alpha; \delta_k(\alpha))$$

and using

$$\begin{aligned} \frac{\|x^k - x^k(\alpha; 0)\|^2}{\alpha^2} &\le \|g^k\|^2 = \|\tilde{d}^k\|^2 + \|d^{k+}\|^2 \\ &\le (\lambda_2 + 1)\|\nabla f(x^k)\|^2 \le (\lambda_2 + 1)B^2, \end{aligned} \tag{36}$$

we have

$$\begin{aligned} &\frac{1}{\alpha}[\nabla f(\zeta^k(\alpha)) - \nabla f(x^k)]^T [x^k - x^k(\alpha; \delta_k(\alpha))] \\ &\quad \ge -\|\nabla f(\zeta^k(\alpha)) - \nabla f(x^k)\| \left[B\sqrt{\lambda_2 + 1} + \frac{1}{\alpha}\delta_k(\alpha) \right] \\ &\quad \ge -\|\nabla f(\zeta^k(\alpha)) - \nabla f(x^k)\| \left[B\sqrt{\lambda_2 + 1} + C_1\alpha^{\tau/2-1} \right] = O(\alpha). \end{aligned}$$

When $d^k \neq 0$, it follows from (35) that

$$\lim_{\alpha \to 0} \frac{f(x^k) - f(x^k(\alpha; \delta_k(\alpha)))}{\alpha} \geq \bar{\sigma} d^{kT} D^k d^k > 0.$$

On the other hand, when $d^k = 0$,

$$\begin{aligned} &\frac{f(x^k) - f(x^k(\alpha; \delta_k(\alpha)))}{\alpha} \\ &\geq \frac{\bar{\sigma}}{\alpha^2} \|x^k(\alpha; \delta_k(\alpha)) - x^k\|^2 + O(\alpha) \\ &\geq \frac{\bar{\sigma}}{\alpha^2} \|x^k(\alpha; 0) - x^k\|^2 - \frac{2\bar{\sigma}}{\alpha^2} \delta_k(\alpha) \|x^k(\alpha; 0) - x^k\| + O(\alpha). \end{aligned} \tag{37}$$

A straightforward application of Lemma 4.2(a) shows that

$$\frac{1}{\alpha} \|x^k(\alpha; 0) - x^k\| \geq \|x^k(1, 0) - x^k\|.$$

Also, from Lemma 4.2(b), we have for $\alpha \in (0, \bar{\alpha})$ that

$$\|x^k(\alpha; 0) - x^k\| \leq \alpha \|d^{k+}\| \leq \alpha B.$$

Using these inequalities, together with (10), we have from (37) that

$$\begin{aligned} \frac{f(x^k) - f(x^k(\alpha; \delta_k(\alpha)))}{\alpha} &\geq \bar{\sigma} \|x^k(1; 0) - x^k\|^2 - \frac{2\bar{\sigma} B}{\alpha} \delta_k(\alpha) + O(\alpha) \\ &\geq \bar{\sigma} \|x^k(1; 0) - x^k\|^2 - 2\bar{\sigma} B C_1 \alpha^{\tau/2 - 1} + O(\alpha). \end{aligned}$$

Taking the limit, we have

$$\lim_{\alpha \to 0} \frac{f(x^k) - f(x^k(\alpha; \delta_k(\alpha)))}{\alpha} \geq \bar{\sigma} \|x^k(1; 0) - x^k\|^2 > 0.$$

In either case, there is an $\hat{\alpha} \leq \bar{\alpha}$ with the desired property. □

For the main result of this section, we need to be more specific about the choice of ϵ_k. We now assume that

$$\epsilon_k = \min(\epsilon, \hat{c}_k \hat{\epsilon}(x_k)), \tag{38}$$

where there is a constant $\hat{B}$ such that $\hat{c}_k \in [1, \hat{B}]$, and $\hat{\epsilon}(x)$ is a continuous function of x that is zero only when x is critical.

THEOREM 4.5. *Suppose that ϵ_k satisfies condition* (38), *that* (A) *holds, and that* (B) *holds for $x^k(\alpha, 0)$, for all $\alpha \in [0, 1]$ and all k sufficiently large. Then every accumulation point x^k generated by the algorithm is critical.*

Proof. The proof is quite similar to the proof of Proposition 2 of Gafni and Bertsekas [5]. Some modifications are necessary because of the inexactness in $x^k(\alpha)$ and because of the need for the quantity $\bar{\sigma}$ in Lemma 4.4. We include most of the details here, and refer the reader to [5] for the remainder.

Suppose for contradiction that there is a noncritical point x^* and a subsequence $\mathcal{K}$ such that $\lim_{k \in \mathcal{K}} x^k = x^*$. If α_k denotes the step length used in the step from x^k to x^{k+1}, (11) implies that

$$\lim_{k \in \mathcal{K}} \alpha_k d^{kT} D^k d^k = 0, \tag{39}$$

$$\lim_{k\in\mathcal{K}} \frac{1}{\alpha_k}\|x^k(\alpha_k;\delta_k(\alpha_k)) - (x^k + \alpha_k \tilde{d}^k)\|^2 = 0. \tag{40}$$

Taking a subsequence, if necessary, assume that

$$\lim_{k\in\mathcal{K}} \alpha_k = \alpha^*$$

for some $\alpha^* \in [0,1]$.

Two cases arise. First assume that $\alpha^* > 0$. Then from (39), $d^k \overset{k\in\mathcal{K}}{\to} 0$, so $\tilde{d}^k \overset{k\in\mathcal{K}}{\to} 0$ and $d^{k+} \overset{k\in\mathcal{K}}{\to} -\nabla f(x^k)$. Also from (40),

$$\lim_{k\in\mathcal{K}} \|x^k(\alpha_k;\delta_k(\alpha_k)) - (x^k + \alpha_k \tilde{d}^k)\| = 0, \tag{41}$$

and so from (9),

$$\lim_{k\in\mathcal{K}} \delta_k(\alpha_k) = 0.$$

Using this limit together with (41), we get

$$x^*(\alpha^*, 0) = P(x^* - \alpha^* \nabla f(x^*)) = x^*,$$

which implies that x^* is critical.

For the second case, take $\alpha^* = 0$. Then for $k \in \mathcal{K}$ sufficiently large, the test (11) will fail at least once; thus, using the notation

$$\alpha_k^- = \frac{\alpha_k}{\beta},$$

we have that

$$\begin{aligned} &f(x^k) - f(x^k(\alpha_k^-;\delta_k(\alpha_k^-))) \\ &\quad < \sigma\left\{\alpha_k^- d^{kT} D^k d^k + \frac{1}{\alpha_k^-}\|x^k(\alpha_k^-;\delta_k(\alpha_k^-)) - (x^k + \alpha_k^- \tilde{d}^k)\|^2\right\}. \end{aligned} \tag{42}$$

Since, by (38), ϵ_k is bounded away from zero and since it follows from (36) that $\|g^k\|$ is bounded above, we have

$$\liminf_{k\in\mathcal{K}} \epsilon_k/\|g^k\| > 0. \tag{43}$$

Hence, setting $\bar{\sigma} = (\sigma+1)/2$, Lemma 4.4 can be applied to find an $\bar{\alpha} > 0$ such that (23) holds for $\alpha \in (0, \bar{\alpha}]$. Moreover, closer examination of the proof of Lemma 4.4 shows that, because of (43), the value of $\bar{\alpha}$ can be chosen independently of x^k for k sufficiently large. Now, since $\lim_{k\in\mathcal{K}} \alpha_k^- = 0$, we have for k sufficiently large that

$$\begin{aligned} &\nabla f(x^k)^T[x^k - x^k(\alpha_k^-;\delta_k(\alpha_k^-))] \\ &\quad \geq \frac{\sigma+1}{2}\left\{\alpha_k^- d^{kT} D^k d^k + \frac{1}{\alpha_k^-}\|x^k(\alpha_k^-;\delta_k(\alpha_k^-)) - (x^k + \alpha_k^- \tilde{d}^k)\|^2\right\}. \end{aligned} \tag{44}$$

Using the mean value theorem and combining (42) and (44), we have

$$\begin{aligned} &\frac{1-\sigma}{2}\left\{\alpha_k^- d^{kT} D^k d^k + \frac{1}{\alpha_k^-}\|x^k(\alpha_k^-;\delta_k(\alpha_k^-)) - (x^k + \alpha_k^- \tilde{d}^k)\|^2\right\} \\ &\quad \leq \nabla f(x^k)^T[x^k - x^k(\alpha_k^-;\delta_k(\alpha_k^-))] - f(x^k) + f(x^k(\alpha_k^-;\delta_k(\alpha_k^-))) \\ &\quad = [\nabla f(x^k) - \nabla f(\zeta^k)]^T[x^k - x^k(\alpha_k^-;\delta_k(\alpha_k^-))] \end{aligned} \tag{45}$$

for some ζ^k on the line joining x^k to $x^k(\alpha_k^-, \delta_k(\alpha_k^-))$. Note that

$$\frac{1}{\alpha_k^-}\|x^k - x^k(\alpha_k^-;\delta_k(\alpha_k^-))\| \le \frac{1}{\alpha_k^-}\|x^k - x^k(\alpha_k^-;0)\| + \frac{\delta(\alpha_k^-)}{\alpha_k^-} \le \|g^k\| + C_1(\alpha_k^-)^{\tau/2-1},$$

which is bounded because of (36). Hence the right-hand side of (45) is $o(\alpha_k^-)$, and dividing both sides of (45) by α_k^-, we have that

$$\lim_{k\in\mathcal{K}} d^{kT}D^k d^k = 0, \tag{46}$$

$$\lim_{k\in\mathcal{K}} \frac{1}{(\alpha_k^-)^2}\|x^k(\alpha_k^-;\delta_k(\alpha_k^-)) - (x^k + \alpha_k^-\tilde{d}^k)\|^2 = 0. \tag{47}$$

From (46) $\lim_{k\in\mathcal{K}} d^k = 0$, so $\lim_{k\in\mathcal{K}} \tilde{d}^k = 0$. Since, in addition, $\tilde{d}^k \in T^k$, we have that $x^k + \alpha_k^-\tilde{d}^k \in X$ for k sufficiently large. Lemma 4.2(a) can be applied to show that

$$\begin{aligned}
&\frac{1}{\alpha_k^-}\|x^k(\alpha_k^-;0) - (x^k + \alpha_k^-\tilde{d}^k)\| \\
&= \frac{1}{\alpha_k^-}\|P((x^k + \alpha_k^-\tilde{d}^k) + \alpha_k^- d^{k+}) - (x^k + \alpha_k^-\tilde{d}^k)\| \\
&\ge \|P((x^k + \alpha_k^-\tilde{d}^k) + d^{k+}) - (x^k + \alpha_k^-\tilde{d}^k)\|.
\end{aligned} \tag{48}$$

Meanwhile, Lemma 4.2(b) implies that

$$\frac{1}{\alpha_k^-}\|x^k(\alpha_k^-;0) - (x^k + \alpha_k^-\tilde{d}^k)\| \le \|d^{k+}\| \le B. \tag{49}$$

Taking the sequence in (47) and using Lemma 4.1, (10), (48), and (49), we have

$$\begin{aligned}
&\frac{1}{(\alpha_k^-)^2}\|x^k(\alpha_k^-;\delta_k(\alpha_k^-)) - (x^k + \alpha_k^-\tilde{d}^k)\|^2 \\
&\ge \frac{1}{(\alpha_k^-)^2}\|x^k(\alpha_k^-;0) - (x^k + \alpha_k^-\tilde{d}^k)\|^2 \\
&\quad - \frac{2\delta_k(\alpha_k^-)}{\alpha_k^-}\left(\frac{1}{\alpha_k^-}\right)\|x^k(\alpha_k^-;0) - (x^k + \alpha_k^-\tilde{d}^k)\| \\
&\ge \|P((x^k + \alpha_k^-\tilde{d}^k) + d^{k+}) - (x^k + \alpha_k^-\tilde{d}^k)\|^2 - 2BC_1(\alpha_k^-)^{\tau/2-1}.
\end{aligned} \tag{50}$$

Since the second term in this expression approaches zero, it follows from (50) that in the limit,

$$P(x^* - \nabla f(x^*)) = x^*,$$

and so x^* is critical, again giving a contradiction. □

5. Local convergence. For the exact algorithm, the local convergence analysis is quite simple because when convergence occurs to a local minimum that satisfies the "standard" assumptions, the iterates eventually all lie on the manifold defined by the constraints that are active at the solution. This does not occur in our case, in which the iterates remain in the interior of X. We thus need to ensure that the distance of the iterates to the active manifold is decreasing sufficiently quickly so as not to interfere with the (rapid) convergence in the tangent direction. Fortunately, some inherent properties of the path following projection algorithm prove to be useful here.

In this section we prove R-quadratic convergence of an algorithm in which D^k is a reduced Hessian. Much of the analysis is devoted to showing that step lengths of $\alpha_k = 1$ are used for all sufficiently large k. We start by defining a scheme for choosing ϵ_k, and we then state an active set identification result. Eventual unit step length is established in a sequence of lemmas and in Theorem 5.6. We conclude with the main rate-of-convergence result in Theorem 5.7.

In addition to the assumptions made in the preceding sections, we use the following:

(D) x^* is a strict local minimum that is nondegenerate; that is,

$$-\nabla f(x^*) \in ri\, N(x^*; X),$$

where $ri\, N(x^*; X)$ is the interior of $N(x^*; X)$ relative to the affine hull of $N(x^*; X)$.

(E) ϵ_k is defined as

$$\epsilon_k = \min(\epsilon, \hat{c}_k \epsilon(x_k)),$$

where $\epsilon > 0$ is a positive constant,

$$\epsilon(x) = \|x - P(x - \nabla f(x))\|,$$

and $\hat{c}_k \in [1, \hat{B}]$ for some $\hat{B} < \infty$. ($\hat{c}_k$ is a "random" quantity and need not be a function of x^k.)

If assumption (B) also holds at x^*, then assumption (D) implies that there are *unique* scalars $y_i^* > 0$ such that

$$-\nabla f(x^*) = \sum_{i \in \mathcal{A}} y_i^* a_i, \tag{51}$$

where $\mathcal{A}$ is as defined in §1. For later reference we introduce the notation

$$\bar{A} = [a_i]_{i \in \mathcal{A}}, \quad \bar{A} \in R^{n \times r}, \quad r \leq n.$$

Orthonormal matrices $Z \in R^{n \times (n-r)}$ and $Y \in R^{n \times r}$ can be defined such that $Z^T \bar{A} = 0$ and $Z^T Y = 0$.

Our relaxed definition of ϵ_k is motivated by the fact that calculation of $x - P(x - \nabla f(x))$ involves a projection onto X and hence will be carried out inexactly by the algorithm of §3. The following scheme can be used.

Algorithm to calculate ϵ_k.

Step 1: Given some constant $\hat{C} \in (0, 1)$, apply the algorithm of §2 to find $P(x^k - \nabla f(x^k))$, terminating when the duality gap $\epsilon_P^2/2$ satisfies the inequality

$$\epsilon_P \leq (1 - \hat{C})\|\hat{x}^k - x^k\|,$$

where $\hat{x}^k$ is the latest estimate of the solution.

Step 2: Set $\epsilon_k = \min(\epsilon, 2\|\hat{x}^k - x^k\|)$.

With the notation $\hat{x}^{k*} = P(x^k - \nabla f(x^k))$, Lemma 4.1 and the conditions on ϵ_P can be used to show that

$$\hat{C} \leq 1 - \frac{\epsilon_P}{\|\hat{x}^k - x^k\|} \leq \frac{\|\hat{x}^{k*} - x^k\|}{\|\hat{x}^k - x^k\|} \leq 1 + \frac{\epsilon_P}{\|\hat{x}^k - x^k\|} \leq 2 - \hat{C}.$$

Hence,

$$2\|\hat{x}^k - x^k\| = \hat{c}_k \|\hat{x}^{k*} - x^k\|, \quad \text{where } \hat{c}_k \in \left[\frac{2}{2-\hat{C}}, \frac{2}{\hat{C}}\right],$$

and so the requirements of assumption (D) are satisfied.

From this definition of ϵ_k, the following active set identification result can be proved.

LEMMA 5.1. *Suppose that assumptions* (A), (D), *and* (E) *hold and that* (B) *holds for* $x^k(\alpha, 0)$, *for* $\alpha \in [0,1]$ *and all* k *sufficiently large. Assuming that* x^* *is a limit point of the sequence* $\{x^k\}$, *we have* $\lim_{k\to\infty} x^k = x^*$ *and* $\mathcal{I}^k = \mathcal{A}$ *for all* k *sufficiently large.*

Proof. The result follows from Lemma B.1 of Gafni and Bertsekas [5]; trivial modifications are required because of our relaxed definition of ϵ_k. The assumption (B) in [5] corresponds to our assumption (C) (see [1, Thm. 2.8]). □

We next show that the step lengths do not vanish as $k \to \infty$.

LEMMA 5.2. *Under the assumptions of Lemma* 5.1, *there is* $\hat{\alpha} > 0$ *such that*

$$\alpha_k \geq \hat{\alpha}$$

for all k *sufficiently large.*

Proof. From Lemma 5.1 we have that for k sufficiently large, $\mathcal{I}^k = \mathcal{A}$. Since $d^k = P_{T^k}(-\nabla f(x^k)) \to 0$, it follows that $\|g^k\| \to \|\nabla f(x^*)\| \leq B$. In Lemma 4.4, we are free to set ϵ_k uniformly equal to a constant $\tilde{\epsilon} > 0$, which is chosen so that

$$i \notin \mathcal{I}^k = \mathcal{A} \Rightarrow a_i^T x^k \leq b_i - 2\tilde{\epsilon}\|a_i\|$$

for all sufficiently large k. Hence $\epsilon_k/\|g^k\|$ is bounded away from zero. Now, given any $\bar{\sigma} \in (0,1)$, we can apply Lemma 4.4 to find an $\bar{\alpha}(\bar{\sigma}) > 0$ such that for all $\alpha \in (0, \bar{\alpha}(\bar{\sigma})]$,

$$\nabla f(x^k)^T [x^k - x^k(\alpha; \delta_k(\alpha))] \;\geq\; \bar{\sigma}\left[\alpha d^{kT} D^k d^k + \frac{1}{\alpha}\|x^k(\alpha; \delta_k(\alpha)) - (x^k + \alpha \tilde{d}^k)\|^2\right].$$

If we use L as an upper bound on $\nabla^2 f(x)$ for x in some neighborhood of x^*, it follows exactly as in Gafni and Bertsekas [5] that

$$\begin{aligned} &f(x^k) - f(x^k(\alpha; \delta_k(\alpha))) \\ &\quad \geq \alpha(\bar{\sigma} - L\lambda_2\alpha) d^{kT} D^k d^k + \left(\frac{\bar{\sigma}}{\alpha} - L\right) \|x^k(\alpha; \delta_k(\alpha)) - (x^k + \alpha \tilde{d}^k)\|^2. \end{aligned}$$

If we choose

$$\tilde{\alpha} = \sup_{\bar{\sigma} \in [\sigma, 1)} \min\left(\bar{\alpha}(\bar{\sigma}), \frac{\bar{\sigma} - \sigma}{L\lambda_2}, \frac{\bar{\sigma} - \sigma}{L}\right),$$

it follows from the line-search mechanism (11) that

$$\alpha_k \geq \hat{\alpha} \stackrel{\text{def}}{=} \frac{\tilde{\alpha}}{\beta},$$

and the result follows, since clearly $\hat{\alpha} > 0$. □

The next result follows easily from Lemma 5.2, [5, Lemma B.2], and the analysis of Dunn [2].

LEMMA 5.3. *Under the assumptions of Lemma* 5.1, *we have for all k sufficiently large that*

$$x^k(\alpha; 0) = x^* + Zv^k(\alpha) \in x^* + T^k$$

for all $\alpha \in [\alpha_k, 1]$, *where* $v^k(\alpha) \in R^{n-r}$. *Also,*

$$(x^k + \alpha(\tilde{d}^k + d^{k+})) - x^k(\alpha; 0) = \bar{A}\bar{y}^k(\alpha) \in N(x^*; X) \tag{52}$$

for all $\alpha \in [\alpha_k, 1]$, *where* $\bar{y}^k(\alpha) \in R^r$ *has* $\bar{y}^k_i(\alpha) > C_2\alpha$ *for* $i = 1, \cdots, r$ *and some constant* $C_2 > 0$.

Proof. We prove only the last statement concerning the lower bound on $\bar{y}^k(\alpha)$. Since $d^k \to 0$ and $d^{k+} \to -\nabla f(x^k)$, we can combine (51) and (52) to obtain

$$x^k(\alpha; 0) - (x^k + \alpha\bar{A}y^*) + \bar{A}\bar{y}^k(\alpha) \to 0, \tag{53}$$

where $y^* = \{y^*_i\}_{i\in\mathcal{A}}$. Since $x^{k+1} - x^k \to 0$ we have from (9) that $\delta_k(\alpha_k) \to 0$. Hence,

$$\begin{aligned} 0 \le \|x^k(\alpha_k; 0) - x^k\| &\le \|x^k(\alpha_k; \delta_k(\alpha_k)) - x^k(\alpha_k; 0)\| + \|x^{k+1} - x^k\| \\ &\le \delta_k(\alpha_k) + \|x^{k+1} - x^k\| \to 0. \end{aligned}$$

Now by Lemma 4.2, and since $\alpha \in [\alpha_k, 1]$,

$$\frac{1}{\alpha}\|x^k(\alpha; 0) - x^k\| \le \frac{1}{\alpha_k}\|x^k(\alpha_k; 0) - x^k\|.$$

Since $\alpha_k \ge \hat{\alpha}$, it follows from this inequality that $x^k(\alpha; 0) - x^k \to 0$. Hence from (53), using the full rank of $\bar{A}$, we have that

$$\bar{y}^k(\alpha) \to \alpha y^*.$$

Since $y^* > 0$, the result follows. □

COROLLARY 5.4. *Under the assumptions of Lemma* 5.1, *we have for all k sufficiently large that*

$$Z^Ts^{k+} = 0 \quad \textit{where } s^{k+} = x^k(1; 0) - (x^k + \tilde{d}^k).$$

In addition, $a_i^T(x^k + s^{k+}) = b_i$ *for* $i \in \mathcal{A}$.

Proof. The statement $Z^Ts^{k+} = 0$ follows from the second expression in Lemma 5.3 by setting $\alpha = 1$ and noting that $Z^Td^{k+} = 0$. For the second part, note that

$$x^k + s^{k+} = x^k(1; 0) - \tilde{d}^k \in x^* + T^k$$

from the first expression in Lemma 5.3 and the fact that $\tilde{d}^k \in T^k$. Hence $a_i^T(x^k + s^{k+}) = a_i^Tx^* = b_i$, as required. □

For the remainder of this section we use the following notational conventions:

- $\gamma_k = \delta_k(\alpha_k)^2/2$ is the final duality gap for the step from x^k to x^{k+1}.
- The error in the approximate unit step is separated into two components:

$$x^k(1; \delta_k(1)) - x^k(1; 0) = e^k = \tilde{e}^k + e^{k+},$$

where $\tilde{e}^k = P_{T^k}(e^k) = ZZ^Te^k$ and $e^{k+} = YY^Te^k$.

- δ_k denotes $\delta_k(1)$.

A technical result is needed before establishing eventual unit steps.

LEMMA 5.5. *Suppose that assumption* (C) *and the assumptions of Lemma* 5.1 *hold and that the "special case" in the projection algorithm (i.e., $x^k - \alpha_k g^k \in X$) occurs only finitely often. Then, for k sufficiently large, there are positive constants C_3, C_4, and C_5 such that*

$$(d^{k+})^T s^{k+} \geq C_3 \gamma_{k-1} \|d^{k+}\|,$$
$$\|s^{k+}\| \leq C_4 \gamma_{k-1}.$$

Further, if $\alpha_k = 1$, we have that

$$\|e^{k+}\| \leq C_5 \gamma_k,$$
$$|d^{kT} \tilde{e}^k| \leq \delta_k \|d^k\|.$$

Proof. Assume k is large enough that the "special case" never occurs after iteration $k-1$. Assume further that α_{k-1} and all subsequent step lengths are bounded below by $\hat{\alpha}$, as in Lemma 5.2. Recall that $x^k = x^{k-1}(\alpha_{k-1}; \delta_{k-1}(\alpha_{k-1}))$. Let ν^{k-1} and y^{k-1} be the final values of the ν and y variables in the projection algorithm of §3, which was used to compute x^k. We start by finding bounds on elements of ν^{k-1} in terms of γ_{k-1}; these are needed for the first three inequalities.

As discussed in the proof of Lemma 5.3, $\delta_k(\alpha_k) \to 0$; that is, the projection subproblem is solved more and more accurately. Recall that the matrix equation in (20) holds at every iteration of the projection algorithm. The first part of this equation yields

$$x^k - (x^{k-1} + \alpha_{k-1}(\tilde{d}^{k-1} + d^{(k-1)+})) + Ay^{k-1} + qy^{k-1}_{m+1} = 0. \tag{54}$$

From the second part of the equation and the choice of $\tilde{\epsilon}$ in the proof of Lemma 5.2,

$$\nu_i^{k-1} = b_i - a_i^T x^k \geq 2\tilde{\epsilon}\|a_i\| > 0 \quad \text{for } i \notin \mathcal{A}.$$

Since $\gamma_{k-1} = \delta_{k-1}(\alpha_{k-1})^2/2 = \sum_{i=1}^{m+1} \nu_i^{k-1} y_i^{k-1}$, we have for $i \notin \mathcal{A}$ that

$$0 < y_i^{k-1} \leq \frac{\gamma_{k-1}}{\nu_i^{k-1}} \leq \frac{\gamma_{k-1}}{2\tilde{\epsilon}\|a_i\|} \to 0. \tag{55}$$

Since we have assumed that ν^*_{m+1} is bounded away from zero for all projection subproblems,

$$y^{k-1}_{m+1} \to 0. \tag{56}$$

Now, using $k-1$ instead of k in (52) and setting $\alpha = \alpha_{k-1}$,

$$x^{k-1}(\alpha_{k-1}; 0) - (x^{k-1} + \alpha_{k-1}(\tilde{d}^{k-1} + d^{(k-1)+})) + \bar{A}\bar{y}^{k-1}(\alpha_{k-1}) = 0. \tag{57}$$

Comparing (54) with (57), we have

$$x^{k-1}(\alpha_{k-1}; 0) - x^k = Ay^{k-1} + qy^{k-1}_{m+1} - \bar{A}\bar{y}^{k-1}(\alpha_{k-1}).$$

Using (55), (56), and the full rank of $\bar{A}$, and noting that $\|x^k - x^{k-1}(\alpha_{k-1}; 0)\| \leq \delta_{k-1}(\alpha_{k-1}) \to 0$, we have that for some constant $C_2 > 0$,

$$y_i^{k-1} \to \bar{y}_i^{k-1}(\alpha_{k-1}) \geq C_2 \alpha_{k-1} \quad \text{for } i \in \mathcal{A}.$$

Hence, for k sufficiently large, with $\alpha_{k-1} \geq \hat{\alpha}$, there is a constant $C_{2L} > 0$ such that

$$y_i^{k-1} \geq C_{2L} \quad \text{for } i \in \mathcal{A}.$$

Also, by full rank of $\bar{A}$ and boundedness of ∇f, there is a $C_{2U} > 0$ such that

$$y_i^{k-1} \leq C_{2U}.$$

By assumption (C), we have for $i \in \mathcal{A}$ that

$$\nu_i^{k-1} \geq \frac{\gamma_{k-1}}{\mu y_i^{k-1}} \geq \frac{\gamma_{k-1}}{C_{2U}\mu}.$$

Also,

$$\nu_i^{k-1} \leq \frac{\gamma_{k-1}}{y_i^{k-1}} \leq \frac{\gamma_{k-1}}{C_{2L}}.$$

From these last two expressions, we can define positive constants C_{7L} and C_{7U} such that

$$C_{7L}\gamma_{k-1} \leq \nu_i^{k-1} \leq C_{7U}\gamma_{k-1} \quad \text{for } i \in \mathcal{A}. \tag{58}$$

For the first result, note from $d^{k+} \rightarrow -\nabla f(x^*)$ and $Z^T d^{k+} = 0$ that $d^{k+} = \bar{A}t^k$, where $t^k \rightarrow y^*$. Hence, $t^k > 0$ for k sufficiently large. From Corollary 5.4, (20), and (58), we have for $i \in \mathcal{A}$ that

$$a_i^T s_i^{k+} = b_i - a_i^T x^k = \nu_i^{k-1} \geq C_{7L}\gamma_{k-1}.$$

Hence, noting that $\|t^k\| \geq \|d^{k+}\|/\|\bar{A}\|$, we have

$$(d^{k+})^T s^{k+} = (\bar{A}t^k)^T s^{k+} = t^{kT}(\bar{A}^T s^{k+}) \geq C_{7L}\gamma_{k-1}\|t^k\| \geq C_3\gamma_{k-1}\|d^{k+}\|$$

for $C_3 = C_{7L}/\|\bar{A}\|$, giving the first result.

For the second result, we have from Lemma 5.3, Corollary 5.4, and (20) that for some $u^k \in \mathbf{R}^r$

$$s^{k+} = \bar{A}u^k \quad \text{and} \quad \bar{A}^T s^{k+} = [\nu_i^{k-1}]_{i \in \mathcal{A}}.$$

Hence,

$$\bar{A}^T \bar{A} u^k = [\nu_i^{k-1}]_{i \in \mathcal{A}},$$

which, by full rank of $\bar{A}$, boundedness of ρ_i, and (58), gives

$$\|u^k\| \leq C_8\gamma_{k-1}$$

for some constant $C_8 > 0$. Since

$$\|s^{k+}\| \leq \|\bar{A}\|\|u^k\|,$$

the result follows by setting $C_4 = C_8\|\bar{A}\|$.

For the third inequality, we again use (20) and Lemma 5.3 to deduce that for $i \in \mathcal{A}$,

$$\nu_i^k = b_i - a_i^T x^k(1;\delta_k) = a_i^T[x^k(1;0) - x^k(1;\delta_k)] = -a_i^T e^{k+}.$$

Now $e^{k+} = \bar{A}v^k$ for some v^k, so an argument identical to that of the preceding paragraph can be used to give the result.

The fourth inequality follows simply from

$$|d^{kT}\tilde{e}^k| \le \|d^k\|\|\tilde{e}^k\| \le \|d^k\|\|e^k\| \le \delta_k\|d^k\|. \qquad \square$$

THEOREM 5.6. *Suppose that assumptions* (A), (C), (D), *and* (E) *hold and that assumption* (B) *holds in a neighborhood of* x^*. *Suppose that* $Z^T\nabla^2 f(x^*)Z$ *is positive definite and that for k sufficiently large, the tangent component of the step is given by*

$$\tilde{d}^k = Z(Z^T\nabla^2 f(x^k)Z)^{-1}Z^T d^k.$$

Suppose there is a nonnegative sequence $\{\xi_k\}$ *such that* $\lim_{k\to\infty}\xi_k = 0$ *and that, in addition to* (9) *and* (10), *the sequence* $\{\delta_k\}$ *satisfies*

$$(59) \qquad \|d^k\|\delta_k \le \xi_k\gamma_{k-1}, \qquad \delta_k^2 \le \xi_k\gamma_{k-1}.$$

Assume that $\sigma < 0.5$ *in* (11). *Then* $\alpha_k = 1$ *for all sufficiently large k.*

Proof. First, we consider the special case of $x^* \in \operatorname{int} X$, for which we have $\nabla f(x^*) = 0$. By Lemma 5.1, the two-metric gradient projection method reduces to Newton's method when k is sufficiently large. Consequently, $d^k = -\nabla f(x^k)$ and $\tilde{d}^k = -(\nabla^2 f(x^k))^{-1}\nabla f(x^k)$. By Lemma 5.3, $x^k(1;0) = x^k + \tilde{d}^k \in \operatorname{int} X$ for k sufficiently large. Correspondingly, exactness of the projection yields $\gamma_k = \delta_k \equiv 0$ for all such k. Now

$$\begin{aligned} f(x^k) - f(x^k(1;\delta_k)) &= f(x^k) - f(x^k + \tilde{d}^k) \\ &= -\nabla f(x^k)^T\tilde{d}^k - \tfrac{1}{2}(\tilde{d}^k)^T\nabla^2 f(x^k)\tilde{d}^k + o(\|\tilde{d}^k\|^2) \\ &= \tfrac{1}{2}d^{kT}(\nabla^2 f(x^k))^{-1}d^k + o(\|d^k\|^2). \end{aligned}$$

The second term on the right-hand side of (11) is zero, so for k sufficiently large, (11) is satisfied for $\alpha_k = 1$.

In the remaining case, $x^* \in \partial X$; thus, by assumption (C), $\nabla f(x^*) \ne 0$. Moreover, the "special case" does not occur in the projection algorithm for sufficiently large k. (This follows directly from (52), which states in particular that $(x^k - \alpha_k g^k) - P(x^k - \alpha_k g^k) \ne 0$.) Since

$$\nabla f(x^k) = -d^k - d^{k+}$$

and

$$x^k(1;\delta_k) = x^k + \tilde{d}^k + s^{k+} + e^k,$$

we have that

$$\begin{aligned} f(x^k) - f(x^k(1;\delta_k)) &= \nabla f(x^k)^T[x^k - x^k(1;\delta_k)] \\ &\quad -(\tfrac{1}{2})[x^k - x^k(1;\delta_k)]^T\nabla^2 f(x^k)[x^k - x^k(1;\delta_k)] \\ &\quad +o(\|x^k - x^k(1;\delta_k)\|^2) \\ &= [-d^k - d^{k+}]^T[-\tilde{d}^k - s^{k+} - e^k] \\ &\quad -(\tfrac{1}{2})[-\tilde{d}^k - (s^{k+} + e^k)]^T\nabla^2 f(x^k)[-\tilde{d}^k - (s^{k+} + e^k)] \\ &\quad +o(\|x^k - x^k(1;\delta_k)\|^2). \end{aligned}$$

It can be easily shown that $\tilde{d}^{kT}\nabla^2 f(x^k)\tilde{d}^k = d^{kT}\tilde{d}^k$, and so, after some rearrangement,

$$
\begin{aligned}
f(x^k) - f(x^k(1;\delta_k)) = {} & \left\{\tfrac{1}{2}d^{kT}\tilde{d}^k + (s^{k+} + e^k)^T(s^{k+} + e^k)\right\} \\
& +(d^{k+})^T s^{k+} - \nabla f(x^k)^T e^k \\
& -\tfrac{1}{2}[s^{k+} + e^k]^T[\nabla^2 f(x^k) + 2I][s^{k+} + e^k] \\
& -[s^{k+} + e^k]^T\nabla^2 f(x^k)\tilde{d}^k + o(\|x^k - x^k(1;\delta_k)\|^2).
\end{aligned}
\tag{60}
$$

Now Lemma 5.5 can be used to deduce the following inequalities:

$$(d^{k+})^T s^{k+} \geq C_3\gamma_{k-1}\|d^{k+}\| \geq \bar{C}_3\gamma_{k-1},$$

for some $\bar{C}_3 > 0$, since $\|d^{k+}\| \to \|\nabla f(x^*)\| \neq 0$;

$$\left|\nabla f(x^k)^T e^k\right| \leq |d^{kT}\tilde{e}^k| + |(d^{k+})^T e^{k+}| \leq \delta_k\|d^k\| + C_5 B\gamma_k;$$

$$
\begin{aligned}
\tfrac{1}{2}\left|[s^{k+} + e^k]^T[\nabla^2 f(x^k) + 2I][s^{k+} + e^k]\right| & \leq C_{11}\|s^{k+} + e^k\|^2 \\
& \leq C_{12}\gamma_{k-1}^2 + C_{13}\gamma_{k-1}\delta_k + C_{14}\delta_k^2;
\end{aligned}
$$

$$\left|[s^{k+} + e^k]^T\nabla^2 f(x^k)\tilde{d}^k\right| \leq C_{15}\|d^k\|(\gamma_{k-1} + \delta_k).$$

By substituting in (60), we obtain

$$
\begin{aligned}
f(x^k) - f(x^k(1;\delta_k)) \geq {} & \left\{\tfrac{1}{2}d^{kT}\tilde{d}^k + (s^{k+} + e^k)^T(s^{k+} + e^k)\right\} \\
& +\bar{C}_3\gamma_{k-1} - \delta_k\|d^k\| - C_5 B\gamma_k - C_{12}\gamma_{k-1}^2 - C_{13}\gamma_{k-1}\delta_k \\
& -C_{14}\delta_k^2 - C_{15}\|d^k\|\gamma_{k-1} - C_{15}\|d^k\|\delta_k + o(\|\tilde{d}^k + s^{k+} + e^k\|^2) \\
= {} & \left\{\tfrac{1}{2}d^{kT}\tilde{d}^k + (s^{k+} + e^k)^T(s^{k+} + e^k)\right\} \\
& +\gamma_{k-1}\left[\bar{C}_3 - \xi_k - (C_5 B\xi_k/2) - C_{12}\gamma_{k-1}\right. \\
& \qquad \left. -C_{13}\delta_k - C_{14}\xi_k - C_{15}\|d^k\| - C_{15}\xi_k\right] \\
& +o(\|\tilde{d}^k + s^{k+} + e^k\|^2).
\end{aligned}
$$

As $k \to \infty$, the term in square brackets approaches $\bar{C}_3 > 0$; that is, it is positive for sufficiently large k. It is easy to see that the final $o(\|\tilde{d}^k + s^{k+} + e^k\|^2)$ term is eventually dominated by the term in curly brackets. Hence, since $\sigma \in (0, \frac{1}{2})$, we have for k sufficiently large that

$$f(x^k) - f(x^k(1;\delta_k)) \geq \sigma\left\{d^{kT}\tilde{d}^k + \|x^k(1,\delta_k) - (x^k + \tilde{d}^k)\|^2\right\},$$

and so $\alpha_k = 1$ passes the acceptance test (11) and $x^k(1;\delta_k)$ will be accepted as the new iterate. □

The conditions (59) should be imposed only in the final stages of the algorithm, when there is a suspicion that the active manifold has been identified. Otherwise, it could happen that at some early iterate, $x^k - g^k \in \mathrm{int}X$, in which case the projection is performed exactly ($\gamma_k = \delta_k = 0$) and, because of (59), exact projections would be demanded at all subsequent iterations.

A result similar to Theorem 5.6 can be stated for the alternative acceptance test (12), and it can be proved in an almost identical fashion.

We can now prove the final result.

THEOREM 5.7. *Suppose that the assumptions of Theorem* 5.6 *hold and that the sequence* $\{\gamma_k\}$ *converges Q-quadratically to zero; that is, there is a constant* C_{10} *such that*

$$\delta_k \le C_{10}\gamma_{k-1}. \tag{61}$$

Then the rate of local convergence of the algorithm is R-quadratic.

Proof. In the case $x^* \in \operatorname{int} X$, we actually obtain Q-quadratic convergence, since the algorithm eventually reduces to Newton's method. We therefore focus on the case of $x^* \in \partial X$.

By setting $\xi_k = \max(\|d^k\|, \delta_k)$, it is easy to see that (61) implies (59), and so Theorem 5.6 applies. By the definition of s^{k+},

$$x^k + \tilde{d}^k - x^k(1;0) = -s^{k+}.$$

Multiplying through by Z^T and using the definition of $\tilde{d}^k$, we obtain

$$Z^T(x^k - x^k(1;0)) - (Z^T\nabla^2 f(x^k)Z)^{-1}Z^T\nabla f(x^k) = 0. \tag{62}$$

By optimality of x^*, $Z^T\nabla f(x^*) = 0$, so by Taylor series expansion, and since $ZZ^T + YY^T = I$,

$$\begin{aligned} &Z^T(\nabla f(x) + \nabla^2 f(x)(x^* - x)) = O(\|x - x^*\|^2)\\ &\Rightarrow Z^T\nabla f(x^k) - Z^T\nabla^2 f(x^k)ZZ^T(x^k - x^*)\\ &\quad = Z^T\nabla^2 f(x^k)YY^T(x^k - x^*) + O(\|x^k - x^*\|^2). \end{aligned} \tag{63}$$

Multiplying (63) by $(Z^T\nabla f(x^k)Z)^{-1}$ and adding to (62), we have

$$\|Z^T(x^* - x^k(1;0))\| = O(\|Y^T(x^k - x^*)\|) + O(\|x^k - x^*\|^2). \tag{64}$$

Recall that $x^k = x^{k-1}(1;\delta_{k-1})$ and that by Lemma 5.3

$$Y^T(x^{k-1}(1;0) - x^*) = 0 \tag{65}$$

for all sufficiently large k. Hence, using the third inequality in Lemma 5.5, we have

$$\begin{aligned} \|Y^T(x^k - x^*)\| &= \|Y^T(x^{k-1}(1;\delta_{k-1}) - x^{k-1}(1,0))\|\\ &= \|e^{(k-1)+}\|\\ &\le C_{20}\gamma_{k-1}. \end{aligned} \tag{66}$$

From (64)–(66),

$$\begin{aligned} \|x^* - x^{k+1}\| &\le \|x^* - x^k(1;0)\| + \|e^k\|\\ &\le C_{20}\gamma_{k-1} + \delta_k + O(\|x^k - x^*\|^2). \end{aligned} \tag{67}$$

Now $\delta_k \le C_{10}\gamma_{k-1}$, and so we can choose a constant $C_{21} \ge \max(1, C_{20} + C_{10})$ such that

$$\|x^* - x^{k+1}\| \le C_{21}\max(\gamma_{k-1}, \|x^* - x^k\|^2). \tag{68}$$

Given any $\tau < C_{21}^{-1}$, we can choose an integer $\bar{k}$ sufficiently large that

$$\gamma_{k-1} \le \tau^2 \quad \text{and} \quad \|x^k - x^*\| \le \tau \quad \text{for all } k \ge \bar{k}.$$

An inductive argument based on (68) then shows that

$$\|x^{\bar{k}+j} - x^*\| \le \bar{\tau}_j,$$

where

$$\bar{\tau}_0 = \tau, \qquad \bar{\tau}_{j+1} = C_{21}\bar{\tau}_j^2.$$

Clearly the sequence $\{\bar{\tau}_j\}$ is Q-quadratically convergent, so the result follows. □

Results similar to Theorems 5.6 and 5.7 could be proved for other choices of $\tilde{d}^k$, for example, where $\tilde{d}^k$ is a quasi-Newton or inexact Newton method step rather than the reduced Newton step. These would be of practical importance in applications in which it is difficult to compute or factor the reduced Hessian.

Finally, we note that it may be efficient to include a second "local" phase in the basic algorithm of §2. When it appears that the active constraint set has been identified, the current iterate could be projected onto the appropriate manifold (placing it on ∂X). Standard methods for equality-constrained nonlinear programming could then be applied to identify the minimum on this manifold. However, it is likely that the basic algorithm would also be quite efficient in this situation because, as the final few iterates are close together, a good starting point for the projection would be readily available.

Acknowledgments. I thank David Stewart for helpful discussions during the preparation of this paper and the referees and editor for their comments on the first draft.

REFERENCES

[1] J. V. Burke and J. J. Moré, *On the identification of active constraints*, SIAM J. Numer. Anal., 25 (1988), pp. 1197–1211.

[2] J. C. Dunn, *On the convergence of projected gradient processes to singular critical points*, J. Optim. Theory Appl., 55 (1987), pp. 203–216.

[3] ———, *Gradient projection methods for systems optimization problems*, Control and Dynamic Systems, 29 (1988), pp. 135–195.

[4] ———, *A projected Newton method for minimization problems with nonlinear inequality constraints*, Numer. Math., 53 (1988), pp. 377–409.

[5] E. M. Gafni and D. P. Bertsekas, *Two-metric projection methods for constrained optimization*, SIAM J. Control Optim., 22 (1984), pp. 936–964.

[6] C.-G. Han, P. Pardalos, and Y. Ye, *Computational aspects of an interior point algorithm for quadratic programming problems with box constraints*, in Large-Scale Numerical Optimization, T. F. Coleman and Y. Li, eds., SIAM Publications, Philadelphia, PA, 1990, pp. 92–112.

[7] M. Kojima, S. Mizuno, and A. Yoshise, *An $o(\sqrt{n}l)$ iteration potential reduction algorithm for linear complementarity problems*, Res. Report B–217, Department of Information Sciences, Tokyo Institute of Technology, Tokyo, Japan, 1988.

[8] O. L. Mangasarian and T.-H. Shiau, *Error bounds for monotone linear complementarity problems*, Math. Programming, 36 (1986), pp. 81–89.

[9] S. J. Wright, *Interior point methods for optimal control of discrete-time systems*, Tech. Report MCS–P229–0491, Mathematics and Computer Science Division, Argonne National Laboratory, Argonne, IL, April 1991.

[10] Y. Ye, K. O. Kortanek, J. A. Kaliski, and S. Huang, *Near-boundary behavior of primal-dual potential reduction algorithms for linear programming*, Working Paper Number 90–9, College of Business Administration, University of Iowa, Iowa City, IA, 1990.

[11] Y. Zhang, J. E. Dennis, and R. A. Tapia, *On the superlinear and quadratic convergence of primal-dual interior point linear programming algorithms*, Tech. Report TR90–6, Department of Mathematical Sciences, Rice University, Houston, TX, 1990.

SIAM J. OPTIMIZATION
Vol. 2, No. 3, pp. 474–504, August 1992

008

FACETS FOR POLYHEDRA ARISING IN THE DESIGN OF COMMUNICATION NETWORKS WITH LOW-CONNECTIVITY CONSTRAINTS*

MARTIN GRÖTSCHEL†, CLYDE L. MONMA‡, AND MECHTHILD STOER†

Abstract. This paper addresses the important practical problem of designing survivable fiber optic communication networks. This problem can be formulated as a minimum-cost network design problem with certain low-connectivity constraints. Previous work presented structural properties of optimal solutions and heuristic methods for obtaining "near-optimal" network designs. Some facet-inducing inequalities for the convex hull of the solutions to this problem are given. A companion paper describes computational results on real-world telephone network design problems with a cutting plane method based on this work. These computational results are summarized in the last section of this paper.

Key words. network design, network survivability, connectivity, polyhedral combinatorics

AMS(MOS) subject classifications. 05C40, 90C27, 90B12

1. Introduction.

A recent trend in communication networks is the emergence of fiber optic technology as one of the major components in the "network of the future." This transmission medium is cost-effective and reliable, and provides very high transmission capacity. This combination promises to usher in new telecommunication services requiring large amounts of bandwidth. At the same time, the unique characteristics of this technology imply the need for new network design approaches. (See [CFLM] for more details.)

Survivability is an important factor in the design of communication networks. Network survivability is used here to mean the ability to restore service in the event of a catastrophic failure of a network component, such as the complete loss of a transmission link, or the failure of a switching node. Service could be restored by routing traffic through other existing network links and nodes, assuming that the design of the network has provided for this additional connectivity. Clearly, a higher level of redundant connectivity results in a greater network survivability and a greater overall network cost. This leads to the problem of designing a minimum-cost network that meets certain required connectivity constraints.

Survivability is a particularly important issue for fiber networks. The high capacity of fiber facilities results in much more sparse network designs with larger amounts of traffic carried by each link than is the case with traditional bandwidth-limited technologies. This increases the potential damage to network services due to link or node failures. It is necessary to trade off the potential for lost revenues and customer goodwill against the extra costs required to increase the network survivability. Recent works on methods for designing survivable fiber communication networks by [CMW] and [MS] conclude that (1) survivability is an important issue for fiber networks, (2) "two-connected" topologies provide a high level of survivability in a cost-effective manner, and (3) good heuristic methods exist for quickly generating "near-optimal"

*Received by the editors July 31, 1990; accepted for publication (in revised form) May 31, 1991.

†Konrad-Zuse-Zentrum für Informationstechnik Berlin, Heilbronner Str. 10, 1000 Berlin 31, Germany.

‡Mathematics, Information Sciences and Operations Research, Bell Communications Research, 445 South Street, Morristown, New Jersey 07960.

networks. In particular, it was determined that a network topology should provide for at least two diverse paths between certain "special" offices, thus providing for protection against any single link or single node failure for traffic between these offices. These special offices represent high revenue–producing offices and other offices that require a higher level of network survivability.

We now formalize the network design problems that are being considered in this paper. A set of **nodes** V is given that represents the locations of the switches (offices) that must be interconnected into a network in order to provide the desired services. A collection E of **edges** is also specified that represents the possible pairs of nodes between which a direct transmission link can be placed. We let $G = (V, E)$ be the (undirected) graph of possible direct link connections. The graph G may have parallel edges but contains no loops. (Thus we assume throughout this paper that all graphs considered are loopless. But they may have parallel edges. Graphs without parallel edges are called **simple.**)

Given a graph $G = (V, E)$ and $W \subseteq V$, the edge set $\delta(W) := \{ij \in E \mid i \in W, j \in V \backslash W\}$ is called the **cut** (induced by W). (We will write $\delta_G(W)$ to make clear—in case of possible ambiguities—with respect to which graph the cut induced by W is considered.)

For $W, W' \in V$ with $W \cap W' = \emptyset$ we define $[W : W'] := \{ij \in E \mid i \in W, j \in W'\}$. So $\delta(W) = [W : V \backslash W]$. For $W \subseteq V$, we denote by $G[W]$ the subgraph of G induced by W and by $E(W)$ its edge set $\{ij \in E \mid i, j \in W\}$. G/W is the graph obtained from G by contracting the nodes in W to a new node w (retaining parallel edges). We call the reverse operation of replacing the shrunk node w by the original node set W the **expansion** of w in G/W to G. We will denote by $G - v$ the graph obtained by removing the vertex v and all incident edges from G, and by $G - F$ the graph obtained by removing the edge set F from G (we write $G - f$ instead of $G - \{f\}$). If $G - v$ has more connected components than G for some node v, we will call v an **articulation node** of G. Similarly, if $G - e$ has more connected components than G, we will call edge e a **bridge** of G.

Each edge $e \in E$ has a **fixed cost** c_e of establishing the direct link connection. The cost of establishing a network $N = (V, F)$ consisting of a subset $F \subseteq E$ of edges is $c(F) := \sum_{e \in F} c_e$, i.e., it is the sum of the costs of the individual links contained in F. The goal is to build a minimum-cost network so that the required survivability conditions, which we describe below, are satisfied. We note that the cost here represents setting up the topology for the communication network and includes placing conduits in which to lay the fiber cables, placing the cables into service, and other related costs. We do not consider costs that depend on how the network is implemented, such as routing, multiplexing, and repeater costs. Although these costs are also important, it is usually the case that a topology is first designed and then these other costs are considered in a second stage of optimization.

For any pair of distinct nodes $s, t \in V$, an $[s, t]$-**path** P is a sequence of nodes and edges $(v_0, e_1, v_1, e_2, \cdots, v_{l-1}, e_l, v_l)$, where each edge e_i is incident with the nodes v_{i-1} and v_i $(i = 1, \cdots, l)$, where $v_0 = s$ and $v_l = t$, and where no node or edge appears more than once in P. A collection $P_1, P_2, \cdots, P_k$ of $[s, t]$-paths is called **edge-disjoint** if no edge appears in more than one path, and is called **node-disjoint** if no node (except for s and t) appears in more than one path. (Remark: In order to be consistent with standard graph theory we do not consider two parallel edges as two node-disjoint paths.)

The **survivability conditions** require that the network satisfy certain edge- and node-connectivity requirements. In particular, each node $s \in V$ has an associated nonnegative integer r_s, which represents its **connectivity requirement**. This means that for each pair of distinct nodes $s, t \in V$, the network $N = (V, F)$ to be designed has to have at least

$$r(s,t) := \min\{r_s, r_t\}$$

edge-disjoint (or node-disjoint) $[s,t]$-paths. These conditions ensure that some communication path between s and t will survive a prespecified level of edge (or node) failures. The levels of survivability specified depend on the relative importance placed on maintaining connectivity between different pairs of offices.

The fiber optic network design problems that arise in practice and that we are addressing in this paper have three types of offices. The so-called "special" offices have connectivity requirement 2 while "ordinary" offices have connectivity requirement 1. An office with connectivity requirement 0 is called "optional" since it need not be part of the network to be designed.

Figure 1.1 shows an example network. Special offices are indicated by squares, ordinary offices by circles. Optional offices do not occur. The lines (thin, bold, and dashed) represent the possible direct links from which the minimum-cost survivable network must be designed. The network obtained by removing the dashed lines, i.e., the graph formed by the union of bold and thin lines, represents a feasible network. It consists of a two-connected part (the bold lines) containing all special nodes, in which every pair of nodes is linked by at least two node-disjoint paths and a collection of trees (the thin lines), which link the remaining nodes into the two-connected part.

Thus in the remainder of this paper we consider the case where the connectivity requirements satisfy

$$r_s \in \{0,1,2\} \quad \text{for all } s \in V.$$

Nodes of connectivity requirement 0 (respectively, 1, 2) will also be called nodes of **type 0** (respectively, **type 1, 2**). Let us define the **2ECON** problem (respectively, **2NCON** problem) to be the network design problem where between each pair of distinct nodes s and t at least $\min(r_s, r_t)$ edge-disjoint (respectively, node-disjoint) paths are required.

Given $G = (V, E)$, we extend the connectivity requirement function r to functions operating on sets by setting

$$\begin{aligned} r(W) &:= \max\{r_s \mid s \in W\} \text{ for all } W \subseteq V, \quad \text{and} \\ \operatorname{con}(W) &:= \max\{r(s,t) \mid s \in W, t \in V \backslash W\} \\ &= \min\{r(W), r(V \backslash W)\} \quad \text{for all } W \subseteq V, \quad \emptyset \neq W \neq V. \end{aligned}$$

Let us now introduce, for each edge $e \in E$, a variable x_e and consider the vector space $\mathbb{R}^E$. Every subset $F \subseteq E$ induces an **incidence vector** $\chi^F = (\chi_e^F)_{e \in E} \in \mathbb{R}^E$ by setting $\chi_e^F := 1$ if $e \in F$, and $\chi_e^F := 0$ otherwise. Vice versa, each 0/1-vector $x \in \mathbb{R}^E$ induces a subset $F^x := \{e \in E \mid x_e = 1\}$ of the edge set E of G. For any subset of edges $F \subseteq E$, we define $x(F) := \sum_{e \in F} x_e$. We can now formulate the 2NCON network design problem introduced above as the following integer linear

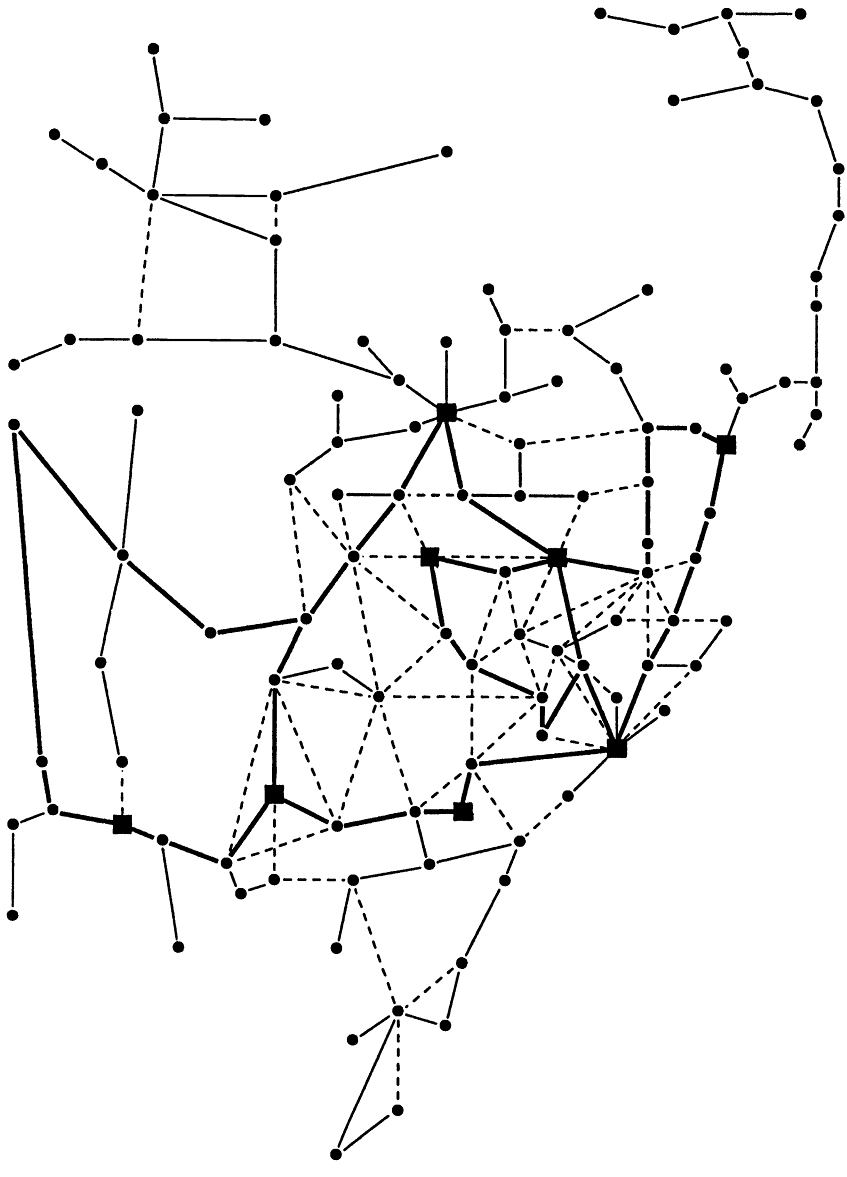


FIG. 1.1

program:

$$\min \sum_{ij \in E} c_{ij} x_{ij}$$

$$(1.1)\quad \begin{array}{lll} & \text{subject to} & \\ & \text{(i) } x(\delta(W)) \geq \mathrm{con}(W) & \text{for all } W \subseteq V, \emptyset \neq W \neq V; \\ & \text{(ii) } x(\delta_{G-z}(W)) \geq 1 & \text{for all } z \in V, \text{ and for all } W \subseteq V \backslash \{z\},\ \emptyset \neq W \neq \\ & & V \backslash \{z\} \text{ with } r(W) = 2 \text{ and } r\big(V \backslash (W \cup \{z\})\big) = 2; \\ & \text{(iii) } 0 \leq x_{ij} \leq 1 & \text{for all } ij \in E; \\ & \text{(iv) } x_{ij} \text{ integral} & \text{for all } ij \in E. \end{array}$$

It follows from Menger's theorem that, for every feasible solution x of (1.1), the subgraph $N = (V, F^x)$ of G defines a network satisfying the two-connected survivability requirements for the 2NCON problem. Removing (ii), we have an integer linear program for the 2ECON network design problem. (Note that in the case $r = \{0,1\}^V$, inequalities (i), (iii), and (iv) of (1.1) characterize the Steiner tree problem.) An inequality of type (i) is called a **cut inequality**, one of type (ii) is called a **node-cut inequality**, and one of type (iii) is called a **trivial inequality**.

The main objective of this paper is to study the 2ECON and 2NCON network design problems from a polyhedral point of view to see which inequalities are suitable choices for a cutting plane approach, i.e., we want to find a tighter LP-relaxation than the one obtained by dropping the integrality constraints (iv) of (1.1) for the 2ECON and 2NCON network design problems. To do this we define the following polytopes. Let $G = (V, E)$ be a graph and let $r \in \{0,1,2\}^V$ be given with $r_v = 2$ for at least two nodes. Then

$$\begin{aligned} \mathrm{2NCON}(G; r) &:= \mathrm{conv}\{x \in \mathbb{R}^E \mid x \text{ satisfies (i), (ii), (iii), (iv) of (1.1)}\}, \\ \mathrm{2ECON}(G; r) &:= \mathrm{conv}\{x \in \mathbb{R}^E \mid x \text{ satisfies (i), (iii), (iv) of (1.1)}\} \end{aligned}$$

are the polytopes associated with the 2NCON and 2ECON network design problems. (Above, "conv" denotes the convex hull operator.) We say that $F \subseteq E$ is **feasible** for one of these polytopes if χ^F is.

Related problems have been investigated previously. A general integer linear programming approach to network design problems with connectivity requirements is presented in [GM] along with a preliminary study of these problems from a polyhedral point of view. We shall make several references to this work in what follows. [CFN] study the dominant of the 2ECON$(G; r)$ polytope in the special case where $r = \{2\}^V$. [MMP] study the 2ECON$(G; r)$ and 2NCON$(G; r)$ polytopes in the special case where $r = \{2\}^V$, and G is a complete graph with the edge weights satisfying the triangle inequality. They show that in this case the optimization problems are the same over both polytopes and then give a certain type of "characterization" of the optimal solutions.

Let us now introduce some connectivity functions and some notation concerning "essential" edges and dimension of polyhedra. Let $G = (V, E)$ and $r \in \{0,1,2\}^V$ be given; we say that $e \in E$ is **essential with respect to** 2ECON$(G; r)$ if 2ECON$(G - e; r) = \emptyset$; similarly we say e is **essential with respect to** 2NCON$(G; r)$ if 2NCON $(G - e; r) = \emptyset$. In other words, e is essential if its deletion results in a graph such that one of the survivability requirements cannot be satisfied. We denote the set of edges of E that are essential with respect to 2ECON$(G; r)$ by **2EES**$(G; r)$, and the set of edges that are essential with respect to 2NCON$(G; r)$ by **2NES**$(G; r)$. Clearly, for all subsets $F \subseteq E \backslash 2\text{EES}(G;r)$, $2\text{EES}(G;r)$ is contained in $2\text{EES}(G-F;r)$ (similarly with $2\text{NES}(G;r)$). Let $\dim(S)$ denote the **dimension** of a set $S \subseteq \mathbb{R}^n$, i.e., the maximum number of affinely independent elements in S minus 1. One of the results proved in [GM] says that the polyhedron $2\text{ECON}(G;r)$ is full-dimensional if and only if $2\text{EES}(G;r)$ is empty, and also that $2\text{NCON}(G;r)$ is full-dimensional if and only if $2\text{NES}(G;r)$ is empty.

Let $G = (V, E)$ be a graph and $W \subseteq V$ with $|W| \geq 2$, and let $G' = (V, E')$ be the simple graph underlying G. We set

$\lambda(G, W) :=$ minimum cardinality of a subset F of E, such that two nodes of W are disconnected in $G - F$;

$\kappa(G, W) :=$ minimum cardinality of a set $S \cup F$, where $S \subseteq V$ and $F \subseteq E'$, such that two nodes of W are disconnected in $G' - (S \cup F)$.

If $|W| < 2$, then $\lambda(G, W)$ and $\kappa(G, W)$ are defined as ∞. If G with node set V_G is a subgraph of some graph H with node set V_H and $W \subseteq V_H$ we will also write $\lambda(G, W)$ instead of $\lambda(G, W \cap V_G)$. We will use these functions frequently in two special situations. To shorten notation in these cases, we introduce the following definitions:

$$\lambda_i(G) := \lambda(G, V_i), \qquad \kappa_i(G) := \kappa(G, V_i),$$

where $V_i := \{v \in V \mid r_v \geq i\}$, $i = 0, 1, 2$. So $\lambda_0(G)$ is nothing but the edge-connectivity of G, and $\kappa_0(G)$ is the node-connectivity of G.

Throughout this paper we make the following assumptions:

(1.2)

Let $G = (V, E)$ and $r \in \mathbb{Z}_+^V$ be given.

(i) $r \in \{0, 1, 2\}^V$ and at least two different nodes s, t satisfy $r_s = r_t = 2$;

(ii) if we consider the 2ECON problem we assume G to be two-node connected and $\lambda_2(G) \geq 3$;

(iii) if we consider the 2NCON problem we assume G to be two-node connected and $\kappa_2(G) \geq 3$.

We will say that (G, r) satisfies (1.2) and mean that the graph $G = (V, E)$ and the vector $r \in \mathbb{Z}_+^V$ of connectivity types satisfy conditions (i), (ii), and (iii). If (i) does not hold, then the 2ECON and the 2NCON problem reduces to the Steiner tree problem for which more specialized investigations can be (and have been) made. We want to exclude this case from the present investigation. It also does not occur in the practical problems we have in mind. One consequence of (ii) and (iii) of (1.2) is that the 2ECON or 2NCON problem contains no essential edges; hence the associated polyhedron is full-dimensional. We further justify assumptions (ii) and (iii) in §2.

In §§2 and 4 we present some decomposition and lifting results that simplify the later discussions. In §3 we investigate which of the basic inequalities given in (1.1) define facets for 2ECON, respectively, 2NCON. In §§5–8 we present several classes of facet-inducing inequalities for 2ECON and 2NCON. These include partition, node-partition, two-cover, and comb inequalities.

We will not discuss the separation problems associated with the classes of inequalities introduced in this paper. Let us just mention here that the cut and node-cut inequalities can be checked in polynomial time, but for all other classes of inequalities to be presented in this paper the separation problem is NP-hard (as is shown in [GMS]). Based on the polyhedral investigations presented in this paper we have designed cutting plane algorithms for the 2ECON and the 2NCON problems. A short summary of our computational results is given in §9. The details can be found in [GMS] and [S].

2. Decomposition. The problem of finding a cost-minimal network for the 2ECON problem can be decomposed into at least two independent problems if the underlying graph G contains an articulation node v disconnecting two nodes of type at least 1. The subproblems are solved on the two-node–connected components of G with the same cost function and the same connectivity types r; only the connectivity type of the articulation node v may have to be adjusted. The 2ECON problem may also be decomposed into independent subproblems if G contains two edges e, f, such that in $G-\{e,f\}$ two nodes of type 2 are disconnected. Another simple decomposition is possible for the 2NCON problem if the graph G contains two nodes u, v so that in $G-\{u,v\}$ two nodes of type 2 are disconnected. These and other more complicated decompositions are described in more detail in [GMS].

Observe that using the above decompositions, any 2ECON or 2NCON problem with essential edges may be decomposed into problems without essential edges. This is the reason why we restrict ourselves to graphs G and connectivity types r for which our general assumptions (ii) and (iii) of (1.2) hold. This implies also that 2ECON$(G;r)$ and 2NCON$(G;r)$ are full-dimensional [GM].

There is another (technical) reason why we restrict ourselves to full-dimensional polyhedra here. If polyhedra are not full-dimensional, proofs often become more involved technically and statements about nonredundancy of certain systems become quite ugly due to the necessity to exclude equivalent inequalities. This is also true in our case. It is not difficult to derive the results for the lower-dimensional cases from the results presented later. But the statements of these theorems are often rather complicated and we want to avoid unnecessary technicalities.

3. Basic facets. In this section we investigate under which conditions the cut inequalities (1.1)(i), the node-cut inequalities (1.1)(ii), and the trivial inequalities (1.1)(iii) define facets for 2ECON$(G;r)$, respectively, 2NCON$(G;r)$.

An inequality $a^T x \leq a$ is **valid** with respect to a polyhedron P if $P \subseteq \{x \mid a^T x \leq \alpha\}$; the set $F_a := \{x \in P \mid a^T x = \alpha\}$ is called the **face** of P defined by $a^T x \leq \alpha$. If $\dim(F_a) = \dim(P) - 1$ and $F_a \neq \emptyset$, then F_a is a **facet** of P and $a^T x \leq \alpha$ is called **facet-defining** or **facet-inducing.**

The following theorem follows from Theorem 3.3 in [GM] and characterizes which of the trivial inequalities (1.1)(iii) define facets.

THEOREM 3.1. *Let* (G,r) *satisfy* (1.2).

(a) $x_e \leq 1$ *defines a facet of* 2ECON$(G;r)$ *and of* 2NCON$(G;r)$ *for all* $e \in E$.

(b) $x_e \geq 0$ *defines a facet of* 2ECON$(G;r)$ (*respectively,* 2NCON$(G;r)$) *for* $e \in E$, *if and only if for every edge* $f \neq e$ *the polytope* 2ECON$(G-\{e,f\};r)$ (*respectively,* 2NCON$(G-\{e,f\};r)$) *is nonempty.*

The next theorem characterizes the **cut inequalities** (1.1)(i) that define facets.

THEOREM 3.2. *Let* (G,r) *satisfy* (1.2) *and let* $W \subseteq V$ *with* $\emptyset \neq W \neq V$.

(a) *Suppose* $\text{con}(W) = 2$. *Then* $x(\delta(W)) \geq 2$ *defines a facet of* 2ECON$(G;r)$ *if and only if*

(a_1) $G[W]$ *and* $G[V \backslash W]$ *are connected;*

(a_2) $\lambda_1(G[W]) \geq 2$ *and* $\lambda_1(G[V \backslash W]) \geq 2$;

(a_3) e *is a bridge of* $G[W]$. *Then* f *is a bridge of* $G[V \backslash W]$, U, U' *are the node sets of the two components of* $G[W]-e$, *and* $\bar{U}$, $\bar{U}'$ *are the node sets of the two* *components of* $G[V\backslash W] - f$; *and if* $r(U) = r(\bar{U}) = 2$ (*implying* $r(U') = r(\bar{U}') = 0$), *then* $|[U : \bar{U}]| \geq 1$.

(b) *Suppose* $\mathrm{con}(W) = 1$. *Then* $x(\delta(W)) \geq 1$ *defines a facet of* $2\mathrm{ECON}(G;r)$ *if and only if*

(b_1) $G[W]$ *and* $G[V\backslash W]$ *are connected;*

(b_2) $\lambda_1(G[W]) \geq 2$ *and* $\lambda_1(G[V\backslash W]) \geq 2$;

(b_3) $\lambda_2(G[W]) \geq 3$ *and* $\lambda_2(G[V\backslash W]) \geq 3$.

(c) *Suppose* $\mathrm{con}(W) = 0$. *Then* $x(\delta(W)) \geq 0$ *does not define a facet of* $2\mathrm{ECON}(G;r)$ *or of* $2\mathrm{NCON}(G;r)$.

(d) *Suppose* $\mathrm{con}(W) = 2$. *Then* $x(\delta(W)) \geq 2$ *defines a facet of* $2\mathrm{NCON}(G;r)$ *if and only if*

(d_1) *the conditions* (a_1), (a_2), *and* (a_3) *of* (a) *are satisfied;*

(d_2) $\kappa_2(G[W]) \geq 2$ *and* $\kappa_2(G[V\backslash W]) \geq 2$;

(d_3) u *is an articulation node of* $G[W]$ *and* $\bar{u}$ *is an articulation node of* $G[V\backslash W]$, *and if* U *and* $\bar{U}$ *are node sets of components of* $G[W]-u$ *and* $G[V\backslash W]-\bar{u}$, *respectively, such that* $r(U) = r(\bar{U}) = 2$, *then* $|[U : \bar{U}]| \geq 1$, *and* (*because of* (d_2)) *all other components of* $G[W] - u$ *and* $G[V\backslash W] - \bar{u}$ *do not contain nodes of type* 2;

(d_4) *neither for* $S = W$ *nor for* $S = V\backslash W$ *does the following situation occur: There are an edge* $e \in E(S)$ *and nodes* $w_1, w_2 \in S$ (*not necessarily distinct*) *and a node* $w_3 \in V\backslash S$ *so that there exists a component* (S_1, E_1) *of* $(G[S] - e) - w_1$, *a component* (S_2, E_2) *of* $(G[S] - e) - w_2$, *and a component* (S_3, E_3) *of* $G[V\backslash S] - w_3$ *with* $r(S_1) = r(S_2) = r(S_3) = 2$, $S_1 \cap S_2 = \emptyset$, *such that in* $G - e$ *there is no edge between* S_i *and* S_j *for all* $i \neq j$, $i, j \in \{1, 2, 3\}$ (*see Fig.* 3.1 *for an illustration; dashed lines denote nonexisting edges*);

(d_5) *for* $S = W$ *and* $S = V\backslash W$ *the following has to hold: if* $V\backslash S$ *has exactly two neighbors in* S, *then one of these two nodes is the only node of type* 2 *in* S.

(e) *Suppose* $\mathrm{con}(W) = r(W) = 1$. *Then* $x(\delta(W)) \geq 1$ *defines a facet of* $2\mathrm{NCON}(G;r)$ *if and only if*

(e_1) *the conditions* (b_1), (b_2), *and* (b_3) *of* (b) *are satisfied;*

(e_2) $\kappa_2(G[V\backslash W] - e) \geq 2$ *for all* $e \in E(V\backslash W)$.

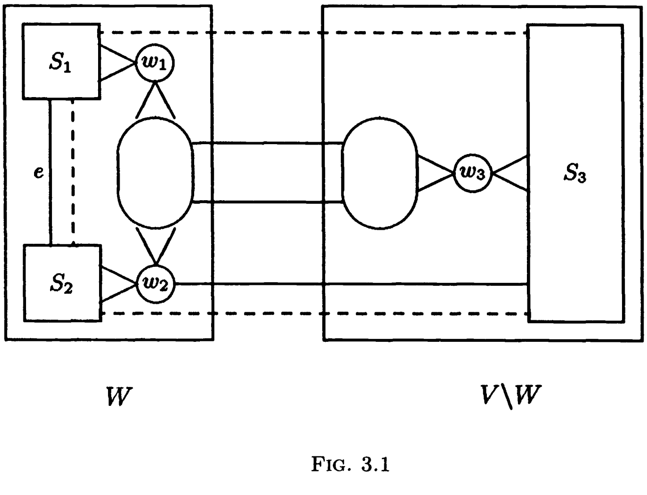


FIG. 3.1

Proof. We give a proof of (d). (The proofs of (a) in the general case, (b), and (e) use the same ideas and are thus omitted. (c) is trivial.)

We first show that if one of the conditions (d_1)–(d_5) is not satisfied, then the cut inequality $x(\delta(W)) \geq 2$ does not define a facet. Necessity of (d_1) is seen easily (see, e.g., Corollary 6.7 of [GM]). Suppose (d_2) is violated. Let u be an articulation node of $G[W]$, and let (S_1, E_1), (S_2, E_2) be two components of $G[S] - u$ with $r(S_1) = r(S_2) = 2$. Then $x(\delta(W)) \geq 2$ can be written as the sum of the node-cut inequalities $x(\delta_{G-u}(S_1)) \geq 1$ and $x(\delta_{G-u}(S_2)) \geq 1$ plus possibly some nonnegativity constraints. Therefore, $x(\delta(W)) \geq 2$ does not define a facet. If (d_3) is violated there are nodes u, $\bar{u}$ and node sets U, $\bar{U}$ with the indicated properties and $[U : \bar{U}] = \emptyset$. In this case the cut inequality can be written as the sum of two other node-cut inequalities $x(\delta_{G-u}(U)) \geq 1$ and $x(\delta_{G-\bar{u}}(\bar{U})) \geq 1$. Hence $x(\delta(W)) \geq 2$ does not define a facet.

Now suppose we have the situation excluded by (d_4) for $S = W$. In this case, it is not possible to construct a feasible solution with $x(\delta(W)) = 2$ and $x_e = 0$, because any feasible set not using e would either have node w_3 as an articulation node or use three edges of $\delta(W)$. Therefore, all feasible sets C with $|C \cap \delta(W)| = 2$ have to use e, so the face defined by $x(\delta(W)) \geq 2$ is contained in the face defined by $x_e \leq 1$. Since 2NCON$(G; r)$ is full-dimensional, these faces cannot be the same. Therefore, $x(\delta(W)) \geq 2$ does not define a facet.

Suppose (d_5) is violated. Let the two neighbor nodes of $V \backslash S$ in S be called u and v. If, in contradiction to (d_5), there is at least one node of type 2 in $S \backslash \{u, v\}$ or $r_u = r_v = 2$, then $x(\delta(W)) \geq 2$ can be written as the sum of the two node-cut inequalities $x(\delta_{G-u}(S \backslash \{u\})) \geq 1$ and $x(\delta_{G-v}(S \backslash \{v\})) \geq 1$.

Now let the conditions of (d) be satisfied for some inequality $a^T x := x(\delta(W)) \geq 2$. Let $b^T x \geq \beta$ be a facet-defining inequality such that the face F_a induced by $a^T x \geq 2$ in 2NCON$(G; r)$ is contained in the facet F_b induced by $b^T x \geq \beta$. Our aim is to show that b is a positive multiple of a, which implies that F_a is identical with the facet F_b.

Let us first state some conditions under which for a given e, $f \in \delta(W)$, the incidence vector of $C_{e,f} := E(W) \cup E(V \backslash W) \cup \{e, f\}$ is feasible for 2NCON$(G; r)$ and hence in $F_a \subseteq F_b$. Assume that both W and $V \backslash W$ contain more than one node of type 2. (In the other case, the proof has to be modified a little.) (1) If e, f are to induce a feasible $C_{e,f}$ they may not have a common endpoint (unless this is the only node of type 2 in W or $V \backslash W$, which we excluded). (2) If we denote the two endpoints of e and f in W with u and v, respectively, then for any node s of type 2 in W there must exist an $[s, u]$-path and an $[s, v]$-path that are node-disjoint; the same for $V \backslash W$.

We can rewrite these conditions in the following way: Let U denote a two-node-connected component of $G[W]$ containing some node of type 2 of $G[W]$. Note that by condition (d_2), U must then contain all nodes of type 2 in W. Now remove from U the set of all articulation nodes of $G[W]$. Let a node set $\bar{U}$ (in $G[V \backslash W]$) be defined in the same way as U in $G[W]$. Condition (2) says that e and f may not be incident to the same component of $G[W] - U$ and $G[V \backslash W] - \bar{U}$. All in all, e and f must constitute a matching of size 2 in the graph G' derived from G by shrinking all components of $G[W] - U$ and $G[V \backslash W] - U$ and deleting all edges except those in $\delta(W)$. The maximum matching possible in this graph has size at least 3, otherwise there are two nodes covering all edges in G', which translates to condition (1.2)(iii), or (d_3) or (d_5) of Theorem 3.2 being violated.

Now we are ready to show that b_e has the same value γ for all $e \in \delta(W)$. Assume that both W and $V \backslash W$ contain more than one node of type 2. G' has a matching with three edges, say, e, f, and g. Since the incidence vectors of $C_{e,f}$, $C_{f,g}$, and $C_{g,e}$ lie in F_b, we have $b_e = b_f = b_g = \gamma$. For any fixed edge $t \in \delta(W) \backslash \{e, f, g\}$ either $\{t, e\}$, $\{t, f\}$, or $\{t, g\}$ constitute a matching in G', say, $\{t, e\}$. Therefore, the incidence vectors of both $C_{t,e}$ and $C_{f,e}$ are in F_b, and we have $b_t = b_f = \gamma$. This way we can prove $b_t = \gamma$ for all $t \in \delta(W)$.

To prove $b_e = 0$ for all $e \in E(W)$ we need to construct a set $C \subseteq E$ with $\chi^C \in F_a$ and $e \notin C$ for some fixed $e \in E(W)$. Since $\chi^{C \cup \{e\}}$ is also in F_a we know $b_e = 0$. Assuming again that both W and $V \backslash W$ have at least two nodes of type 2, we try for a given $e = v_1 v_2 \in E(W)$ to find f, $g \in \delta(W)$ constituting a matching of G', so that $C := C_{f,g} \backslash \{e\}$ is feasible for 2NCON$(G; r)$. If $\kappa_2(G[W] - e) \geq 2$, we can find such f, $g \in \delta(W)$ inducing a feasible $C_{f,g} \backslash \{e\}$ in G by similar arguments as above. Since the incidence vectors of $C_{f,g} \backslash \{e\}$ and $C_{f,g}$ are in F_b, we have $b_e = 0$.

Now suppose $\kappa_2(G[W] - e) = 1$. Consider the tree structure of the two-node–connected components and the articulation points of $G[W] - e$. Since $\kappa_2(G[W]) \geq 2$ and $\kappa_2(G[W] - e) = 1$, the endnodes v_1 and v_2 of e lie in two different two-node–connected components. Furthermore, there is a $[v_1, v_2]$-path in $G[W] - e$ that touches all two-node–connected components containing nodes of type 2 and all articulation nodes of type 2. Let w_1 be an articulation node of $G[W] - e$ so that the component of $(G[W] - e) - w_1$ containing one endnode v_1 of e also contains some node of type 2 (possibly $= v_1$), and so that the node set S_1 of this component is as small as possible with respect to this property. Similarly, find an articulation node w_2 and a component of $(G[W] - e) - w_2$ with node set S_2 containing v_2 and some node of type 2, so that $|S_2|$ is as small as possible. S_1 and S_2 are disjoint. Since G satisfies (1.2) there has to be some edge $f \in \delta(W)$ leaving S_1 and an edge $g \in \delta(W)$ leaving S_2. Since $V \backslash W$ has two nodes of type 2, condition (d_4) ensures that f and g may be chosen without common endpoint, such that in $G[V \backslash W]$ there is no articulation point separating the two endpoints of f and g from some node of type 2. Because S_1 and S_2 are connected in $G[W] - e$ by a path touching all two-node–connected components of $G[W] - e$ containing nodes of type 2, the set $C := C_{f,g} \backslash \{e\}$ defined above is feasible. Therefore, $b_e = 0$. So we have proved that $b = \gamma a$. Since F_b cannot define a facet if $b \leq 0$, we have $\gamma > 0$. So $x(\delta(W)) \geq 2$ and $b^T x \geq \beta$ define the same facet $F_a = F_b$. □

The following theorem characterizes which of the **node-cut inequalities** (1.1)(ii) define facets for 2NCON$(G; r)$.

THEOREM 3.3. *Let* (G, r) *satisfy* (1.2) *and let a node* $z \in V$, *and a set* $W \subseteq V \backslash \{z\}$ *with* $\emptyset \neq W \neq V \backslash \{z\}$ *and* $r(W) = 2$, $r(V \backslash (W \cup \{z\})) = 2$ *be given. Denote by* V_i, $i = 1, 2$, *the set of nodes in* V *of type at least* i, *and let* $\bar{W} := V \backslash (W \cup \{z\})$.

The node-cut inequality $x(\delta_{G-z}(W)) \geq 1$ *defines a facet of* 2NCON$(G; r)$ *if and only if*

(a) $G[W]$ *is connected;*

(b) $\lambda\big(G[W \cup \{z\}], V_1 \cup \{z\}\big) \geq 2$;

(c) $\lambda_2(G[W]) \geq 2$;

(d) $u \in W$ *is an articulation node of* $G[W \cup \{z\}]$ *separating two nodes of* $V_2 \cup \{z\}$, *and* $S \subseteq W$ *is the node set of a component of* $G[W \cup \{z\}] - u$ *with* $r(S) = 2$; *then* $[W \backslash S : \bar{W}] = \emptyset$ *and* $r\big(W \backslash (S \cup \{u\})\big) \leq 1$;

(e) *the following situation does not occur: there are an edge* $e \in E(W)$ *and two nodes* $w_1, w_2 \in W$ (*not necessarily distinct*), *so that, for* $i = 1, 2$, $\big(G[W \cup \{z\}] - e\big) - w_i$ *has a component with node set* S_i *where* $S_i \subseteq W$ *and* $r(S_i) = 2$; *furthermore,* $S_1 \cap S_2 = \emptyset$ *and* $e \in [S_1 : S_2]$ (*see Fig.* 3.2 *for an illustration; dashed lines denote nonexisting edges*);

(f) *conditions* (a), $\cdots$, (e) *also hold when* W *is replaced by* $\bar{W}$.

Proof. The proof is analogous to the proof of Theorem 3.2. □

4. Lifting theorems. We now present conditions under which valid inequalities (respectively, facets) for the 2ECON and 2NCON polytopes on a graph $\hat{G}$ can be lifted

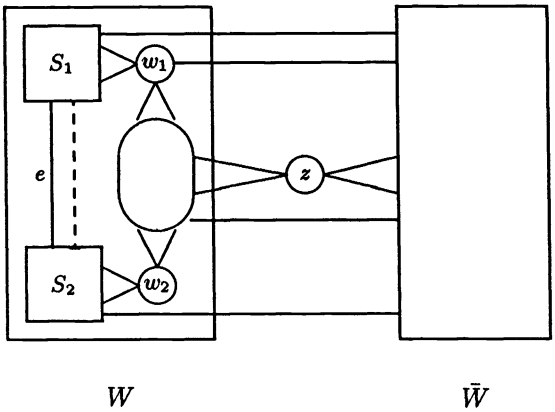


FIG. 3.2

to valid inequalities (respectively, facets) for higher-dimensional 2ECON and 2NCON polytopes on a graph G that contains $\hat{G}$ as a subgraph. These results simplify the proofs to be presented in the next sections.

Some of the results can be treated for the 2ECON and 2NCON polytopes simultaneously. Thus we introduce a slightly more general network design model that combines edge- **and** node-connectivity features. Let $G = (V, E)$ be a graph, $r \in \{0, 1, 2\}^V$ be the vector of connectivity types, and Z be some subset of V. (In this section we do not necessarily assume that $(G; r)$ satisfies (1.2).) We define the **2CON(Z)** problem to be the network design problem where between each pair of distinct nodes s and t at least $\min(r_s, r_t)$ edge-disjoint paths are required that have no node of $Z\backslash\{s,t\}$ in common. Note that for $Z = \emptyset$ only edge-disjoint paths are required, so in this case 2CON(Z) is the 2ECON problem. For $Z = V$ this is the 2NCON problem. This general model is introduced only for technical reasons. Throughout the rest of this paper we will be interested only in the cases $Z = \emptyset$ and $Z = V$.

The 2CON(Z) problem can be formulated as an integer linear program in the following way:

$$\min \sum_{ij \in E} c_{ij} x_{ij}$$

subject to

$$(4.1)\quad \begin{array}{lll} \text{(i)} & x(\delta(W)) \geq \text{con}(W) & \text{for all } W \subseteq V, \emptyset \neq W \neq V; \\ \text{(ii)} & x(\delta_{G-z}(W)) \geq 1 & \text{for all } z \in Z, \text{ and for all } W \subseteq V\backslash\{z\},\ \emptyset \neq W \neq \\ & & V\backslash\{z\} \text{ with } r(W) = 2 \text{ and } r(V\backslash(W \cup \{z\})) = 2; \\ \text{(iii)} & 0 \leq x_{ij} \leq 1 & \text{for all } ij \in E; \\ \text{(iv)} & x_{ij} \text{ integral} & \text{for all } ij \in E. \end{array}$$

The polytope **2CON**$(G; Z; r)$ is then defined as the convex hull of all $x \in \mathbb{R}^E$ that satisfy (i),$\cdots$,(iv) of (4.1). As mentioned above, 2CON$(G; \emptyset; r)$ = 2ECON$(G; r)$ and 2CON$(G; V; r)$ = 2NCON$(G; r)$.

The polytope 2CON$(G; Z; r)$ is not necessarily full-dimensional. In the later sections we only apply the results of this section in the case dim(2CON$(G; Z; r)$) = $|E|$. So we can avoid treating all the technicalities arising in the low-dimensional case, and we thus assume throughout this section that 2CON$(G; Z; r)$ has dimension $|E|$.

In Lemma 4.2 we derive valid inequalities for the $2\text{CON}(G;Z;r)$ polytope from valid inequalities for the $2\text{CON}(G/W;Z;r)$ polytope.

LEMMA 4.2. *Consider the* $2\text{CON}(G;Z;r)$ *polytope and let* $W \subseteq V \backslash Z$. *Let the node* w *in* G/W *that represents node set* W *inherit its connectivity type from* W *by* $r_w := \text{con}(W)$. *If* $\hat{a}^T x \geq b$ *is a valid inequality for* $2\text{CON}(G/W;Z;r)$ *where* $W \subseteq V \backslash Z$, *then* $a^T x \geq b$ *is valid for* $2\text{CON}(G;Z;r)$, *where*

$$a_e = \hat{a}_e \quad \textit{for } e \in E(G/W) \quad \textit{and} \quad a_e = 0 \quad \textit{for } e \in E(W).$$

We say that $a^T x \geq b$ *is obtained from* $\hat{a}^T x \geq b$ *by expanding* w *to* W.

Proof. We first remark that the lemma is true for any of the inequalities (i), (ii), or (iii) of (4.1). The reason is that the expansion of any inequality of type (i), (ii), or (iii) is again of the same type. (Note that since $Z \cap W = \emptyset$, a shrunk node w can never be chosen as a node z in a node-cut inequality (ii).)

Since $2\text{CON}(G/W;Z;r)$ is the convex hull of the integral solutions of (i), (ii), and (iii) of (4.1) every valid inequality for $2\text{CON}(G/W;Z;r)$ can be obtained by taking nonnegative combinations of the inequalities (i), (ii), and (iii), rounding the left- and right-hand sides up and recursively repeating this procedure. (This so-called cutting plane proof is described in [Chv]; see also [Sch, Cor. 23.2b].) It is easy to see that such a validity proof of $\hat{a}^T x \geq b$ from the inequalities (i), (ii), and (iii) of (4.1) for $2\text{CON}(G/W;Z;r)$ yields a validity proof of $a^T x \geq b$ by applying the same nonnegative combinations and rounding operations to the associated expanded inequalities, since combining and rounding expanded inequalities produces expanded inequalities. □

The following lemma gives a technical condition for an "expanded" inequality derived by Lemma 4.2 to define a facet of $2\text{CON}(G;Z;r)$.

LEMMA 4.3. *Consider the* $2\text{CON}(Z)$ *problem given by* (G,r) *satisfying* (1.2)(i) *and let* W *be a subset of* $V \backslash Z$ *with* $G[W]$ *connected. Let the node* w *in* G/W (*representing the node set* W) *inherit its connectivity type from* W *by* $r_w := \text{con}(W)$. (*Note that* $(G/W,r)$ *does not necessarily satisfy* (1.2)(i).)

Let the inequality $\hat{a}^T x \geq \alpha > 0$ *be valid for* $2\text{CON}(G/W;Z;r)$ *and let* $a^T x \geq \alpha$ *be the inequality* (*valid for* $2\text{CON}(G;Z;r)$) *obtained from* $\hat{a}^T x \geq \alpha > 0$ *by expanding node* $w \notin Z$ *to* $W \subseteq V(G)$.

Denote by F_a *the face of the polytope* $P := 2\text{CON}(G;Z;r)$ *induced by* $a^T x \geq \alpha$ *and by* $F_{\hat{a}}$ *the face of the polytope* $\hat{P} := 2\text{CON}(G/W;Z;r)$ *induced by* $\hat{a}^T x \geq \alpha$.

F_a *is a facet of* P *if and only if the following conditions hold:*

(a) *For any* $e \in E(W)$ *there exists a set* $\hat{C} \subseteq E(G/W)$ *with* $\chi^{\hat{C}} \in F_{\hat{a}}$ *so that the incidence vector of* $\hat{C} \cup E(W) \backslash \{e\}$ *lies in* F_a.

(b) *There exist* $s := |E(G/W)|$ *sets* $C_i \in E(G/W)$, $i = 1, \cdots, s$, *with* $\chi^{C_i} \in F_{\hat{a}}$ *so that*

(b$_1$) $\chi^{C_i \cup E(W)} \in F_a$, *and*

(b$_2$) *the* χ^{C_i} *are affinely independent.*

Proof. Suppose that (a) and (b) are satisfied. We want to show that F_a is a facet. (Note that (b) implies that $F_{\hat{a}}$ is a facet.) Let $b^T x \geq \beta$ define a facet F_b of P that contains F_a. For any $e \in E(W)$, condition (a) provides a set C with $e \notin C$ and $\chi^C \in F_b$. Therefore, $\chi^{C \cup \{e\}} \in F_b$ and $b_e = 0$ also. Condition (b) implies that vector b has to satisfy $b^T \chi^{C_i \cup E(W)} = \beta$ for $i = 1, \cdots, s$. Since we have just proved b^T to be $(0, \hat{b}^T)$ with $\hat{b} \in \mathbb{R}^{E(G/W)}$, this means $\hat{b}^T \chi^{C_i} = \beta$ for $i = 1, \cdots, s$. The affine independence of the $\dim(\hat{P})$ vectors χ^{C_i} implies that $\hat{b}^T x \geq \beta$ defines a facet of $\hat{P}$, necessarily the same as $F_{\hat{a}}$. Therefore, $(\hat{b}^T, \beta)$ is a positive multiple of $(\hat{a}^T, \alpha)$, and (b^T, β) is a positive multiple of (a^T, α). So F_a defines a facet.

On the other hand, if we know that $a^T x \geq \alpha$ defines a facet of P, then for each $e \in E(W)$ there must exist a set C with $e \notin C$ and $\chi^C \in F_a$; otherwise $F_a \subseteq \{x \in P : x_e = 1\}$. If we shrink node set W to node w in the graph defined by C we arrive at a set $\hat{C} := C \backslash E(W)$ whose incidence vector satisfies $\hat{a}^T x = \alpha$ and is feasible for $\hat{P}$ because $\text{con}(w) = \text{con}(W)$. w may be an articulation node in C, but this does not matter because $w \notin Z$. The set $\hat{C} \cup E(W) \backslash \{e\}$ is feasible for P because it contains the feasible set C and because $G[W]$ is connected. Therefore, (a) is satisfied.

If F_a is a facet of P, there exist $|E|$ affinely independent vertices χ^{C_i} in F_a, where $C_i \subseteq E$ is feasible for P, for $i = 1, \cdots, |E|$. We set $x_i := \chi^{C_i}$ for $i = 1, \cdots, |E|$. There must be a subset of $|E(G/W)|$ affinely independent vectors among $\hat{x}_1, \cdots, \hat{x}_{|E|}$, where $\hat{x}_i$ is derived from x_i for $i = 1, \cdots, |E|$ by deleting the components $e \in E(W)$. The $\hat{x}_i$, for $i = 1, \cdots, |E|$, are feasible for $\hat{P}$ because the deletion of the $E(W)$-components of a vector x in $\{0,1\}^E$ is equivalent to the contraction of W in the subgraph (V, F^x) of G defined by x. So the affinely independent subset of $\{\hat{x}_i : i = 1, \cdots, |E|\}$ satisfies (b_1) and (b_2). □

The conditions of Lemma 4.3 can be used to derive some conditions on $G[W]$ that are of a more graph-theoretical nature and sufficient for an "expanded" inequality to define a facet of $2\text{ECON}(G; r)$.

LEMMA 4.4. *Consider the* 2ECON *problem given by* (G, r) *satisfying* (1.2)(i). *Let* $W \subseteq V$ *with* $\emptyset \neq W \neq V$, *and let* w (*of type* $\text{con}(W)$) *be the node of* G/W *representing* W. *Consider an inequality* $\hat{a}^T x \geq b$ *that is facet-defining for the polytope* $2\text{ECON}(G/W; r)$, *and consider the inequality* $a^T x \geq b$ (*valid for* $2\text{ECON}(G; r)$) *derived from* $\hat{a}^T x \geq b$ *by expanding node* w *to* W.

If $G[W]$ *is* $\max\{2, r(W) + 1\}$*-edge-connected then* $a^T x \geq b$ *defines a facet of* $2\text{ECON}(G; r)$.

Proof. Let F_a and $F_{\hat{a}}$ be defined as in Lemma 4.3. We will check conditions (a) and (b) of Lemma 4.3. The connectivity conditions on $G[W]$ imply that for any $e \in E(W)$ and $\hat{C} \subseteq E(G/W)$ that is feasible for $2\text{ECON}(G/W; r)$, the sets $\hat{C} \cup E(W) \backslash \{e\}$ and $\hat{C} \cup E(W)$ are feasible for $2\text{ECON}(G; r)$. Since $F_{\hat{a}}$ is a facet, there are enough affinely independent $\chi^{\hat{C}}$ to satisfy condition (b) of Lemma 4.3. □

Usually much weaker conditions on the edge-connectivity of $G[W]$ are already sufficient for an expanded inequality $a^T x \geq b$ to define a facet of $2\text{ECON}(G; r)$. But this leads to further technicalities concerning assumptions on the structure of the graph and properties of $\hat{a}^T x \geq b$; see, for instance, Theorem 3.2(a) and (b).

The next lemma gives a sufficient condition for an expanded inequality to define a facet of $2\text{NCON}(G; r)$. (Note that any inequality valid for $2\text{CON}(G; Z; r)$ is also valid for $2\text{NCON}(G; r)$.)

LEMMA 4.5. *Consider the* 2NCON *problem given by* (G, r) *satisfying* (1.2)(i). *Let* $Z \subseteq V$ *and* $W \subseteq V \backslash Z$ *with* $\emptyset \neq W \neq V$ *and* $r(W) = 1$, *and let* w (*of type* 1) *be the node of* G/W *representing* W. *Consider an inequality* $\hat{a}^T x \geq b$ *that is valid for the polytope* $2\text{CON}(G/W; Z; r)$ *and facet-defining for* $2\text{NCON}(G/W; r)$.

If $G[W]$ *is two-edge-connected, then the inequality* $a^T x \geq b$ *derived from* $\hat{a}^T x \geq b$ *by expanding node* w *to* W *defines a facet of* $2\text{NCON}(G; r)$.

Proof. First, $a^T x \geq b$ is valid for $2\text{CON}(G; Z; r)$ by Lemma 4.2 and hence for $2\text{NCON}(G; r)$. To prove that $a^T x \geq b$ also defines a facet of $2\text{NCON}(G; r)$, we apply Lemma 4.3 with $\hat{P} := 2\text{CON}(G/W; V; r)$ and $P := 2\text{CON}(G; V; r)$. Conditions (a) and (b) are still sufficient for $a^T x \geq b$ to define a facet of P, because of the fact that $w \notin Z$ is not used in the sufficiency part of Lemma 4.3. So we have to check (a) and (b) of Lemma 4.3, which is easy. □

Our final lifting result presents conditions under which a valid inequality for

$2\mathrm{CON}(G;\ Z;\ r)$ on a complete graph $G = (V, E)$ can be extended to the graph with a new node w of type at least 1 added, along with all of the edges incident between w and V; we denote such a graph by $G + w$.

LEMMA 4.6. *Consider the* $2\mathrm{CON}(Z)$ *problem given by a graph* $\hat{G}$ *and node types* r *satisfying* (1.2)(i), *where* $\hat{G} = (\hat{V}, \hat{E})$ *is a complete graph with two parallel edges* uv *for each* $u,\ v \in V$ *with* $u \neq v$. *Let* $\hat{a}^T x \geq \hat{b}$ *be a valid inequality for* $2\mathrm{CON}(\hat{G}; Z; r)$ *with* $\hat{a} \geq 0$. *Let* $W \subseteq \hat{V} \backslash Z$ *be a node set with* $r(W) = 2$ *and* α *some nonnegative value so that either* $\hat{a}_e = \alpha$ *for all* $e \in \hat{E}(W)$ *or* $|W| = 1$.

We define an inequality $a^T x \geq b$ *on the graph* $G := \hat{G} + w$ *with* $r_w := 1$ *by setting*

$$\begin{aligned} b &:= \hat{b} + \alpha, \\ a_e &:= \hat{a}_e \quad \textit{for all } e \in \hat{E}, \\ a_{uw} &:= \alpha \quad \textit{for all } u \in W, \\ a_{uw} &:= \beta_u := \max\{\alpha, \max\{\hat{a}_{uv},\ v \in W\}\} \quad \textit{for all } u \notin W. \end{aligned}$$

If $a_{uw} + a_{vw} \geq \alpha + a_{uv}$ *for all distinct nodes* $u, v \notin W$, *then* $a^T x \geq b$ *is valid for* $2\mathrm{CON}(G; Z; r)$.

Note that in Lemma 4.6 the restriction to complete graphs is no restriction at all, because any inequality valid for $2\mathrm{CON}(G; Z; r)$, where G is a complete graph, is also valid if G is replaced by some subgraph (V, F). In the lemma we need completeness of $\hat{G}$ to compute the β_u correctly. Also, we can restrict ourselves without loss of generality to $\hat{a} \geq 0$ because it is easy to see that any inequality $\hat{a}^T x \geq \hat{b}$ that is facet-defining for $2\mathrm{CON}(\hat{G}; Z; r)$ (except $-x_e \geq -1$) has nonnegative coefficients.

Proof. We will assume that $r_w = 1$, because validity of an inequality in this case implies its validity if $r_w = 2$. Assume further that $a^T x \geq b$ is not valid, i.e., that there exists an edge set C that is feasible for $2\mathrm{CON}(G; Z; r)$ and does not satisfy $a^T \chi^C \geq b$.

(1) If there is an edge $uw \in C$ with $u \in W$, we contract node set $\{u, w\}$ to node u. The resulting subgraph of $\hat{G}$ with edge set $C \backslash \{uw\}$ is feasible for $2\mathrm{CON}(\hat{G}; Z; r)$. Note that $\hat{a}_{vu} \leq a_{vw}$ for all $v \in V$. Therefore, $\hat{a}^T \chi^{C/\{u,w\}} \leq a^T \chi^C - a_{uw} < b - \alpha = \hat{b}$. But then $\hat{a}^T x \geq \hat{b}$ is not valid for $2\mathrm{CON}(\hat{G}; Z; r)$, a contradiction.

If C uses no edge of $[W : \{w\}] \cap C$, we will show how to replace C by some set C' containing an edge in $[W : \{w\}]$, such that $a^T \chi^{C'} \leq a^T \chi^C < b$. So we can apply the argumentation above to derive a contradiction to the validity of $\hat{a}^T x \geq \hat{b}$.

(2) Suppose all edges of $\delta(w) \cap C$ were bridges of (V, C). Since w is connected to W in C, there must be a bridge uw of (V, C), which separates w from some node $v \in W$. The set $C' := (C \backslash \{uw\}) \cup \{vw\}$ is feasible for $2\mathrm{CON}(G; Z; r)$ and contains an edge of $[W : \{w\}]$. Moreover, $a^T \chi^{C'} = a^T \chi^C - a_{uw} + a_{vw} \leq a^T \chi^C < b$.

Now suppose there are edges of $\delta(w) \cap C$ that are not bridges of (V, C). Define U as the set of nodes that are incident to nonbridges of C (the so-called two-connected part of C). U must contain all nodes of type 2. By assumption, w belongs to U.

(3) Assume that w is not an articulation node of (V, C) disconnecting two nodes of type 2. The case that w is an articulation node is treated separately. Since $r(W) = 2$, there exists a node $s \in W$ of type 2, and since s and w are in U, there exist two edge-disjoint $[s, w]$-paths in C that do not coincide in any node $z \in Z$. Let $u, v \in U$ be the nodes adjacent to w on these two paths. If $u = v$, we eliminate one of the two wv-edges. This can be done without destroying feasibility of C because w is not an articulation node separating two nodes of type 2 and because $r_w = 1$. Also, $a^T \chi^C$ does not increase with this operation, since $a \geq 0$. Now we are either in the case that $\delta(w) \cap C$ contains only bridges of (V, C) (proceed with part (2) of the proof), or we construct two other $[s, w]$-paths that lead to different nodes $u \neq v$.

Now we show that $C' := (C\backslash\{uw, vw\}) \cup \{ws, uv\}$ is also feasible. Clearly C' is connected, so we only have to check for bridges and articulation nodes. Suppose that e is a bridge of (V, C') separating two nodes of type 2. In $C'\backslash\{e\}$, node s is connected to u and v by at least one of the two edge-disjoint paths and edge uv. If $e \neq ws$, all four nodes w, s, u, v lie in the same component (S, F) of $(V, C'\backslash\{e\})$. Since $C' \cap \delta(S) = C \cap \delta(S)$, edge e is also a bridge in (V, C) separating two nodes of type 2. So e must be ws. But $(V, C) - w$ is a subgraph of $(V, C') - w$, and w is not an articulation node of (V, C). Now suppose that $z \in Z$ is an articulation node that separates two nodes of type 2 in (V, C') but not in (V, C). $z = s$ need not be considered, because $s \notin Z$. The remaining cases lead to a contradiction similar to the case in which e is a bridge. So C' is feasible for $2\text{CON}(\hat{G}; Z; r)$. C' also satisfies $a^T\chi^{C'} < b$ because $a^T\chi^{C'} = a^T\chi^C - (a_{uw} + a_{vw} - a_{uv}) + a_{ws} \leq a^T\chi^C - \alpha + \alpha$.

(4) The last remaining case in our transformation of C is the case in which w is an articulation node of (V, C) separating two nodes of type 2. Let $u, v \in U$ be nodes adjacent to w lying on different sides of $(V, C) - w$. Replace C by $C' = (C\backslash\{uw, vw\}) \cup \{uv\}$. C' is feasible, $a^T\chi^{C'} \leq a^T\chi^C$, and $(V, C') - w$ contains one component less than $(V, C) - w$. Ultimately, we reach a set C' where w does not separate any nodes of type 2, and we can apply one of the earlier cases. Thus, we have proved that if $a^Tx \geq b$ is not valid for $2\text{CON}(G; Z; r)$, then also $\hat{a}^Tx \geq \hat{b}$ is not valid for $2\text{CON}(\hat{G}; Z; r)$. □

The next theorem gives sufficient conditions for $a^Tx \geq b$ to define a facet.

THEOREM 4.7. *Consider the situation in Lemma* 4.6, *where we have an inequality* $\hat{a}^Tx \geq \hat{b}$ *valid for* $2\text{CON}(K_n; Z; r)$, *and where* (K_n, r) *satisfies* (1.2)(i). *Let* W, w *with* $r_w = 1$, $\alpha \geq 0$, *be defined as in Lemma* 4.6. *Let* $a^Tx \geq b$ *be the inequality derived from* $\hat{a}^Tx \geq \hat{b}$ *by the formula in Lemma* 4.6. *Furthermore, let* $G = (V, E)$ *be a subgraph of* K_{n+1} *with* $n + 1$ *nodes, and define* $\hat{G}$ *as* $G - w$.

Then, for any $Z' \supseteq Z$, *the inequality* $a^Tx \geq b$ *defines a facet of* $2\text{CON}(G; Z'; r)$ *if the following conditions hold:*

(a) $\hat{a}^Tx \geq \hat{b}$ *defines a facet of* $2\text{CON}(\hat{G}; Z'; r)$;

(b) *for all* $u \notin W$ *with* $uw \in E$ *and* $a_{uw} > \alpha$ *there exists a node* $v \in W$ *with* $a_{uw} = a_{uv}$ *and* $uv, vw \in E$;

(c) *there exist two distinct nodes* u, v *with* $a_{uv} = a_{vw} = a_{uw} = \alpha$, *and* uv, vw, $uw \in E$;

(d) *all nodes* u *with* $uw \in E$ *and* $a_{uw} = \alpha$ *have type at least* 1.

Proof. First, note that $a^Tx \geq b$ is valid for $2\text{CON}(G; Z'; r)$ because it is valid for $2\text{CON}(G; Z; r)$ by Lemma 4.6. We prove the theorem by exhibiting $|E|$ affinely independent vectors in $F_a := \{x \in 2\text{CON}(G, Z'; r) \mid a^Tx = b\}$. Let $F_{\hat{a}} := \{x \in 2\text{CON}(\hat{G}, Z'; r) \mid \hat{a}^Tx = \hat{b}\}$.

Let $f = vw$ be an edge $a_f = \alpha$ and $r_v \geq 1$. This edge exists by condition (d). Then any set $\hat{C} \subseteq \hat{E}$ feasible for $F_{\hat{a}}$ can be enlarged to a set $C \subseteq E$ feasible for F_a by adding f. This way we can create $|\hat{E}|$ affinely independent vectors in F_a. Now we want to exhibit $|\delta(w)| - 1$ sets C_k with $\chi^{C_k} \in F_a$. The C_k are characterized by the fact that C_k contains an edge $e_k \in \delta(w)\backslash\{f\}$ that is not contained in any of the previous C_i, $i = 1, \cdots, k - 1$. This fact implies that each χ^{C_k} will be affinely independent from all χ^{C_i}, $i = 1, \cdots, k - 1$, and the $|\hat{E}|$ vectors already found in F_a. The C_k are constructed as follows. Order the edges in $\delta(w)\backslash\{f\}$ as e_1, e_2, etc., by increasing a_e-values, so that the edges e with $a_e = \alpha$ come first. Now, for an edge $e_k \in \delta(w)\backslash\{f\}$ with $a(e_k) = \alpha$ and $e_k = v_kw$, let $\hat{C} \subseteq \hat{E}$ be a set with incidence vector in $F_{\hat{a}}$ and set $C_k := \hat{C} \cup \{e_k\}$. For an edge $e_k = uw \in \delta(w)\backslash\{f\}$ with $a_{uw} > \alpha$, let uv be the edge with $a_{uv} = a_{uw}$, existing by condition (b). Let $\hat{C} \subseteq \hat{E}$ be a set with $\chi^{\hat{C}} \in F_{\hat{a}}$, which uses uv. The incidence vector of the set $C_k := (\hat{C} \setminus \{uv\}) \cup \{uw, wv\}$ is then in F_a. Therefore, we can create the proposed $|\delta(w)| - 1$ sets C_k.

We still have to exhibit one more vector in F_a that is independent from all the others. By condition (c) there is a triangle $uv, vw, uw \in E$ with $a_e = \alpha$ for all triangle edges. As before, there is a set $\hat{C}$ with $\chi^{\hat{C}} \in F_{\hat{a}}$ and $uv \in \hat{C}$. The incidence vector of $C := (C \setminus \{uv\}) \cup \{uw, wv\}$ is affinely independent of all the others already found in F_a, because these all satisfy $x\big([\{w\} : S]\big) = 1$ for $S = \{u \mid uw \in E, a_{uw} = \alpha\}$. So we have found $|E|$ affinely independent vectors in F_a. □

5. Partition inequalities for 2ECON and 2NCON.

5. Partition inequalities for 2ECON and 2NCON. In this section we introduce a class of inequalities that is motivated by the partition inequalities for the connected subgraph polytope (see [GM]), and that generalizes cut inequalities.

DEFINITION 5.1. Let $G = (V, E)$ be a graph and $r \in \{0, 1, 2\}^V$. We call a collection $W_1, \cdots, W_p$ of subsets of V a **proper partition** of V if

— $W_i \neq \emptyset$, $i = 1, \cdots, p$,

— $W_i \cap W_j = \emptyset$, $1 \leq i < j \leq p$,

— $\cup_{i=1}^{p} W_i = V$,

— $r(W_i) \geq 1$, $i = 1, \cdots, p$.

The **partition inequality** induced by a proper partition $W_1, \cdots, W_p$ is given by

$$(5.2) \qquad \frac{1}{2}\sum_{i=1}^{p} x(\delta(W_i)) \geq \begin{cases} p, & \text{if } r(W_i) = 2 \text{ for at least two node sets } W_i, \\ p-1 & \text{otherwise.} \end{cases}$$

See Fig. 5.1 for an illustration of a partition inequality with four node sets $W_1, \cdots, W_4$. Here and in all following illustrations, node sets W with $r(W) = 2$ are depicted by big squares, and node sets W with $r(W) = 1$ are depicted by big circles. Nodes of types 2 and 1 are depicted by small squares and circles, respectively.

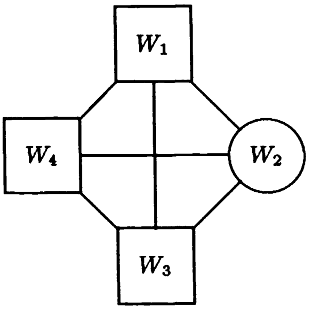


FIG. 5.1

The following observation follows immediately from the definition.

Remark 5.3. Any partition inequality (5.2) induced by a proper partition is valid for 2ECON$(G; r)$ and 2NCON$(G; r)$.

Note that a partition inequality induced by a proper partition with $p = 2$ is nothing but a cut inequality $x(\delta(W)) \geq \text{con}(W)$. The next observation indicates that we cannot expect to obtain a useful characterization of those partition inequalities that define facets.

Remark 5.4. Checking whether a partition inequality supports 2ECON$(G;r)$ or 2NCON$(G;r)$ is NP-complete.

Proof. The problem is obviously in NP. Let $G=(V,E)$ be a graph and $r_v=2$ for all $v\in V$. Then the sets $\{w\}$, $w\in V$, form a proper partition of V and the induced partition inequality reads $x(E)\geq |V|$. Thus there is a point in 2ECON$(G;r)$ or 2NCON$(G;r)$ that satisfies $x(E)\geq |V|$ with equality if and only if G is Hamiltonian. This implies the remark. □

We will now derive a sufficient condition for a partition inequality to define a facet.

THEOREM 5.5. *Let $G=(V,E)$ be a graph, $r\in\{0,1,2\}^V$, and let $W_1,\cdots,W_p$, $p\geq 3$, be a proper partition (see* (5.1)*). Let $\hat{G}=(\hat{V},\hat{E})$ be the graph $G/W_1/\cdots/W_p$ where the W_i are shrunk to nodes w_i of connectivity type $\hat{r}(w_i):=\text{con}(W_i)$ for $i=1,\cdots,p$. Let V_1 be the set of nodes of type at least* 1 *in $\hat{G}$ and V_2 the set of nodes of type* 2 *in $\hat{G}$. The partition inequality* (5.2) *defines a facet of* 2ECON$(G;r)$ *if*

(a) *$\kappa_2(\hat{G})\geq 3$ and $\kappa_1(\hat{G})\geq 2$;*

(b) *in $\hat{G}$ every node of type* 2 *is adjacent to some node of type* 1;

(c) *$\hat{G}[V_1\backslash V_2]$ is connected;*

(d) *$\hat{G}[V_2]$ is Hamiltonian;*

(e) *$G[W_i]$ is $(r(W_i)+1)$-edge-connected for $i=1,\cdots,p$.*

Proof. The partition inequality (5.2) can be written as $x(\hat{E})\geq t$ for the graph $\hat{G}$, where $t=|\hat{V}|$ or $|\hat{V}|-1$, according to whether $\hat{G}$ contains nodes of type 2 or not. If $\hat{G}$ contains only nodes of type 1 and $\hat{G}$ is two-node–connected (see condition (a)), the partition inequality $x(\hat{E})\geq|\hat{V}|-1$ defines a facet of the polytope of connected subgraphs of $\hat{G}$. This was shown in [GM]. By our lifting Lemma 4.4 and Theorem 5.5(e), we can expand all nodes w_i of $\hat{G}$ successively to node sets W_i, and thus obtain a facet of the 2ECON$(G;r)$ polytope.

Suppose that $\hat{G}$ contains nodes of type 2. First we show that conditions (a)–(d) are sufficient for $x(\hat{E})\geq|\hat{V}|$ to define a facet F of 2NCON$(\hat{G};\hat{r})$. We do this by constructing $|\hat{E}|$ affinely independent vectors in F.

Take some Hamiltonian cycle C of $\hat{G}[V_2]$. Let $G'=(V',E')$ denote the graph $\hat{G}/V_2$. Any tree T spanning the nodes of the shrunk graph G' may be added to C, thus creating a set whose incidence vector is in F. There are at least $|E'|$ such trees with affinely independent incidence vectors. This is true because the inequality $x(E')\geq|V'|-1$ defines a facet of the polytope of connected subgraphs of G' (see [GM]) if G' is two-node–connected. Note that G' **is** two-node–connected because $\hat{G}$ is two-node–connected by condition (a), and because $\hat{G}[V_1\backslash V_2]$ is connected. Hence we can find $|E'|$ affinely independent vectors of the form $\chi^{C\cup T}$ in F.

Now take some cycle edge $e\in C$. With the help of conditions (b) and (c) we can construct a cycle not using e and spanning all nodes of type 2 in $\hat{G}$ by using the path $C\backslash\{e\}$ and a path in G'. This new cycle may be augmented by some trees to a feasible set with incidence vector in F. This vector is affinely independent of all other vectors constructed so far because these all satisfied $x_e=1$. By applying this argument for each cycle edge successively we can construct $|E'|+|C|$ affinely independent vectors in F.

For any other edge $e=uv\in\hat{E}(V_2)\backslash C$ we want to construct a cycle spanning all nodes of type 2, and using e but no other edge of $\hat{E}(V_2)\backslash C$. This can be done easily by starting with u, going to v, running in some direction along the cycle C to the neighbor of u, taking a path in E' to the neighbor of v on C that is not already visited, and running along the other half of C to the starting point u. This cycle can be augmented to a set with incidence vector in F. This vector is affinely independent of all the others exhibited so far because all of those satisfied $x_e = 0$. So we have $|E'| + |\hat{E}(V_2)| = |\hat{E}|$ affinely independent vectors in F. This proves that $x(\hat{E}) \geq |\hat{V}|$ defines a facet of $2\text{NCON}(\hat{G}; \hat{r})$, and hence of $2\text{ECON}(\hat{G}; \hat{r})$.

Our partition inequality (5.2) in G can be obtained from $x(\hat{E}) \geq |\hat{V}|$ by expanding successively the nodes w_i to node sets W_i according to the definition in Lemma 4.2. Because of Theorem 5.5(e) we can apply Lemma 4.4, and thus the partition inequality (5.2) defines a facet of $2\text{ECON}(G; r)$. □

Remark 5.6. The partition inequality (5.2) defines a facet of $2\text{NCON}(G; r)$ if G is complete and no node set W_i with $r(W_i) = 2$ contains exactly two nodes.

Proof. The proof is the same as for Theorem 5.5 except that in the end we use Lemma 4.3 instead of Lemma 4.4. □

In view of Theorem 3.2 (d), which gives quite complicated necessary and sufficient conditions for a cut inequality $x(\delta(W)) \geq 2$ to define a facet of $2\text{NCON}(G; r)$, we did not further investigate necessary and sufficient conditions for a partition inequality (with $p > 2$) to define a facet of $2\text{NCON}(G; r)$.

The next theorem shows which of the sufficient conditions of Theorem 5.5 are actually necessary for a partition inequality to define a facet of $2\text{ECON}(G; r)$.

THEOREM 5.7. *Let* (G, r) *and a proper partition* $W_1, \cdots, W_p$ *with* $p \geq 3$ *be given, and let* $\hat{G}$ *and* $\hat{r}$ *be defined as in Theorem* 5.5. *The partition inequality* (5.2) *defines a facet of* $2\text{ECON}(G; r)$ *only if*

(a) *conditions* (a) *and* (b) *of Theorem* 5.5 *are satisfied;*

(b) $\hat{G}$ *contains nodes of type* 2*; then* $\hat{G}$ *contains a cycle* C *containing all nodes of type* 2*;*

(c) $G[W_i]$ *is connected for* $i = 1, \cdots, p$*;*

(d) $\lambda_1(G[W_i]) \geq 2$ *for* $i = 1, \cdots, p$.

Proof. The necessity of condition (a) of Theorem 5.5 is easily seen. Suppose that condition (b) of Theorem 5.5 is violated and that $|W_i| = 1$ for all $i = 1, \cdots, p$. This implies that $\hat{G} = G$ and that there is a node $v \in V_2$ that is adjacent only to other nodes of type 2 in G. Then any set C that is feasible for $2\text{ECON}(G; r)$ with $|C| = |V|$ has to use exactly two edges of $\delta(v)$. Otherwise C would have at least two cycles, and this would imply $|C| \geq |V| + 1$. So the face induced by the partition inequality $x(E) \geq |V|$ is contained in the face induced by $x(\delta(v)) \geq 2$. But since the partition was supposed to consist of at least three sets, the partition inequality does not define the same face as the cut inequality. If $|W_i| \geq 2$ for some i and if Theorem 5.5(b) is violated, one can argue similarly.

The necessity of conditions (b) and (c) of Theorem 5.7 is easily seen. As for (d), suppose that some $G[W_i]$ contains a bridge e so that $G[W_i] - e$ has two components with node sets U and W, with $r(U) \geq 1$ and $r(W) \geq 1$. In this case our partition inequality can be written as the sum of $x_e \leq 1$ and another partition inequality can be defined by the same partition as above, except that W_i is replaced by U and W. □

From an algorithmic point of view, Remark 5.4 seems to be bad news. Even worse, the separation problem for partition inequalities is NP-complete (see [GMS]). But in practice, using heuristic separation routines, the class of partition inequalities proved to be very useful in the cutting plane algorithm presented in [GMS]. Usually, partitions with a small number of node sets were used there, and for small p it is quite likely that—in our real-world examples—a partition inequality supports $2\text{NCON}(G; r)$.

Moreover, checking the conditions of Theorem 5.7 is easy, and this helps to convert one partition inequality into another partition inequality that induces a face of higher dimension than the first one. Indeed, finding cutting planes that induce faces of dimension as high as possible is of importance in cutting plane algorithms. We noticed this clearly in our computational experiments (see [GMS]).

6. Node-partition inequalities. We now generalize node-cut inequalities to "node-partition inequalities" in the same way as we generalized cut inequalities to partition inequalities in the previous section. These new inequalities will only be valid for $2\mathrm{NCON}(G;r)$, but, in general, not for $2\mathrm{ECON}(G;r)$.

Let $G = (V,E)$ be a graph and $r \in \{0,1,2\}^V$. Let $z \in V$ and let $W_1, \cdots, W_p$ be a proper partition (see (5.1)) of $V \setminus \{z\}$ such that at least two node sets W_i contain nodes of type 2. The following **node-partition inequality** induced by z and $W_1, \cdots, W_p$ is given by

$$(6.1) \qquad \frac{1}{2}\Big(\sum_{i\in I_2} x(\delta_{G-z}(W_i)) + \sum_{i\in I_1} x(\delta_G(W_i)) + x([\{z\} : \cup_{i\in I_1} W_i])\Big) \geq p-1,$$

where $I_k := \{i \in \{1,\cdots,p\} \mid r(W_i) = k\}$, $k = 1, 2$.

In Fig. 6.1 a node partition inequality is depicted with three sets W_i with $r(W_i) = 2$ and two sets W_i with $r(W_i) = 1$. Edges with coefficient 0 are depicted by dashed lines; edges with coefficient 1 are depicted by solid lines.

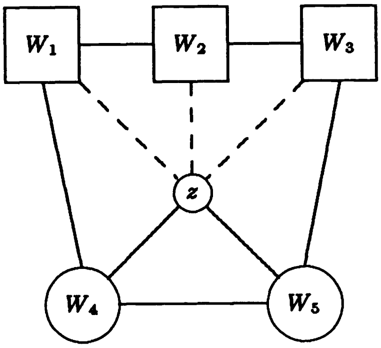


FIG. 6.1

THEOREM 6.2. *The node partition inequality* (6.1) *is valid for* $2\mathrm{NCON}(G;r)$.

Proof. Consider first a node partition inequality induced by a node z and the partition consisting of all node sets $\{v\}$, $v \in V\setminus\{z\}$. Suppose also that $r_v = 2$ for all $v \in V\setminus\{z\}$. This node partition inequality, $x(E(V\setminus\{z\})) \geq |V| - 2$, is valid, because after deletion of a node z the rest of the network should still connect all nodes $v \in V\setminus\{z\}$. Nodes of type 1 can be added successively to $V\setminus\{z\}$ by applying Lemma 4.6 with $Z := \{z\}$, $W := V\setminus\{z\}$, and $\alpha := 1$. With Lemma 4.2 all nodes $v \in V\setminus\{z\}$ can be expanded to node sets. In this way, every node partition inequality is proved to be valid. □

The following theorem gives a sufficient condition for the node partition inequality (6.1) to define a facet of $2\mathrm{NCON}(G;r)$.

THEOREM 6.3. *Consider a node partition inequality* (6.1) *induced by* $W_1, \cdots, W_p$. *Let* $\hat{G}$ *denote the graph* $(G-z)/W_1/\cdots/W_p$, *where the* W_i *are shrunk to nodes* w_i, $i = 1, \cdots, p$. *Let* I_1 *and* I_2 *be defined as in* (6.1). *The node partition inequality* $a^T x \geq p-1$ *defines a facet of* $2\mathrm{NCON}(G;r)$ *if*

(a) $\hat{G}$ *is two-node-connected;*

(b) $G[W_i \cup \{z\}] - e$ *is two-node-connected for all edges* $e \in G[W_i \cup \{z\}]$ *and for all* $i \in I_2$*;*

(c) $G[W_i]$ *is two-edge-connected for all* $i \in I_1$.

Proof. Let conditions (a), (b), and (c) be satisfied. We will show how to construct $|E|$ affinely independent vectors in the face defined by the node partition inequality (6.1).

Let E' be the set of all edges whose coefficients in $a^T x \geq p - 1$ are 0. By condition (a), the graph $\hat{G} = (\hat{V}, \hat{E})$ contains $|\hat{E}|$ spanning trees whose incidence vectors are affinely independent (see Theorem 4.10 in [GM]). Any such tree T of $\hat{G}$ can be augmented by E' to a feasible set $C \subseteq E$ for 2NCON$(G; r)$. Feasibility can be shown as follows. For any two nodes u, $v \in G[W_i \cup \{z\}]$ (where $i \in I_2$) there exist, by condition (b), two node-disjoint paths in (V, C). For $u \in W_i$ and $v \in W_j$ (where $i, j \in I_2$ and $i \neq j$), we construct the following two node-disjoint paths. In $(V, C) - z$, there exists a path from some node $u' \in W_i$ to some node $v' \in W_j$. Let u' and v' have the property that u' is the last node of W_i and v' is the first node of W_j encountered on this path. Since $G[W_i \cup \{z\}]$ is two-node–connected, it contains a $[u, u']$-path and a $[u, z]$-path, which do not have a node except u in common. (If $u = u'$, we only need one path, namely, the $[u, z]$-path.) Similarly, $G[W_j \cup \{z\}]$ contains a $[v, v']$-path and a $[v, z]$-path, which are node-disjoint. From these paths we can construct two node-disjoint $[u, v]$-paths in (V, C). So for all pairs u, v of nodes we can construct the required number of paths in (V, C), which proves feasibility of C. Feasibility is preserved even when some single $e \in E'$ is deleted from C. □

The connectivity conditions given in (b) imply that if $r(W_i) = 2$ for one of the node sets in the partition, then W_i must contain at least three nodes. This is not at all necessary. In fact, there exist facet-defining node-partition inequalities where all node sets in the partition contain exactly one node. Because we need it later on, we state this result as a lemma.

LEMMA 6.4. *Consider a* 2NCON *problem given by* (G, r) *and let* z *be some node of* G. *We suppose that* $G = (V, E)$ *is a graph with at least four nodes and* $r_v = 2$ *for all* $v \in V \backslash \{z\}$. *The node-partition inequality* (6.1) *induced by the partition of* $V \backslash \{z\}$ *into node sets* $\{w\}$ *for* $w \in V \backslash \{z\}$ *defines a facet of* 2NCON$(G; r)$ *if* z *is adjacent to every node in* G.

Proof. This can be proved by considering trees of $G - z$ augmented by certain edges of $\delta(z)$. Note that by (1.2)(iii) the graph G is supposed to be three-node–connected, so there exists a sufficient number of trees of $G - z$. □

Some **necessary** conditions for node-partition inequalities to define facets of 2NCON$(G; r)$ can be derived from Theorem 3.3 for node-cut inequalities.

THEOREM 6.5. *The node-partition inequality* (6.1) *defines a facet of* 2NCON $(G; r)$ *only if*

(a) $G[W_i]$ *is connected for all* $i \in I$*;*

(b) $\lambda_1(G[W_i \cup \{z\}]) \geq 2$ *for all* $i \in I_2$*;*

(c) $\lambda_1(G[W_i]) \geq 2$ *for all* $i \in I_1$*;*

(d) $\lambda_2(G[W_i]) \geq 2$ *for* $i = 1, \cdots, p$.

Proof. The proof is obvious. □

The connectivity conditions given in Theorem 6.5 can be easily checked and are of some practical use in cutting plane algorithms to derive faces of higher dimension.

7. Lifted two-cover inequalities. The motivation for introducing and studying the next class of inequalities derives from the fact that the two-matching inequalities play an important role in solving the traveling salesman problem; see [GP] and [PG].

The roots, however, are Edmonds's results for b-matching polyhedra (see [E]) since a certain (complemented) b-matching problem provides an interesting relaxation of the ECON problem.

Let $G = (V, E)$ be a graph and $r \in \{0, 1, 2\}^V$. Every incidence vector of a feasible solution $F \subseteq E$ to the 2ECON problem satisfies the "star inequalities" $x(\delta(v)) \geq r_v$ for all $v \in V$. And therefore the incidence vector of the complement $\bar{F} := E \backslash F$ of a feasible solution F to the 2ECON problem satisfies

$$(7.1) \qquad \begin{array}{ll} y(\delta(v)) \leq b_v := |\delta(v)| - r_v & \text{for all } v \in V, \\ 0 \leq y_e \leq 1 & \text{for all } e \in E. \end{array}$$

The convex hull of the integral solutions of (7.1) is the **1**-capacitated b-matching polytope of G, where $b = (b_v)_{v \in V} \in \mathbb{Z}^V$. Let us set, for $W \subseteq V$, $b(W) := \sum_{v \in W} b_v$. Edmonds [E] has shown that a complete linear description of the **1**-capacitated b-matching polytope of G is given by the following system

$$(7.2) \qquad \begin{array}{ll} y(\delta(v)) \leq b_v & \text{for all } v \in V, \\ y(E(H)) + y(\bar{T}) \leq \frac{b(H)+|\bar{T}|-1}{2} & \text{for all } H \subseteq V \text{ and all } \bar{T} \subseteq \delta(H) \text{ such} \\ & \text{that } b(H) + |\bar{T}| \text{ is odd,} \\ 0 \leq y_e \leq 1 & \text{for all } e \in E. \end{array}$$

Since $\chi^F = \mathbf{1} - \chi^{\bar{F}}$, we can derive from (7.2) that every incidence vector of a feasible solution to the 2ECON problem satisfies

$$(7.3) \qquad x(E(H)) + x(\delta(H) \backslash T) \geq \frac{\sum_{v \in H} r_v - |T| + 1}{2}$$

for all $H \subseteq V$ and all $T \subseteq \delta(H)$ such that $\sum_{v \in H} r_v - |T|$ is odd. In the transformation from (7.2) to (7.3) we have also set $T := \delta(H \backslash \bar{T})$.

Since $r \in \{0, 1, 2\}^V$, we call inequalities (7.3) **two-cover inequalities**. Note that it follows from Edmonds's result that the two-cover inequalities (7.3) plus the trivial constraints $0 \leq x_e \leq 1$, for all $e \in E$, give a complete description of the two-cover polytope, which is the convex hull of all incidence vectors of edge sets $F \subseteq E$ such that each node $v \in V$ has at least r incident edges.

From the two-cover inequalities we derive a larger class of inequalities as follows. Let $G = (V, E)$ be a graph and $r \in \{0, 1, 2\}^V$. Let $H \neq V$ be a node set, called the **handle**, and $T \subseteq \delta(H)$ an edge set. For each $e \in T$ we denote by T_e the set of the two endnodes of e. The sets T_e, $e \in T$, are called **teeth**. For simplicity we also call the edges $e \in T$ teeth in this section. If an edge $e \in T$ is parallel to some edge $f \in T$, we count T_e and T_f as two sets, even if $T_e = T_f$. Let $H_1, \cdots, H_p$, $p \geq 3$ be a partition of H into nonempty disjoint node sets such that

— $r(H_i) \geq 1$ for $i = 1, \cdots, p$;
— $r(H_i) = 2$ if H_i is intersected by some tooth, $i = 1, \cdots, p$;
— no more than two teeth may intersect any H_i, $i = 1, \cdots, p$;
— $|T| \geq 3$ and odd.

We call

$$(7.4) \qquad x(E(H)) - \sum_{i=1}^{p} x(E(H_i)) + x(\delta(H)) - x(T) \geq p - \left\lfloor \frac{|T|}{2} \right\rfloor$$

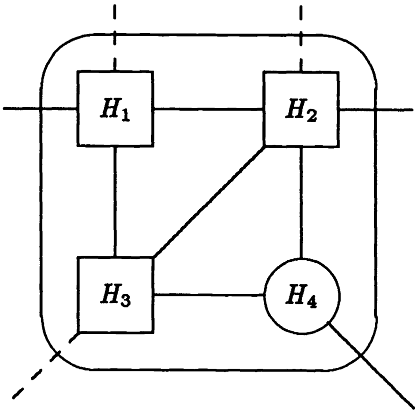


FIG. 7.1

the **lifted two-cover inequality.**

In Fig. 7.1 a handle with four node sets $H_1, \cdots, H_4$ and three teeth (drawn with dashed lines) is depicted, inducing a lifted two-cover inequality with right-hand side 3.

For the case in which $r_v = 2$ for all $v \in V$, Mahjoub [M] has found the same class of inequalities (and calls them "odd wheel inequalities" using a quite different notation).

Note that a lifted two-cover inequality coincides with a two-cover inequality (7.3), if $|H_i| = 1$ and $r(H_i) = 2$ for $i = 1, \cdots, p$. Note also that with each additional H_i with $|H_i| = 1$ and $r(H_i) = 1$ the right-hand side of a lifted two-cover inequality increases by 1, whereas the right-hand side of a two-cover inequality increases only by $\frac{1}{2}$ (on the average). This implies that two-cover inequalities do not support 2ECON$(G; r)$ if H contains nodes of type 1. Nevertheless, if the right-hand side of a two-cover inequality is increased appropriately, these inequalities define facets of 2ECON$(G; r)$ in many cases. This odd behavior may be explained by the fact that in an edge-minimal solution to the two-cover problem the nodes of type 1 may lie on matching edges, whereas in an edge-minimal solution to the 2ECON problem they are connected by a tree (or they lie on some cycle).

Also, the class of lifted two-cover inequalities is not very useful for the 2NCON problem, because they do not define facets in the case in which G is a complete graph and some H_i with incident tooth contains more than one node. In §8 we will introduce a class of inequalities for 2NCON$(G; r)$ that contain the lifted two-cover inequalities with $|H_i| = 1$ as a subclass, and define facets for complete G and $|H_i| \geq 1$. But these will be valid only for 2NCON$(G; r)$.

As in the previous sections, we will derive validity and facet results of lifted two-cover inequalities from validity and facet results of a special class of lifted two-cover inequalities, namely those with $|H_i| = 1$.

THEOREM 7.5. *A lifted two-cover inequality* (7.4) *is valid for* 2ECON$(G; r)$ (*and hence for* 2NCON$(G; r)$).

Proof. First, assume that $|H_i| = 1$ and that all nodes in the handle are of type 2. In this case, we have a two-cover inequality that is valid for the polytope of two-covers, hence for 2ECON$(G; r)$. It is also easy to prove validity in this case by summing up the inequalities:

$$\begin{aligned} x(\delta(v)) &\geq \ \ 2 \quad \text{for all } v \in H,\\ -x_e &\geq -1 \quad \text{for all } e \in T,\\ x_e &\geq \ \ 0 \quad \text{for all } e \in \delta(H)\backslash T, \end{aligned}$$

dividing the result by 2 and rounding the right-hand side up.

Our next step is induction over the number of nodes of type 1 in the handle (but still $|H_i| = 1$). This can be done with the help of Lemma 4.6 by setting $W := H$, $\alpha := 1$, and w as the new node of type 1. The result is a new valid inequality of the form (7.4).

Finally, using Lemma 4.2, we expand the nodes in the handle successively to node sets H_i with coefficients 0 inside H_i, to derive all inequalities of the form (7.4).

Note that when lifting a node w with incident $wv \in T$ to node set W, only one edge of $[W : \{v\}]$ gets coefficient 0; all others have coefficient 1 in the lifted two-cover inequality. (If all edges in $[W : \{v\}]$ had coefficient 0, the obtained inequality would not be valid for 2ECON$(G;r)$, but it would be valid for 2NCON$(G;r)$; see Theorem 8.2.) □

Lifted two-cover inequalities are also valid if we allow an even number of teeth. But they cannot define facets in this case, as can be seen easily.

The following theorem gives a necessary and sufficient condition for a special subclass of lifted two-cover inequalities to define facets of 2ECON$(G;r)$.

THEOREM 7.6. (a) *A lifted two-cover inequality* (7.4) *with* $|H_i| = 1$ *for* $i = 1,\cdots,p$, $|H| = |T|(=p)$, *and* $|V\backslash H| = 1$, *defines a facet of* 2ECON$(G;r)$ *if and only if* $G[H]$ *is hypomatchable* (*i.e., for each node* $v \in H$ *there is a matching of* $G[H]$ *that is incident to all nodes in* H *except* v).

(b) *Let* $G[H]$ *be a complete graph. Then any lifted two-cover inequality* (7.4) *with* $|H_i| = 1$ *for* $i = 1,\cdots,p$, $|H| \geq |T| \geq 3$, *and* $|V\backslash H| = 1$, *defines a facet of* 2ECON$(G;r)$.

Proof. Let F be the face induced by the lifted two-cover inequality in question.

(a) Let F be contained in a facet F_b induced by some inequality $b^T x \geq \beta$. We want to prove that b is a scalar multiple of the left-hand side of the lifted two-cover inequality.

Pick some $v \in H$. Any perfect matching M of $G[H\backslash\{v\}]$ can be enlarged to a set C whose incidence vector is in the face F by adding some edge $e \in \delta(v)\backslash T$ along with all tooth edges. The resulting set $C \cup \{e\} \cup T$ is two-edge–connected and $C \cup \{e\}$ contains $|H| - \lfloor\frac{|T|}{2}\rfloor = \lceil\frac{|T|}{2}\rceil$ edges. By varying $e \in \delta(v)\backslash T$, we achieve $b_e = \alpha_v$ for all $e \in \delta(v)\backslash T$ and some constant α_v. Since $G[H]$ is connected, α_v is the same for all nodes $v \in H$. ($G[H]$ is connected if $G[H]$ is hypomatchable.)

Now we prove that $b_e = 0$ for $e \in T$. Let u be the node in H incident to e and let v be some node in H adjacent to u. The incidence vector of a perfect matching of $G[H\backslash\{v\}]$ plus edge uv plus $T\backslash\{e\}$ lies in F_b. Since adding edge e does not change the right-hand side, we know $b_e = 0$. Therefore, our lifted two-cover inequality defines a facet.

Suppose now that $E(H)$ is not hypomatchable. With the help of Tutte's theorem we will find a separation of $E(H) \cup (\delta(H)\backslash T)$ into edge sets $E_1, E_2, \cdots, E_s$ so that $x(E_i) \geq k_i$ is valid for 2ECON$(G;r)$ and the sum of the k_i is at least $|H| - \lfloor\frac{|T|}{2}\rfloor$. This is done as follows: since for some node $v \in H$ the graph $G[H\backslash\{v\}]$ has no perfect matching, by Tutte's theorem there exists a node set $S \subseteq H\backslash\{v\}$ so that the number of odd components $c_o(G[H\backslash\{v\}]-S)$ of $G[H\backslash\{v\}]-S$ is strictly larger than $|S|$. Since $H\backslash\{v\}$ is an even node set ($|H| = |T|$ is odd), either the number of odd components of $G[H\backslash\{v\}] - S$ is odd and $|S|$ is odd, or both numbers are even. In any case, we know that $c_o(G[H\backslash\{v\}] - S) - |S| \geq 2$. So $c_o(G[H] - (S \cup \{v\}))$, which is the same as $c_o(G[H\backslash\{v\}] - S)$, is still larger than $|S \cup \{v\}|$. For the sake of simplicity, we will rename $S := S \cup \{v\}$. Let H_i be the node set of the ith (odd or even) component of $G - S$. Let T_i denote the subset of teeth incident to H_i and let E_i denote the edge set $E(H_i) \cup (\delta(H_i)\backslash T_i)$. The T_i constitute a partition of $T\backslash\delta(S)$, and the E_i constitute a partition of the edge set $(E(H) - E(S)) \cup (\delta(H)\backslash T)$.

$$x(E_i) \geq k_i := |H_i| - \left\lfloor \frac{|T_i|}{2} \right\rfloor$$

is a valid lifted two-cover inequality (this is valid also for an even number of teeth!). If we take the sum of these inequalities plus the nonnegativity constraints for $e \in E(S)$, we achieve $x(E(H)) + x(\delta(H)\backslash T) \geq k$, where k is the sum of the k_i. In the right-hand side, the $|H_i|$ sum up to $|H| - |S|$, and the $\lfloor |T_i|/2 \rfloor$ sum up to $\frac{1}{2}(|T| - |S| - c_o(G[H] - S))$, so the k_i sum up to

$$|H| - \frac{|T|}{2} + \frac{1}{2}(c_o(G[H] - S) - |S|) \geq |H| - \left\lfloor \frac{|T|}{2} \right\rfloor.$$

Therefore, our lifted two-cover inequality can be written as the sum of at least two other valid inequalities; hence it does not define a facet.

(b) Assume first that H contains only nodes of type 2 (with or without incident teeth). If nodes of type 2 without incident teeth are allowed in the handle, the restriction of a feasible set C whose incidence vector is in F to the edge set $E(H) \cup \delta(H)\backslash T$ is something more complicated than a matching with additional edge. It is rather a collection of node-disjoint paths between pairs of nodes with incident teeth plus one additional path connecting the last node with incident tooth to $V\backslash H$ or to some other path. More exactly, if we set $\bar{r}_v := 2$ minus the number of incident teeth for $v \in H$ and $\bar{r}_z := 0$ for the node $z \notin H$, then $C\backslash T$ meets each node $v \in V$ with exactly $\bar{r}_v$ edges, except for one node that is met by $\bar{r}_v + 1$ edges. $C\backslash T$ is a near-perfect $\bar{r}$-cover of $E(H) \cup (\delta(H)\backslash T)$ (so to speak). To see this, add the $\bar{r}_v$, divide by two, and compare this with the right-hand side of the lifted two-cover inequality. (But not every near-perfect $\bar{r}$-cover of $E\backslash T$ plus T defines a feasible set, as there might be some node-disjoint cycles.)

Since the structure of the feasible sets with incidence vector in F is somewhat unwieldy, we switch to complete graphs. Let F_b be a facet containing F, induced by some valid inequality $b^T x \geq \beta$. First we show $b_e = \alpha_v$ for all $e \in \delta(v)\backslash T$ and all $v \in H$. The connectedness of $G[H]$ will imply that the α_v are the same for all $v \in V$. If $v \in H$ has an incident tooth, construct node-disjoint paths in $G[H]$ connecting pairs of nodes with incident teeth and meeting all nodes of H except v. To this set add any edge $e \in \delta(v)\backslash T$ and T. Since we have freedom in choosing e, we can prove $b_e = \alpha_v$ for all nodes $v \in H$ with incident teeth. If $v \in H$ has no incident tooth, construct node-disjoint paths in $E(H)$ between pairs of nodes with incident teeth plus one path (node-disjoint from all others) between v and the last leftover node with an incident tooth. These paths should meet all nodes in H. Call this collection of paths C. As before, we can add any edge of $\delta(v)$ (except the path edge $C \cap \delta(v)$), plus all teeth, and get a set with incidence vector in F_a. This proves $b_e = \alpha_v$ for all $e \in \delta(v)\backslash C$. But we can construct another set C' the same way as before, only this time it uses a different edge of $\delta(v)$. So we have $b_e = \gamma_v$ for all $e \in \delta(v)\backslash C'$ and some value γ_v. Since $\delta(v)\backslash T$ contains at least three edges, all edges in $\delta(v)\backslash T$ have the same a_e-value

$\gamma_v = \alpha_v$. Proving $b_e = 0$ for the teeth $e \in T$ is easy, so we have that b is identical to the lifted two-cover inequality; therefore it defines a facet.

If H contains nodes of type 1, we use Theorem 4.7 for induction on the number of nodes of type 1 in H in the same way as we used Lemma 4.6 for proving validity of the lifted two-cover inequality. □

Usually the feasible sets of 2ECON$(G; r)$ whose incidence vectors satisfy the lifted two-cover inequality with equality are not feasible for 2NCON$(G; r)$ if $V \backslash H$ consists of only one node, because this node may be an articulation node. But if $V \backslash H$ has sufficiently high connectivity, (7.4) may define a facet of 2NCON$(G; r)$.

Remark 7.7. A lifted two-cover inequality (7.4) with $|H_i| = 1$ for $i = 1, \cdots, p$, $|H| = |T| (= p)$, defines a facet of 2NCON$(G; r)$ if $G[H]$ is hypomatchable, $G[V \backslash H]$ is three-edge–connected, no two teeth are incident to the same node (in $V \backslash H$), and no parallel edges exist.

Proof. The proof is analogous to the proof of Theorem 7.6. □

But usually, as the following remark shows, lifted two-cover inequalities do not define facets for 2NCON$(G; r)$ as soon as $|H_i| \geq 2$ for some H_i with an incident tooth.

Remark 7.8. A lifted two-cover inequality does not define a facet of 2NCON$(G; r)$ if there is a node set H_i and a node $v \in V \backslash H$ so that $[\{v\} : H_i]$ contains a tooth and a nontooth.

(This is the case especially if G is complete and some H_i with incident tooth contains at least two nodes.)

Proof. It can be shown that a feasible set $C \subseteq E$ with 2NCON$(G; r)$ that satisfies such a lifted two-cover inequality with equality never uses the nontooth in $[\{v\} : H_i]$. □

But for the 2ECON problem we can use our lifting lemmas of §4 to derive sufficient conditions for a lifted two-cover inequality with general H_i to define a facet of 2ECON$(G; r)$.

THEOREM 7.9. *Given a lifted two-cover inequality* (7.4), *we will denote by $\hat{G}$ the graph $G/H_1/\cdots/H_p$.*

(a) *If $\hat{G}[H]$ is hypomatchable* (*in the case $p = |T|$*) *or complete* (*in the case $p > |T|$*), *if the $G[H_i]$ for $i = 1, \cdots, p$ are $(r(H_i) + 1)$-edge–connected, and if $G[V \backslash H]$ is* $\max\{2, r(V \backslash H) + 1\}$*-edge–connected, a lifted two-cover inequality defines a facet of* 2ECON$(G; r)$.

(b) *If the lifted two-cover inequality is facet-inducing, then $\hat{G}[H]$ and $G[H_i]$ are connected for $i = 1, \cdots, p$, and $\lambda_1(G[H_i]) \geq 1$ for $i = 1, \cdots, p$. In fact, one can always find $H_1, \cdots, H_p$ with $\lambda_1(G[H_i]) \geq 2$ for $i = 1, \cdots, p$ that induce the lifted two-cover inequality in question.*

Proof. (a) Theorem 7.6 proves the lifted two-cover inequality to be facet-defining for 2ECON$(\hat{G}; r)$. With Lemma 4.4 we can lift this result to 2ECON$(G; r)$.

(b) It is easy to see that the $G[H_i]$ must be connected for all $i = 1, \cdots, p$.

If $\hat{G}[H]$ is not connected, we can split the handle H into two handles H' and H'' to derive two lifted two-cover inequalities whose sum gives the old one. So the old one cannot define a facet.

It remains to show that we can find $H_1, \cdots, H_p$ with $\lambda_1(G[H_i]) \geq 2$ for $i = 1, \cdots, p$ that induce our lifted two-cover inequality.

If H_i has no incident tooth and $\lambda_1(G[H_i]) = 1$, then our lifted two-cover inequality can be written as the sum of another lifted two-cover inequality where H_i is split into at least two other sets plus one constraint $x_e \leq 1$. The same argument is possible if H_i has an incident tooth and $\lambda_2(G[H_i]) = 1$. So in these cases our lifted two-cover inequality cannot define a facet.

It remains to check the case in which H_i has an incident tooth e and $\lambda_1(G[H_i]) = 1$. In this case $G[H_i]$ has a bridge f so that $G[H_i]-f$ decomposes into two components U and W with $r(U) = 1$ and $r(W) \geq 1$. The interesting case is the one where the tooth e is incident to U, because there we cannot simply split H_i into U and W to derive a stronger lifted two-cover inequality. But we can replace H_i by $H_i \backslash U$, H by $H \backslash U$, and the tooth e by the bridge f to derive another lifted two-cover inequality of the same form as the old one. By repeating this procedure of reducing H_i, we can assume that $\lambda_1(G[H_i]) \geq 2$ for all $i = 1, \cdots, p$. □

8. Comb inequalities. The following constraints were motivated, on the one hand, by the comb inequalities for the traveling salesman problem (see [GP]), and on the other hand, they were motivated by the fact that the lifted two-cover inequalities do not generally define facets for the 2NCON problem (see Remark 7.8). We wanted to find a facet containing the face induced by a lifted two-cover inequality in the case in which G is a complete graph and the H_i contain more than one node.

The class of inequalities we came up with in this case are valid for 2NCON$(G; r)$, but not generally for 2ECON$(G; r)$. We will call this class comb inequalities for 2NCON$(G; r)$. These inequalities allow a further generalization using the concept of clique trees. But we will not discuss this here.

Let $H, T_1, \cdots, T_t$ be subsets of V and let $z_i \in T_i \backslash H$, $i = 1, \cdots, t$, be not necessarily distinct nodes (H is called the **handle**, the sets $T_1, \cdots, T_t$ are the **teeth**, and the $z_1, \cdots, z_t$ the **special nodes**) that satisfy the following conditions:

— $t \geq 3$ and odd;

— two teeth have at most one node in common;

— if $T_i \cap T_j \neq \emptyset$, then $T_i \cap T_j = \{z_i\} = \{z_j\}$;

— each tooth T_i intersects the handle H in exactly one node; we denote this node by t_i for $i = 1, \cdots, t$;

— $r_{t_i} = 2$ for $i = 1, \cdots, t$;

— $r_v \geq 1$ for all $v \in H \cup (\cup_{i=1}^t (T_i \backslash \{z_i\}))$.

We denote by V_2 the set of nodes of type 2 in G. The **special comb inequality** is given by

$$
\begin{aligned}
(8.1) \qquad & x(E(H)) + x(\delta(H)) + \textstyle\sum_{i=1}^t x(E(T_i)) \\
& \quad + \textstyle\sum_{i=1}^t x([T_i \backslash (H \cup \{z_i\}) : V \backslash T_i]) - \sum_{i=1}^t x([\{t_i\} : T_i]) \\
& \quad - \textstyle\sum_{i=1}^t x([\{z_i\} : T_i \cap V_2]) \geq |H| + \sum_{i=1}^t (|T_i| - 2) - \lfloor \frac{t}{2} \rfloor .
\end{aligned}
$$

The (general) **comb inequality** is derived from the special comb inequality (8.1) by expanding all nodes $w \in H$ that are not in $\{z_1, \cdots, z_t\}$ to node sets W (see Lemma 4.2). Figure 8.1 gives an illustration of a comb inequality with a handle H consisting of four node sets and three teeth T_i, $i = 1, \cdots, 3$, which has right-hand side 6. Edges with coefficient 0 are drawn with dashed lines, edges with coefficient 1 with solid lines, and edges with coefficient 2 with bold lines.

We note that the comb inequality becomes a lifted two-cover inequality with sets $H_i := \{t_i\}$ if $|T_i| = 2$ and $|E(T_i)| = 1$.

We will prove validity and facet results only for special comb inequalities. With the help of Lemmas 4.2–4.5 one can easily derive validity and facet results for general comb inequalities.

THEOREM 8.2. *A comb inequality* (8.1) *is valid for* 2CON$(G; Z; r)$ *with* $Z = \{z_1, z_2, \cdots, z_t\}$, *and hence it is valid for* 2NCON$(G; r)$.

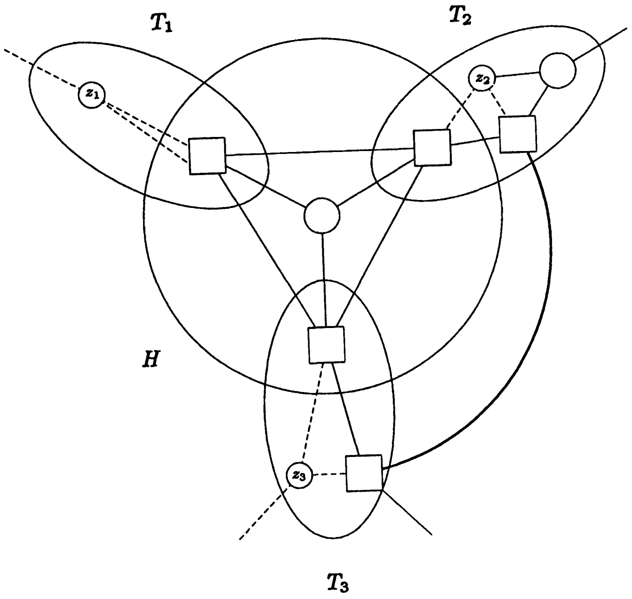


FIG. 8.1

Proof. Assume that all nodes in $\big(H \cup (\cup_{i=1}^t T_i)\big) \backslash \{z_1, \cdots, z_t\}$ are nodes of type 2. Then the left-hand side of the comb inequality (8.1) can be written as $\frac{1}{2}$ times the sum of the following inequalities with subsequent rounding:

(1) for all $v \in H \backslash (\cup_{i=1}^t T_i)$: the cut inequality $x(\delta(v)) \geq 2$;

(2) for all teeth T_i: the node-partition inequality (6.1) induced by z_i and the partition $\{V\backslash T_i, \ \{v\}$ for all $v \in T_i \backslash \{z_i\}\}$; the right-hand side is $|T_i| - 1$;

(3) for all teeth T_i with $r\big(T_i \backslash \{t_i, z_i\}\big) = 2$: the node-partition inequality (6.1) induced by z_i and the partition $\{V \backslash (T_i \backslash \{t_i\}), \ \{v\}$ for all $v \in T_i \backslash \{z_i, t_i\}\}$; the right-hand side is $|T_i| - 2$;

(4) for all teeth T_i with $r\big(T_i \backslash \{t_i, z_i\}\big) = 1$: the partition inequality (5.2) induced by the partition $\{(V \backslash T_i) \cup \{t_i, z_i\}, \ \{v\}$ for $v \in T_i \backslash \{t_i, z_i\}\}$; its right-hand side is $|T_i| - 2$;

(5) some nonnegativity constraints.

The sum of $(\frac{1}{2})$ times the right-hand sides of these inequalities is:

$$\begin{aligned} &|H\backslash \ (\textstyle\bigcup_{i=1}^t T_i)| + \sum_{i=1}^t (|T_i| - \frac{3}{2}) \\ &= |H| - t + \textstyle\sum_{i=1}^t |T_i| - \frac{3t}{2} \\ &= |H| + \textstyle\sum_{i=1}^t (|T_i| - 2) - \frac{t}{2}. \end{aligned}$$

Rounding this up gives the right-hand side of (8.1) exactly.

If the handle contains nodes of type 1, we apply Lemma 4.6 inductively with $W := H$ and $\alpha := 1$. If a tooth T_i contains nodes of type 1, we apply Lemma 4.6

with $W := T_i$ and $\alpha := 1$; this is done in the same way as in the validity proof for node-partition inequalities. □

Note that the comb inequality (8.1) is also valid if the number of teeth t is even. But in this case it does not define a facet, as it can be written as the sum of a comb inequality and node-partition inequality (or a nonnegativity constraint).

Note also that if $H \cup (\cup_{i=1}^{t} T_i) = V$ and $z_1 = z_2 = \cdots = z_t$ and $|T_i| = 2$ for all i, the special comb inequality with right-hand side $|H| - \lfloor \frac{t}{2} \rfloor$ may degenerate into a node-partition inequality with higher right-hand side, namely, $|H| - 1$. In this case the special comb inequality cannot define a facet.

THEOREM 8.3. *The special comb inequality* (8.1) *defines a facet of* 2NCON$(G; r)$ *if* $r_v = 2$ *for all nodes* $v \in V$, *if the* z_i *are all distinct, and if* G *is the complete graph minus all edges with coefficient* 2 *in* (8.1).

Proof. The restriction to nodes of type 2 has only technical reasons, mainly because of Lemma 6.4. The restriction to edges with coefficients 0 and 1 is also introduced only for technical reasons. Once we have proved an inequality to define a facet only on a subset of edges of the complete graph, it is easy to prove it to be facet-defining on the complete graph.

Let F be the face induced by the comb inequality in question, and let F be contained in the face F_b induced by some valid inequality $b^T x \geq \beta$. First we prove $b_e = \alpha_i$ for all edges in $E(T_i) \cup [\{t_i\} : H]$ with coefficient 1 and some α_i. We do this (without loss of generality) for tooth T_1. Suppose that $|T_1| \geq 3$. (For "small" teeth that consist of only one edge, the following proof has to be modified somewhat.) Construct a collection P of node-disjoint paths in $G[H]$ between pairs of nodes t_j, say, between t_2 and t_3, t_4 and t_5, etc. Those paths should meet every node in H except t_1. To this collection of paths P, we may add certain trees in the teeth T_i that are constructed as follows:

(1) For T_1 we take any feasible edge set whose incidence vector lies in the face of 2NCON$(G/(V \backslash T_1); r)$ induced by a certain node-partition inequality on T_1, namely, the one with node $z = z_1$ and node sets $\{v\}$ for all nodes v in T_1 and $\{w\}$ for the shrunk node standing for $V \backslash T_1$ (cf. (2) used in the validity proof in Theorem 8.2). These sets are trees on $T_1 \backslash \{z_1\}$ plus certain edges of $\delta(z_1)$ plus some edge leading from T_1 to $V \backslash T_1$. Note also that the face of 2NCON$(G/(V \backslash T_1); r)$ induced by the node-partition inequality is a facet by Lemma 6.4.

(2) For T_i with $i \neq 1$, we take any feasible edge set whose incidence vector lies in the face of 2NCON$\big(G/((V \backslash T_i) \cup \{t_i\}); r\big)$ induced by (3) or (4) of the validity proof in Theorem 8.2. These objects are mainly trees on $T_i \backslash \{z_i, t_i\}$ plus certain edges in $[\{z_i\} : T_i]$. If $|T_i| = 2$, we just take the edge of tooth T_i.

Finally, we add all edges $z_i z_j$ to this construction.

We claim that this combination of paths in $G[H]$ and trees of T_i is feasible. This can be easily checked. Secondly, we claim that its incidence vector lies in the face induced by the comb inequality; this is true because all inequalities used in the validity proof of the comb inequality are satisfied with equality except one.

Since we have some freedom in the choice of the "tree" in T_1, and we know that the node-partition inequality used for the construction of these "trees" defines a facet of 2NCON$(G/(V \backslash T_i); r)$, we know that $b_e = \alpha_1$ for all nonzero edges in this node-partition inequality, and $b_e = 0$ for all zero edges e. This can be done for all teeth T_i in the same way as shown for tooth T_1.

Now we prove that all edges inside the handle have the same b_e-value. This value must be the same as α_1, α_2, etc. Thus, we know that all edges with coefficient 1 in the comb inequality have the same b_e-value and all edges e with coefficient 0 in the comb inequality have $b_e = 0$.

To prove $b_e = \alpha_v$ for all $e \in \delta_{G[H]}(v)$ and $v \in H$, we just vary our construction of paths in the beginning. This is done in exactly the same way as in the proof of Theorem 7.6(b). To give an example: If $v \in H \backslash T$, then we construct paths between t_1 and v, t_2 and t_3, etc. that are all node-disjoint. These paths should meet all nodes in $G[H]$. In addition to this collection P of paths we construct trees in T_i according to point (2) above. Now we can add any edge $e \in \delta(v) \cap E(H)$ not already in some path to achieve a feasible solution whose incidence vector lies in the face F_a. So $b_e = b_f$ for all $e, f \in (\delta(v) \backslash P) \cap E(H)$. To prove $b_e = b_f$ for all $e, f \in \delta(v) \cap E(H)$, we just choose a collection of paths using another edge of $\delta(v)$.

It is easy to prove that the b_e-value for the e of zero coefficient in the comb inequality is also 0.

So inequality $b^T x \geq \beta$ is identical to the comb inequality (8.1) except for scalar multiplication. Therefore, it defines a facet of $2\text{NCON}(G; r)$. □

The question naturally arises whether there are also "comb" inequalities valid for $2\text{ECON}(G; r)$. We know of such a class, but the validity proof is somewhat ugly. In such a "comb" inequality we have two types of teeth: "simple" teeth consisting of only one edge with coefficient 0, and "large" teeth T with coefficients 0 on edges in $T \backslash H$, and coefficients 1 on the edges leading from $T \backslash H$ to $T \cap H$ and to the "outside." The edges in the handle have coefficients 2. This seems to be more symmetric, and therefore, in a way, nicer than the comb inequalities (8.1).

Also, some other odds and ends of inequalities that do not fit into any of the presented classes are known to us. Some of these are published in Stoer's dissertation [S].

9. Computational results. The theory presented here for the 2ECON and 2NCON polytopes was developed in order to solve problems of the type and size that arise in the design of survivable telephone networks in fiber optic technology. The idea was to design and implement a cutting plane algorithm that uses the inequalities introduced above.

As mentioned before, it unfortunately turned out that—except for the cut and node-cut inequalities—the separation problem for all other classes of inequalities presented here is NP-hard. This means that we can use these classes of inequalities only heuristically. We had to make an experimental investigation of the relative benefit of running various heuristics that determine, for a given point y, an inequality of some class of valid inequalities that is violated by y.

The final outcome of our computational study was a cutting plane code that uses exact separation routines for cut and node-cut inequalities and separation heuristics for partition, node-partition, and lifted two-cover inequalities. For the type and size of practical problems used as our test cases, the other classes of inequalities were of no significant help. We expect, however, that for larger problem sizes and graphs of higher density further inequalities will be needed to achieve satisfactory computational performance. But that will make a more thorough design and investigation of separation heuristics for the other classes of inequalities necessary.

The design and implementation of a practically efficient cutting plane algorithm is a rather tricky and time-consuming task. Its success is based on the proper combination of many details. Some of these are described in [GMS] and [S]. We are unable to outline these here. Our final code showed the following computational characteristics on our test problems.

We obtained the data of seven real networks (nodes, possible direct links, costs of establishing links) from network designers at Bell Communications Research. The sizes ranged from 36 nodes and 65 edges to 116 nodes and 173 edges. For all networks, 2NCON and 2ECON solutions had to be found, but in only one case did these solutions differ. So we had eight test problems available. According to the network designers, these data represent the range of typical practical applications in this area.

We ran our cutting plane algorithm (using a research version of Bixby's LP-code (see [Bix]) and Jünger's Branch and Cut framework (unpublished)) on a SUN 3/60, a 3 MIPS machine. Five of the eight problems were solved to optimality in the cutting plane phase in less than 10 seconds. In the remaining three cases the cutting plane phase finished after at most 31 seconds with an integrality gap of less than 1 percent. In the subsequent branch and cut phases no more than 20 nodes were generated in the branching tree and at most an additional $1\frac{1}{2}$ minutes were needed to find an optimal solution and prove optimality. Further cases, run subsequently, showed similar computational performance. (See [GMS] for more details.)

Considering these computational results, we feel confident in saying that all survivable network design problems of the type and size arising at Bellcore can be solved to optimality with our code in at most a few minutes on a 3 MIPS machine. Thus the theoretical investigation presented here has helped (and helps further) to solve typical instances of a combinatorial optimization problem of significant practical importance.

REFERENCES

[Bix] R. E. Bixby, *Implementing the simplex method: The initial basis,* Tech. Report TR 90-32, Department of Mathematical Sciences, Rice University, Houston, TX, 1991.

[CFLM] R. H. Cardwell, H. Fowler, H. L. Lemberg, and C. L. Monma, *Determining the impact of fiber optic technology on telephone network design,* Bellcore Exchange Magazine, March/April 1988, pp. 27–32.

[CFN] G. Cornuéjols, J. Fonlupt, and D. Naddef, *The traveling salesman problem in a graph and some related integer programs,* Math. Programming, 33 (1988), pp. 1–27.

[CMW] R. H. Cardwell, C. L. Monma, and T. H. Wu, *Computer-aided design procedures for survivable fiber optic networks,* IEEE Selected Areas of Communications, 7 (1989), pp. 1188–1197.

[Chv] V. Chvátal, *Edmonds polytopes and a hierarchy of combinatorial problems,* Discrete Math., 4 (1973), pp. 305–337.

[E] J. Edmonds, *Maximum matching and a polyhedron with* 0, 1*-vertices,* J. Res. Nat. Bur. Standards, B69 (1965), pp. 125–130.

[GM] M. Grötschel and C. L. Monma, *Integer polyhedra associated with certain network design problems with connectivity constraints,* SIAM J. Discrete Math., 3 (1990), pp. 502–523.

[GMS] M. Grötschel, C. L. Monma, and M. Stoer, *Computational results with a cutting plane algorithm for designing communication networks with low-connectivity constraints,* Oper. Res., to appear.

[GP] M. Grötschel and M. W. Padberg, *Polyhedral theory,* in The Traveling Salesman Problem, E. L. Lawler, J. K. Lenstra, A. H. G. Rinnooy Kan, and D. Shmoys, eds., John Wiley, Chichester, U.K., 1985, pp. 251–305.

[M] A. R. Mahjoub, *Two edge connected spanning subgraphs and polyhedra,* Report No 88520-OR, Institut für Operations Research, Universität Bonn, Bonn, Germany, 1988.

[MMP] C. L. Monma, B. S. Munson, and W. R. Pulleyblank, *Minimum-weight two-connected spanning networks,* Math. Programming, 46 (1990), pp. 153–171.

[MS] C. L. Monma and D. F. Shallcross, *Methods for designing communication networks with certain two-connected survivability constraints,* Oper. Res., 37 (1989), pp. 531–541.

[PG] M. W. PADBERG AND M. GRÖTSCHEL, *Polyhedral computations,* in The Traveling Salesman Problem, E. L. Lawler, J. K. Lenstra, A. H. G. Rinnooy Kan, and D. Shmoys, eds., John Wiley, Chichester, U.K., 1985, pp. 307–360.

[S] M. STOER, *Design of survivable networks,* Ph.D. thesis, University of Augsburg, Augsburg, Germany, 1991.

[Sch] A. SCHRIJVER, *Theory of Linear and Integer Programming,* John Wiley, New York, 1986.

SIAM J. OPTIMIZATION
Vol. 2, No. 3, pp. 505–532, August 1992

009

CONVERGENCE OF BROYDEN'S METHOD IN BANACH SPACES*

D. M. HWANG† AND C. T. KELLEY†

Abstract. This paper proves new convergence theorems for convergence of Broyden's method when applied to nonlinear equations in Banach spaces. The convergence is in the norm of the Banach space itself, rather than in the norm of some Hilbert space that contains the Banach space. It is shown that the norms in which q-superlinear convergence takes place are determined by the smoothing properties of the error in the Fréchet derivative approximation and not by the inner product in which Broyden's method is implemented. Among the consequences of the results in this paper are a proof of sup-norm local q-superlinear convergence when Broyden's method is applied to integral equations with continuous kernels, global q-superlinear convergence of the Broyden iterates for singular and nonsingular linear compact fixed point problems in Banach space, a new method for integral equations having derivatives with sparse kernels, and q-superlinear convergence for a new method for integral equations when part of the Fréchet derivative can be explicitly computed. Partitioned variants of the methods and the "bad" Broyden method are also discussed.

Key words. Broyden's method, q-superlinear convergence, quasi-Newton update, Banach space

AMS(MOS) subject classifications. 65J15, 47H17, 49D15

1. Introduction.

This paper considers the solution of equations in Banach space by Broyden's method. We write our equations as

$$F(u) = 0, \tag{1.1}$$

where F is a Lipschitz continuously differentiable map between Banach spaces X and Y. We will consider both linear and nonlinear equations. Broyden's method is a variation of Newton's method in which an approximation, B, to the Fréchet derivative, $F'(u^*)$, is maintained along with an approximation u to a solution u^*. This method has been used with success for discretizations of infinite-dimensional problems in integral equations [23], [20]; fluid mechanics [10]; and optimal control [26]. In this paper we take the position that analysis of the convergence properties of the method should not be done for the discrete finite-dimensional problems alone because the results of such a finite-dimensional analysis can hide features that may depend on how the level of discretization is refined and may lead to conclusions that are valid only in finite dimension. The papers [11], [31], and [24] present examples of these kinds of problems. Direct analysis of the infinite-dimensional problem can also lead to effective preconditioning strategies that produce good convergence properties even for finite-dimensional approximations.

The purpose of this paper is to extend convergence results for Broyden's method in finite-dimensional spaces and Hilbert spaces to the Banach space setting and thereby sharpen convergence results. In particular, our results show how the inner product used in the implementation of Broyden's method and the topology in which convergence takes place are related. The significant consequences of this Banach space analysis for problems considered previously in [23], [20], [26], and [27] are to make the convergence estimates more precise, for example, to provide uniform convergence

* Received by the editors December 7, 1990; accepted for publication (in revised form) July 3, 1991. This research was supported by Air Force Office of Scientific Research grant #AFOSR-ISSA-890044 and National Science Foundation grant #DMS-8900410. Computing activity was also partially supported by an allocation of time from the North Carolina Supercomputing Center.

† North Carolina State University, Department of Mathematics, Box 8205, Raleigh, North Carolina 27695-8205.

results where at best L^2 results had been available. Our technique of analysis is an extension of that in [27] and allows for a direct discussion of linear equations and least squares problems that extends results in [4], [12], [28], and [13]. Our analysis also extends to partitioned variations of Broyden's method and to the so-called "bad" Broyden method.

In this introductory section we describe our setting and establish notation. In §2 we state and prove a basic lemma and use it to derive global and local convergence results for singular and nonsingular linear problems. The lemma is an extension and reformulation of a weak q-superlinear convergence result from [18]. The convergence results in §2 extend those of [4], [12], [13], and [28] for the finite-dimensional case. In the Hilbert space setting, unpublished results [15] give q-superlinear convergence for nonsingular linear systems with a more complex proof that uses the singular value decomposition in a way similar to the result for nonlinear problems given in [16]. Our proof is more direct and uses extensions of the techniques of [28] and [27]. Our proof also extends to allow for a partitioned structure in the sense of [17]. We use this idea to introduce a new q-superlinearly convergent algorithm for linear integral equations whose kernels have a sparse structure. We show how the ideas extend to the "bad" Broyden update.

In §3 we extend local q-superlinear convergence results for nonlinear problems from [3], [16], [18], and [27] for the finite-dimensional and Hilbert space settings to the Banach space case. It may happen that different parts of the nonlinear function have different continuity properties. Our nonlinear method can handle this situation and we apply it to a class of integral equations for which Broyden's method is not directly applicable. The partitioned and "bad" Broyden variants of the algorithm are discussed as well.

In §4 we report on several numerical experiments. The goal of these experiments is to illustrate how Broyden's method produces iterates that converge q-superlinearly in a variety of Banach space norms. We report results for more than one grid size to show that the tables reflect infinite-dimensional properties of the iteration.

We assume that the Banach space X has a continuous inner product $(\cdot,\cdot)$. We let H denote the completion of X in the norm induced by the inner product and denote this norm by $\|\cdot\|_H$. We will multiply the inner product by a constant, if necessary, to ensure that

$$\|u\|_H \le \|u\|_X \tag{1.2}$$

for all $u \in X$. For Banach spaces U and V we let $\mathcal{L}(U,V)$ denote the space of bounded linear operators from U to V and $\mathcal{L}(U) = \mathcal{L}(U,U)$. Similarly $\mathcal{COM}(U,V)$ denotes the space of compact operators from U to V and $\mathcal{COM}(U) = \mathcal{COM}(U,U)$. Both $\mathcal{L}(U,V)$ and $\mathcal{COM}(U,V)$ are Banach spaces with the operator norm.

Unless we explicitly state otherwise we make the following *standard assumptions* [9].

ASSUMPTION 1.1. *A solution, $u^* \in X$, of* (1.1) *exists. F' exists and is Lipschitz continuous in a neighborhood $\mathcal{N}$ of u^*. $F'(u^*)$ has a bounded inverse, $F'(u^*)^{-1} \in \mathcal{L}(Y,X)$.*

We now describe Broyden's method in terms of the transition from a current pair of approximations, (u_c, B_c), to $(u^*, F'(u^*))$, to the next, (u_+, B_+), in the sequence of iterates.

All quasi-Newton methods follow the following general pattern. Given $(u_c, B_c) \in X \times \mathcal{L}(X,Y)$ with B_c nonsingular, compute:

1. the current function value, $F(u_c) \in Y$,
2. the Broyden step, $s = -B_c^{-1}F(u_c) \in X$,
3. the new approximation to u^*, $u_+ = u_c + s \in X$, and
4. the update for B_c, $B_+ \in \mathcal{L}(X,Y)$.

The quasi-Newton method is determined by the formula used to compute B_+ as a function of u_c, u_+, and B_c.

For $\theta_c \in (0,2)$ the Broyden update is

$$B_+ = B_c + \theta_c \frac{(y - B_c s) \otimes s}{\|s\|_H^2}. \tag{1.3}$$

In (1.3),

$$y = F(u_+) - F(u_c)$$

and $\otimes$ denotes the outer product, a rank-one operator in $\mathcal{L}(H,Y)$, defined for $g, h \in H$, and $f \in Y$ by

$$(f \otimes g)h = (g,h)f.$$

The usual Broyden update sets $\theta_c = 1$, but our results on linear problems will use other values of θ_c.

At this point we recall the relation between $\mathcal{L}(X,Y)$ and $\mathcal{L}(H,Y)$ by means of a lemma.

LEMMA 1.1. $\mathcal{L}(H,Y) \subset \mathcal{L}(X,Y)$. *If $A \in \mathcal{L}(X,Y)$ satisfies*

$$\|Au\|_Y \le M\|u\|_H$$

for some $M > 0$ and all $u \in X$, then A can be extended to be a map in $\mathcal{L}(H,Y)$ and

$$\|A\|_{\mathcal{L}(X,Y)} \le \|A\|_{\mathcal{L}(H,Y)} \le M.$$

Proof. The proof of the first assertion follows from the density of X in H and $\|\cdot\|_H \le \|\cdot\|_X$. To prove the second, note that for $u \in X$ and $u \ne 0$ we have

$$\frac{\|Au\|_Y}{\|u\|_X} \le \frac{\|Au\|_Y}{\|u\|_H} \le \|A\|_{\mathcal{L}(H,Y)}. \tag{1.4}$$

Taking suprema over $u \in X$ in (1.4) completes the proof. □

The relation between $\mathcal{L}(X,Y)$ and $\mathcal{L}(H,Y)$ is important because our results will all require that

$$E_0 = B_0 - F'(u^*) \in \mathcal{L}(H,Y). \tag{1.5}$$

In view of Lemma 1.1, (1.5) is equivalent to the existence of $M \ge 0$ such that

$$\|E_0 x\|_Y \le M\|x\|_H$$

for all $x \in X$.

The formulation of the Broyden update in terms of inner and outer products as well as technical details of the analysis of the convergence properties of the methods naturally lead to a Hilbert space or Euclidean finite-dimensional space setting. Such a setting can lead one to overlook significant properties of the iterates. As an example of the advantage of a Banach space analysis, let $r \geq 0$, $k \in C^r([0,1] \times [0,1])$, $g \in X = C^r([0,1])$, and take $(\cdot,\cdot)$ to be the L^2 inner product. The results of §2 will imply that if 1 is not an eigenvalue of the integral operator K defined by

$$(Ku)(x) = \int_0^1 k(x,\xi)u(\xi)\, d\xi,$$

then the Broyden iterates converge to a solution of $u - Ku = g$ from any initial iterate pair of the form (u_0, I), with $u_0 \in C^r$ arbitrary. Moreover, the convergence is global and locally q-superlinear in the norm of C^r. This is a stronger result than global and local q-superlinear convergence in the L^2 norm.

We will describe the behavior of the iterates in terms of the errors in the solution, $e = u - u^*$, and the approximate Fréchet derivative, $E = B - F'(u^*)$. We specify conditions that imply that the Broyden iterates converge locally q-superlinearly to the solution of (1.1). This means that

$$\lim_{n \to \infty} \frac{\|e_{n+1}\|_X}{\|e_n\|_X} = 0. \tag{1.6}$$

Our proofs will use the fact that q-linear convergence of the sequence $\{u_n\}$ and the Dennis–Moré condition,

$$\lim_{n \to \infty} \frac{\|E_n s_n\|_Y}{\|s_n\|_X} = 0, \tag{1.7}$$

imply local q-superlinear convergence [7].

As a final note in this introductory section, we point out that the Broyden iterates for $F(u) = 0$ are the same as those for $B_0^{-1}F(u) = 0$ and we could, therefore, assume that $X = Y$. Hence, for the theorems about problems with nonsingular Fréchet derivatives, the assumptions about the boundedness or compactness of E_0 as a map from H to Y can be replaced by similar assumptions on boundedness or compactness of $B_0^{-1}E_0$ as a map from H to X. We choose to state our results directly in terms of E_0. For the differential equations problems in the examples we will express the convergence criteria in terms of $B_0^{-1}E_0$ so that we can use the most familiar form of the equation.

2. Linear problems. We begin this section with a lemma of a type that is often used in q-superlinear convergence proofs for quasi-Newton methods. It is a weak form of the Frobenius norm estimates found in [3] and was first used in [18] to obtain weak q-superlinear convergence, proved again in a modified form in [29], and used in [27]. We state and prove the result in a slightly different form designed for use in this paper. Its statement as a separate lemma, its application to linear problems via the parameter θ as in [28], and its applicability to the analysis of partitioned updates, are the novel features of our version.

LEMMA 2.1. *Let H be a Hilbert space with inner product $(\cdot,\cdot)$, let $0 < \hat{\theta} < 1$, and let*

$$\{\theta_n\}_{n=0}^{\infty} \subset (\hat{\theta}, 2 - \hat{\theta}).$$

Let $\{\epsilon_n\}_{n=0}^{\infty} \subset H$ be such that

$$\sum_n \|\epsilon_n\|_H < \infty,$$

and let $\{\eta_n\}_{n=0}^{\infty}$ be a set of vectors in H such that $\|\eta_n\|_H$ is either 1 *or* 0 *for all n. Let $\psi_0 \in H$ be given. If $\{\psi_n\}_{n=1}^{\infty}$ is given by*

$$\psi_{n+1} = \psi_n - \theta_n(\eta_n, \psi_n)\eta_n + \epsilon_n, \tag{2.1}$$

then

$$\lim_{n\to\infty} (\eta_n, \psi_n) = 0. \tag{2.2}$$

Proof. The proof is trivial if $\epsilon_n = 0$ for all n and we give a proof for that case first. The sequence $\{\psi_n\}$ is bounded in H-norm by $\|\psi_0\|_H$ and satisfies

$$\|\psi_{n+1}\|_H^2 = \|\psi_n\|_H^2 - \theta_n(2-\theta_n)(\eta_n, \psi_n)^2.$$

Therefore, for any $M > 0$,

$$\sum_{n=0}^{M} (\eta_n, \psi_n)^2 \leq \frac{\|\psi_0\|_H^2 - \|\psi_{M+1}\|_H^2}{\hat{\theta}^2} \leq \frac{\|\psi_0\|_H^2}{\hat{\theta}^2}.$$

We let $M \to \infty$ to obtain

$$\sum_{n=0}^{\infty} (\eta_n, \psi_n)^2 < \infty,$$

which implies (2.2).

To prove the result for $\epsilon_n \neq 0$ we use the inequality

$$\sqrt{a^2 - b^2} \leq a - \frac{b^2}{2a}, \tag{2.3}$$

which is valid for $a > 0$ and $|b| \leq a$. This inequality is used often in the analysis of quasi-Newton methods [8]. From (2.3) we conclude that if $\psi_n \neq 0$, then

$$\begin{aligned}\|\psi_n - \theta_n(\eta_n, \psi_n)\eta_n\|_H &= \sqrt{\|\psi_n\|_H^2 - \theta_n(2-\theta_n)(\eta_n, \psi_n)^2} \\ &\leq \|\psi_n\|_H - \frac{\theta_n(2-\theta_n)(\eta_n, \psi_n)^2}{2\|\psi_n\|_H}.\end{aligned}$$

Hence if $\psi_n \neq 0$

$$\|\psi_{n+1}\|_H \leq \|\psi_n\|_H - \frac{\theta_n(2-\theta_n)(\eta_n, \psi_n)^2}{2\|\psi_n\|_H} + \|\epsilon_n\|_H. \tag{2.4}$$

Hence

$$(\eta_n, \psi_n)^2 \leq \frac{2\|\psi_n\|_H}{\theta_n(2-\theta_n)}(\|\psi_n\|_H - \|\psi_{n+1}\|_H + \|\epsilon_n\|_H), \tag{2.5}$$

which holds even if $\psi_n = 0$.

From (2.4) and (2.1) we conclude that

$$\|\psi_{n+1}\| \leq \mu,$$

where

$$\mu = \sum_{i=0}^{\infty} \|\epsilon_i\|_H + \|\psi_0\|_H.$$

Hence

$$\begin{aligned}\sum_{n=0}^{M} (\eta_n, \psi_n)^2 &\leq \tfrac{2\mu}{\hat{\theta}^2} \textstyle\sum_{n=0}^{M} (\|\psi_n\|_H - \|\psi_{n+1}\|_H + \|\epsilon_n\|_H) \\ &= \tfrac{2\mu}{\hat{\theta}^2} (\|\psi_0\|_H - \|\psi_{M+1}\|_H + \|\epsilon_n\|_H) \\ &\leq \tfrac{2\mu^2}{\hat{\theta}^2}.\end{aligned}$$

This completes the proof. □

As a first application of Lemma 2.1 we present results for the linear problem

$$F(u) = Au - b = 0, \tag{2.6}$$

where $A \in \mathcal{L}(X, Y)$ and $b \in Y$. In this special case, of course, the standard assumptions are simply that A is nonsingular. We seek a solution $u^* \in X$ by an iterative method that converges in the topology of X.

We let $u_0 \in X$; as we mentioned in §1 we will require

$$E_0 = B_0 - A \in \mathcal{L}(H, Y).$$

We begin this section with consideration of nonsingular A. As a simple corollary of those results on nonsingular linear problems we will obtain Hilbert space extensions of results in [13] for singular systems that show convergence of the Broyden iterates to least squares solutions.

It is clear that bounded deterioration in the sense of [3] holds. We state this as a lemma and we give the simple proof in order to introduce some notation.

LEMMA 2.2. *Let $\theta_c \in [0, 2]$, $u_c \in X$, and $B_c \in \mathcal{L}(X, Y)$ be given. Assume that B_c is nonsingular and that $E_c \in \mathcal{L}(H, Y)$. Then $E_+ \in \mathcal{L}(H, Y)$ and*

$$\|E_+\|_{\mathcal{L}(H,Y)} \leq \|E_c\|_{\mathcal{L}(H,Y)}. \tag{2.7}$$

Proof. First note that linearity of F implies that $y = As$. From the Broyden update (1.3), we see that if (u_c, B_c) is defined, then the error in the new approximate Fréchet derivative E_+ is related to E_c by

$$\begin{aligned}E_+ &= E_c + \theta_c \tfrac{(Au_+ - Au_c - B_c s) \otimes s}{\|s\|_H^2} \\ &= E_c - \theta_c \tfrac{(E_c s) \otimes s}{\|s\|_H^2} \\ &= E_c (I - \theta_c P_s),\end{aligned} \tag{2.8}$$

where P_s is the H-orthogonal projector

$$P_s = \frac{s \otimes s}{\|s\|_H^2}. \tag{2.9}$$

This completes the proof as

$$\|E_+\|_{\mathcal{L}(H,Y)} \leq \|E_c\|_{\mathcal{L}(H,Y)}\|I - \theta_c P_s\|_{\mathcal{L}(H)}$$

and $\|I - \theta_c P_s\|_{\mathcal{L}(H)} = 1$ by orthogonality of P_s and the fact that $0 \leq \theta_c \leq 2$. □

We remark that θ_c can always be selected to make B_+ nonsingular. In [28] one suggestion was

$$\theta_c = \begin{cases} 1, & |\gamma_c| \geq \sigma, \\ \dfrac{1 - \text{sign}(\gamma_c)\sigma}{1 - \gamma_c}, & \text{otherwise,} \end{cases}$$

where

$$\gamma_c = \frac{(B_c^{-1}y, s)}{\|s\|_H^2} = \frac{(B_c^{-1}As, s)}{\|s\|_H^2}$$

and $\sigma \in (0,1)$ is fixed. However, the results in [28] assume only that the sequence $\{\theta_n\}$ satisfies the hypotheses of Lemma 2.1 for some $\hat{\theta} \in (0,1)$ and that θ_c is always chosen so that B_+ is nonsingular.

We can verify the Dennis–Moré condition [7] and q-superlinear convergence under additional assumptions. We begin by applying ideas from [18], [28], and [27] and using Lemma 2.1 to prove a weak q-superlinear convergence result. We let Y^* be the Banach space dual of Y and let $\phi(u)$ denote the action of $\phi \in Y^*$ on $u \in Y$.

LEMMA 2.3. *Assume that $\{\theta_n\}$ satisfies the hypotheses of Lemma 2.1 for some $\hat{\theta} \in (0,1)$. If $E_0 \in \mathcal{L}(H,Y)$ and $\{\theta_n\}$ is such that the operators $\{B_n\}$ are nonsingular, then*

$$\lim_{n\to\infty} \phi\left(\frac{E_n s_n}{\|s_n\|_H}\right) = 0 \tag{2.10}$$

for all $\phi \in Y^*$.

Proof. Let $\phi \in Y^*$ be given. Note first that if $E \in \mathcal{L}(H,Y)$ and $E^* \in \mathcal{L}(Y^*,H)$ is the adjoint of E in $\mathcal{L}(H,Y)$, then for all $u \in Y$, $v \in H$, and $\phi \in Y^*$,

$$(v, E^*\phi) = \phi(Ev).$$

Since

$$E_{n+1}^*\phi = (I - \theta_n P_n)E_n^*\phi,$$

we may invoke Lemma 2.1 with

$$\eta_n = s_n/\|s_n\|_H, \quad \psi_n = E_n^*\phi, \quad \text{and} \quad \epsilon_n = 0$$

to conclude that

$$\begin{aligned} 0 &= \lim_{n\to\infty} (\eta_n, \psi_n) \\ &= \lim_{n\to\infty} \left(\frac{s_n}{\|s_n\|_H}, E_n^*\phi\right) \qquad (2.11) \\ &= \lim_{n\to\infty} \phi\left(\frac{E_n s_n}{\|s_n\|_H}\right). \end{aligned}$$

This completes the proof. □

In the main result of this section note that the proof of norm q-superlinear convergence follows the lines of that in [28] in that convergence itself is a consequence of the strong Dennis–Moré condition. The strong Dennis–Moré condition, as in [27], is implied by the weak condition and a compactness assumption on E_0.

An important concept in the proof of the next theorem, as in [27], is that of collective compactness [1]. We say that a sequence of linear operators $\{E_n\} \subset \mathcal{L}(U, V)$ between Banach spaces U and V is collectively compact if

$$\bigcup_n E_n \mathcal{B}_U(0:1)$$

is a precompact set in V. Here

$$\mathcal{B}_U(u_0 : \rho) = \{u \in U \mid \|u - u_0\|_U \leq \rho\}.$$

THEOREM 2.4. *Assume that $\{\theta_n\}$ satisfies the hypotheses of Lemma* 2.1 *for some $\hat{\theta} \in (0,1)$. If $E_0 \in \mathcal{COM}(H,Y)$ and $\{\theta_n\}$ is such that the operators $\{B_n\}$ are nonsingular, then* (1.7) *holds and $\{u_n\}$ converges to u^* globally and q-superlinearly in the norm of X.*

Proof. Let $s_n = u_{n+1} - u_n$ and let $P_n = P_{s_n}$. Since

$$E_n = E_0 \prod_{i=0}^{n-1} (I - \theta_i P_i),$$

the family $\{E_n\}$ is a collectively compact family of maps in $\mathcal{L}(H,Y)$ since $\|I - \theta_i P_i\|_{\mathcal{L}(H)} = 1$,

$$\bigcup_n E_n \mathcal{B}_H(0:1) = \bigcup_n E_0 \prod_{i=0}^{n-1} (I - \theta_i P_i) \mathcal{B}_H(0:1) \subset E_0 \mathcal{B}_H(0:1),$$

and $E_0 \mathcal{B}_H(0 : 1)$ is precompact in Y by assumption. Hence the sequence $\{\zeta_n\} = \{E_n s_n / \|s_n\|_H\}$ is precompact in Y. Therefore there is a Y-norm convergent subsequence $\zeta_{n_j} \to \zeta^*$.

By Lemma 2.3, $\phi(\zeta^*) = 0$ for all $\phi \in Y^*$ and therefore $\zeta^* = 0$. As $\{\zeta_{n_j}\}$ was an arbitrary convergent subsequence of $\{\zeta_n\}$, we can conclude that $\zeta_n \to 0$ in the Y-norm. This completes the proof of (1.7) as

$$\lim_{n\to\infty} \frac{\|E_n s_n\|_Y}{\|s_n\|_X} \leq \lim_{n\to\infty} \frac{\|E_n s_n\|_Y}{\|s_n\|_H} = \lim_{n\to\infty} \|\zeta_n\|_Y = 0,$$

by (1.2). As in [28], from (1.7) we can conclude convergence, which must be q-superlinear since (1.7) holds. □

A standard local convergence result is an easy corollary of Lemma 2.2 and Theorem 2.4.

COROLLARY 2.5. *For all $\bar{\delta} \in (0,1)$ there is δ such that if $\|E_0\|_{\mathcal{L}(H,Y)} < \delta$, then the Broyden iterates* (*with $\theta_n = 1$ for all n*) *converge q-linearly to u^* with q-factor $\sigma \leq \bar{\delta}$. If, in addition, $E_0 \in \mathcal{COM}(H,Y)$, then the convergence is q-superlinear.*

We conclude this section with remarks on the so-called bad Broyden method, linear least squares problems, and a partitioned form of Broyden's method.

2.1. The bad Broyden method. In this section we assume that $X = Y$. As mentioned above, by replacing F by $B^{-1}F$ one can see that this assumption can be made with no loss of generality. We do not make this assumption throughout the paper because the replacement of F by $B^{-1}F$ often leads to an unfamiliar form of the equation to be solved. Here, however, we will require an inner product structure on Y.

The bad Broyden method, so named because of its inferior performance in practice [8], is, for $\theta = 1$, the least change secant update [9] to B^{-1} that preserves the inverse secant equation $B^{-1}y = s$. The update is given by

$$B_+^{-1} = B_c^{-1} + \theta_c \frac{(s - B_c^{-1}y) \otimes y}{\|y\|_H^2}. \tag{2.12}$$

Let $\tilde{E} = B^{-1} - A^{-1}$; then

$$\tilde{E}_+ = \tilde{E}_c(I - \theta_c P_y),$$

where

$$P_y = \frac{y \otimes y}{\|y\|_H^2}.$$

The following theorem has a proof exactly like that of Theorem 2.4.

THEOREM 2.6. *Let the standard assumptions hold with $X = Y$, and assume that $\tilde{E}_0 \in \mathcal{COM}(H, X)$. Then if the sequence $\{\theta_n\}$ satisfies the hypotheses of Lemma 2.1 for some $\hat{\theta} \in (0,1)$ and is such that the operators B_n generated by (2.12) are nonsingular we have*

$$\lim_{n \to \infty} \frac{\|\tilde{E}_n y_n\|_X}{\|y_n\|_X} = 0.$$

Global and local q-superlinear convergence follow from Theorem 2.6 if the operators B_n^{-1} are uniformly bounded.

COROLLARY 2.7. *Let the assumptions of Theorem 2.6 hold and assume that the sequence $\{\|B_n\|_{\mathcal{L}(X)}\}$ is bounded. Then the bad Broyden iterates converge globally and q-superlinearly to $u^* = A^{-1}b$ in the norm of X.*

Proof. Let $M_B = \max\{\|B_n\|_{\mathcal{L}(X)}\}$ and let $E_n = B_n - A$. Note that $y = Au_+ - Au_c$. Since

$$\tilde{E}_n y_n = (B_n^{-1} - A^{-1})As_n = B_n^{-1}(A - B_n)s_n = -B_n^{-1}E_n s_n,$$

we have

$$\|E_n s_n\|_X \leq M_B \|\tilde{E}_n y_n\|_X.$$

Moreover,

$$\|s_n\|_X \geq \|A\|_{\mathcal{L}(X)}^{-1} \|y_n\|_X$$

and so

$$\frac{\|E_n s_n\|_X}{\|s_n\|_X} \leq M_B \|A\|_{\mathcal{L}(X)} \frac{\|\tilde{E}_n y_n\|_X}{\|y_n\|_X},$$

which completes the proof. □

2.2. Linear least squares problems. In finite dimension [13], a scheme based on Broyden's method was applied to systems of linear equations with singular coefficient matrices. The iterates converged to a least squares solution in $2N$ steps, where N was the number of unknowns. In this subsection we show that in the case of Broyden's method the iterates in [13] for $Au = b$ are the same as those given by Broyden's method itself when applied to the normal equation $A^*Au = A^*b$. Here A^* is the adjoint of A in the Hilbert space sense. We conclude from this observation that Theorem 2.4 is applicable and obtain an extension of the results in [13] to a Hilbert space setting. In our infinite-dimensional setting, of course, finite termination of the iteration is replaced by q-superlinear convergence.

We assume that $X = Y = H$ in this section. The proofs would be little changed if we assumed that X and Y were possibly different Hilbert spaces. Let $R(\cdot)$ denote range and $N(\cdot)$ denote null space of an operator. We assume that A has closed range $R(A)$. The iteration considered in [13] began with data $u_0 \in H$ and $G_0 \in \mathcal{L}(H)$ such that $R(G_0) = R(A^*)$, $N(G_0) = N(A^*)$. The operator G is intended to approximate the pseudo-inverse of A, $(A^*A)^{-1}A^*$, where $(A^*A)^{-1}$ is the inverse of the restriction of A^*A to $N(A)^\perp$, the orthogonal compliment of $N(A)$ in H. The iteration from (u_c, G_c) to (u_+, G_+) is

$$(2.13) \qquad s = -G_c(Au_c - b), \quad u_+ = u_c + s, \quad G_+ = G_c + (s - G_c y) \otimes v,$$

where, for Broyden's method,

$$v = \frac{\theta_c G^* s}{(1-\theta_c)\|s\|_H^2 + \theta_c(s, G_c y)}.$$

In the case of nonsingular A, (2.13) is simply the update on $G = B^{-1}$ induced by (1.3). Note that in this iteration the component of u_0 in $N(A)$ is never updated because $s \in R(A^*) = N(A)^\perp$, hence convergence is a question of convergence in the space $N(A)^\perp$. With initial data $u_0 \in N(A)^\perp$, one would hope to converge to the minimum norm least squares solution.

One iteration that does converge to the minimum norm least squares solution is Broyden's method applied to the normal equation in $N(A)^\perp$. The iterates (updating $\tilde{G} = B^{-1} \approx (A^*A)^{-1}$ and $\tilde{u} \approx (A^*A)^{-1}A^*b$) are given by

$$(2.14) \quad \tilde{s} = -\tilde{G}(A^*A\tilde{u}_c - A^*b), \quad \tilde{u}_+ = \tilde{u}_c + \tilde{s}, \quad \tilde{G}_+ = \tilde{G}_c + (\tilde{s} - \tilde{G}_c\tilde{y}) \otimes \tilde{v}.$$

Here $\tilde{y} = A^*A(\tilde{u}_+ - \tilde{u}_c) = A^*A\tilde{s}$ and

$$\tilde{v} = \frac{\theta_c \tilde{G}_c^* \tilde{s}}{(1-\theta_c)\|\tilde{s}\|_H^2 + \theta_c(\tilde{s}, \tilde{G}_c\tilde{y})}.$$

We now have the following lemma.

LEMMA 2.8. *Let A have null space $N(A)$ and closed range $R(A)$. Assume that $G_0 = \tilde{G}_0 A^*$ has range $N(A)^\perp = R(A^*)$, that $\tilde{u}_0 = u_0$, and the sequence $\{\theta_n\}$ is selected so that either of the sequences of maps $\{G_n\}$ or $\{\tilde{G}_n\}$ are nonsingular maps in $\mathcal{L}(N(A)^\perp)$. Then the sequences given by* (2.13) *and* (2.14) *are identical.*

Proof. From comparison of the two iterations above it is clear that if $u_c = \tilde{u}_c \in N(A)^\perp$ and $G_c = \tilde{G}_c A^*$, then

$$\tilde{s} = -\tilde{G}_c A^*(Au_c - b) = -G_c(Au_c - b) = s \in N(A)^\perp$$

and therefore $u_+ = \tilde{u}_+$. Therefore

$$\tilde{y} = A^*A\tilde{s} = A^*As = A^*y,$$

$$(s, G_c y) = (\tilde{s}, \tilde{G}_c A^* y) = (\tilde{s}, \tilde{G}_c \tilde{y}),$$

and

$$G_c^* s = (\tilde{G}_c A^*)^* \tilde{s} = A\tilde{G}_c^* s.$$

Hence $A\tilde{v} = v$ and therefore

$$(s - G_c y) \otimes v = (\tilde{s} - \tilde{G}_c A^* y) \otimes A\tilde{v} = ((\tilde{s} - \tilde{G}_c A^* y) \otimes \tilde{v})A^*.$$

Hence $G_+ = \tilde{G}_+ A^*$. This completes the proof. □

In view of Lemma 2.8 we may apply Theorem 2.4 for linear equations to least squares problems. We state this as Theorem 2.9.

THEOREM 2.9. *Assume that the assumptions of Lemma* 2.8 *hold and that* $G_0 - (A^*A)^{-1}A^* = Q_0 A^*$ *with* $Q_0 \in \mathcal{COM}(N(A)^\perp)$. *Assume that* $u_0 = u_0^N + u_0^R$, *with* $u_0^N \in N(A)$ *and* $u_0^R \in N(A)^\perp$. *Then the iterates given by* (2.13) *converge globally and q-superlinearly to* $(A^*A)^{-1}A^*b + u_0^N$.

2.3. Partitioned Broyden's method. Consider, for example, a linear system of integral equations,

$$u(x) - \int_\Omega k(x,\xi)u(\xi)\,d\xi - f(x) = 0,$$

where $\Omega \subset R^N$ is compact and an R^M-valued solution $u^* \in C(\Omega; R^M)$ is sought. Suppose k is an $M \times M$ matrix-valued function with a sparse structure

$$k(x,\xi)_{ij} = 0, \quad \text{if } j \notin \mathcal{I}_i. \tag{2.15}$$

Here $\mathcal{I}_i \subset \{1, \cdots, M\}$ is the set of indices of nonzero elements of row i of $k(x,\xi)$. We assume that $\mathcal{I}_i$ is independent of x and ξ. If we apply Broyden's method to this problem with $B_0 = I$, then

$$B_n = I + K_n,$$

where K_n is an integral operator with kernel

$$\kappa_n(x,\xi) \in C(\Omega \times \Omega; R^{M\times M}).$$

The Broyden update formula, with $H = L^2$, becomes

$$\kappa_{n+1}(x,\xi) = \kappa_n(x,\xi) + \frac{(y - B_n s)(x)s(\xi)^T}{\|s\|^2_{L^2(\Omega;R^M)}},$$

which does not reflect the sparsity of k. An update that does reflect that sparse structure could be based on the Schubert algorithm [30] applied to the matrix entries of κ. That update would take the form

$$\kappa_+(x,\xi)_{ij} = \begin{cases} \kappa_c(x,\xi)_{ij} + \frac{(y-B_c s)_i(x) s_j(\xi)}{\sum_{j\in\mathcal{I}_i}\int_\Omega s_j(t)^2\,dt}, & j \in \mathcal{I}_i, \\ 0, & j \notin \mathcal{I}_i. \end{cases} \tag{2.16}$$

We introduce some notation that will allow us to use the theory developed above to show that the iterates determined by (2.16) converge globally and locally q-superlinearly in $X = C(\Omega; R^M)$. We require the following assumption.

ASSUMPTION 2.1. *There are $\mathcal{E} \subset \mathcal{L}(H,Y)$, $\{\Pi_i\}_{i=1}^M \subset \mathcal{L}(Y)$, and $\{\pi_i\} \subset \mathcal{L}(X) \cap \mathcal{L}(H)$ such that*

$$\begin{aligned}
&\textstyle\sum_{i=1}^M \Pi_i = I, \\
&\Pi_i K = \Pi_i K \pi_i \in \mathcal{E} \quad \textit{for all } K \in \mathcal{E}, \\
(2.17)\qquad &\pi_i^2 = \pi_i, \\
&\Pi_i (u \otimes v)\pi_i \in \mathcal{E} \quad \textit{for all } u, v \in H, \textit{ and} \\
&(\pi_i u, v) = (u, \pi_i v) \quad \textit{for all } u, v \in H.
\end{aligned}$$

We consider updates of the form

$$\Pi_i B_+ = \Pi_i B_c + \theta_c \Pi_i \left((y - B_c s) \otimes \pi_i s\right) \left(\|\pi_i s\|_H^2\right)^+. \tag{2.18}$$

In (2.18) we use the superscript "+" for the standard pseudoinverse notation

$$\alpha^+ = \begin{cases} \frac{1}{\alpha}, & \text{if } \alpha \neq 0, \\ 0, & \text{if } \alpha = 0. \end{cases}$$

In the case of the integral equation above the update (2.16), $\mathcal{E}$ could be the space of integral operators with kernels in $C(\Omega \times \Omega; R^{M\times M})$ with sparsity pattern given by (2.15). The operators Π_i and π_i are given by

$$(\Pi_i u)_j(x) = \begin{cases} u_i(x), & \text{if } i = j, \\ 0, & \text{otherwise,} \end{cases}$$

for $u \in Y$ and

$$(\pi_i v)_j(x) = \begin{cases} v_j(x), & j \in \mathcal{I}_i, \\ 0, & j \notin \mathcal{I}_i \end{cases}$$

for $v \in H$.

If we consider methods with updates given by (2.18) with derivative errors in $\mathcal{E}$, then for $i = 1, \cdots, M$ we have by (2.17) that

$$\Pi_i E_+ = \Pi_i E_c - \theta_c \Pi_i (E_c s \otimes \pi_i s)(\|\pi_i s\|_H^2)^+ = \Pi_i E_c (I - \theta_c P_s^i), \tag{2.19}$$

where

$$P_s^i = \begin{cases} P_{\pi_i s}, & \text{if } \pi_i s \neq 0, \\ 0, & \text{if } \pi_i s = 0, \end{cases}$$

and as in (2.9),

$$P_{\pi_i s} = \frac{(\pi_i s) \otimes (\pi_i s)}{\|\pi_i s\|_H^2}.$$

We can now apply Lemma 2.1 for each $i = 1, \cdots, M$ to obtain Lemma 2.10.

LEMMA 2.10. *Let Assumption* 2.1 *hold and assume that* $\{\theta_n\}$ *satisfies the hypotheses of Lemma* 2.1 *for some* $\hat{\theta} \in (0,1)$. *Then if* $E_0 \in \mathcal{E}$ *and* $\{\theta_n\}$ *is such that the operators* $\{B_n\}$ *are nonsingular, then*

$$\lim_{n\to\infty} \phi(\Pi_i E_n s_n)\|\pi_i s_n\|_H^+ = 0 \tag{2.20}$$

for all $\phi \in Y^*$ *and* $i = 1, \cdots, M$.

Proof. Apply Lemma 2.1 for each $i = 1, \cdots, M$ to

$$\psi_n = (\Pi_i E_n)^* \phi, \quad \eta_n = \pi_i s_n \|\pi_i s_n\|_H^+, \quad \text{and} \quad \epsilon_n = 0$$

to conclude the result. □

The collective compactness argument used in Theorem 2.4 allows us to verify the strong Dennis–Moré condition from (2.20). We state this as Theorem 2.11.

THEOREM 2.11. *Assume that Assumption* 2.1 *holds, that* $\{\theta_n\}$ *satisfies the hypotheses of Lemma* 2.1 *for some* $\hat{\theta} \in (0,1)$, *that* $E_0 \in \mathcal{COM}(H,Y) \cap \mathcal{E}$, *and that* $\{\theta_n\}$ *is such that the operators* $\{B_n\}$ *are nonsingular; then* (1.7) *holds and* $\{u_n\}$ *converges to* u^* *globally and q-superlinearly in the norm of* X.

Proof. As in the proof of Theorem 2.4, (2.19) implies that the maps

$$\Pi_i E_n = \Pi_i E_0 \prod_{j=0}^{n-1} (I - \theta_j P_j^i)$$

form a collectively compact family for each $i = 1, \cdots, M$. Hence

$$\|\pi_i s_n\|_H^+ \Pi_i E_n s_n = \|\pi_i s_n\|_H^+ \Pi_i E_n \pi_i s_n \to 0$$

in the norm of Y. Hence there is $\{\delta_n\} \subset [0,\infty)$ with $\delta_n \to 0$ such that

$$\|\Pi_i E_n s_n\|_Y \le \delta_n \|\pi_i s_n\|_H$$

for all $i = 1, \cdots, M$. By Assumption 2.1

$$\begin{aligned} \|E_n s_n\|_Y = \left\| \sum_{i=1}^M \Pi_i E_n s_n \right\|_Y &\le \delta_n \sum_{i=1}^M \|\pi_i s_n\|_H \\ &\le M\delta_n \|s_n\|_H \le M\delta_n \|s_n\|_X. \end{aligned}$$

This completes the proof. □

3. Nonlinear problems.

3. Nonlinear problems. Unlike the results in §2 the results here are local convergence results. We consider nonlinear equations of the form (1.1) satisfying the standard assumptions. In order to prove results about Broyden's method, however, we must extend the standard assumptions and specify some properties of F' relative to the norm of the Hilbert space H. To this end, we make the following assumptions, which we refer to as the *extended standard assumptions.*

ASSUMPTION 3.1. *The standard assumptions hold and* F *can be split as*

$$F(u) = F_C(u) + F_A(u). \tag{3.1}$$

In (3.1)*, F_C is defined and Lipschitz continuously differentiable in X, with Lipschitz constant denoted by γ. F_A is also defined and Lipschitz continuously differentiable in X, and*

$$F_A'(u) \in \mathcal{L}(H, Y) \tag{3.2}$$

for each $u \in \mathcal{N}$. F_A' is a Lipschitz continuous map from $\mathcal{N}$ to $\mathcal{L}(H, Y)$ with Lipschitz constant γ_H.

Note that F_C may have no continuity or differentiability properties with respect to the H norm. F_C' will be computed or approximated by means independent of the Broyden updates. We will approximate only F_A' by Broyden's method and take into account in the analysis the accuracy in F_C'. Our formalism will be like that in [9], but the details will differ because one of our goals is to isolate the part of F' that is not continuous in the H norm. In [9] the issue was to separate the parts of F' that can be computed from those that are to be approximated by a quasi-Newton method. Here, therefore, we split the function, whereas in [9] the derivative was split.

As an example of a problem with such a splitting consider the integral equation from [21]:

$$F(u)(x) = h(u(x)) - \int_0^1 k(x, \xi, u(\xi))\, d\xi = 0.$$

On the space, $X = C[0,1]$ with $H = L^2[0,1]$. Here $F_C(u)(x) = h(u(x))$ is, in general, not a Fréchet differentiable map on L^2 and one could take advantage of the fact that in $C[0,1]$, F_C is Fréchet differentiable and $F_C'(u)v(x) = h'(u(x))v(x)$, which could be easily computed explicitly or by differences. In the case of approximation by a quadrature rule, for example, F_C' could be represented as a diagonal matrix. The integral part F_A' would then be updated by Broyden's method. We return to this equation later.

For nonlinear problems we set $\theta_c = 1$ and let the update formula for $A \approx F_A'(u^*)$ be

$$A_+ = A_c + \frac{(y^\# - A_c s) \otimes s}{\|s\|_H^2}. \tag{3.3}$$

In (3.3) $y^\#$ is intended to reflect information obtained from the most recent iterate. In the finite-dimensional analysis in [9], $y^\#$ could be any vector satisfying

$$\|y^\# - y_D^\#\|_{R^n} = O(\|e_c\|_{R^n} \|s\|_{R^n}) \tag{3.4}$$

where $y_D^\#$, the "default choice," is

$$y_D^\# = y - C_+ s$$

where C_+ is the "computed part" of F', which in our context is a sufficiently accurate approximation to F_C'. The choice of $y_D^\#$ makes the secant equation $B_+ s = (A_+ + C_+)s = y$ hold. In the Banach space setting considered here, however, we must sacrifice the secant equation and use

$$y^\# = y_B^\# = y - F_C(u_+) + F_C(u_c) = F_A(u_+) - F_A(u_c). \tag{3.5}$$

As we will see from the analysis to follow, any choice that differs from the choice in (3.5) by $O(\|e\|_X \|s\|_H)$ will suffice as well. Since $\|y_D^\# - y_B^\#\| = O(\|e\|_X \|s\|_X)$ and

$\|s\|_X$ cannot be bounded in terms of $\|s\|_H$, the choice $y^{\#} = y^{\#}_D$ will not work unless $X = H$.

The proof of local q-superlinear convergence follows the outline of the proof of the Dennis–Moré condition in §2 after we prove q-linear convergence. This task is very similar to the finite-dimensional case and requires only use of Lemma 1.1. We must introduce some notation. We let

$$B = C + A,$$

where $A \approx F'_A(u^*)$ is updated by (3.3) and $C \approx F'_C(u^*)$ is computed by other means. We let

$$E^C = C - F'_C(u^*) \quad \text{and} \quad E^A = A - F'_A(u^*).$$

Note that

$$y^{\#} - A_c s = F_A(u_+) - F_A(u_c) - F'_A(u^*)s - E^A_c s = -E^A_c s + \Delta^A_c s, \tag{3.6}$$

where

$$\Delta^A_c = \int_0^1 F'_A(u_c + ts) - F'_A(u^*)\,dt. \tag{3.7}$$

Hence

$$E^A_+ = E^A_c(I - P_s) + \Delta^A_c P_s. \tag{3.8}$$

As in the finite-dimensional case (e.g., see [8]) we use the extended standard assumptions to conclude that

$$\|\Delta^A_c\|_{\mathcal{L}(H,Y)} \leq \frac{\gamma_H}{2}(\|e_+\|_X + \|e_c\|_X). \tag{3.9}$$

The estimate (3.9) is basic to the remainder of this section. With (3.9) in hand we obtain the following theorem on q-linear convergence. The proof is exactly the same as in finite dimensions and we omit it.

THEOREM 3.1. *Let the extended standard assumptions hold. Let $\sigma \in (0,1)$ and let $\bar{\delta} > 0$. Then there is $\delta > 0$ such that if*

$$\|E^C_n\|_{\mathcal{L}(X,Y)} < \delta \quad \textit{for all } n, \quad \|e_0\|_X < \delta, \quad \textit{and} \quad \|E^A_0\|_{\mathcal{L}(H,Y)} < \delta, \tag{3.10}$$

then the Broyden iterates converge q-linearly to u^* in X with q-factor at most σ and

$$\|E_n\|_{\mathcal{L}(H,Y)} < \bar{\delta}. \tag{3.11}$$

The weak q-superlinear convergence result follows immediately from Lemma 2.1.

LEMMA 3.2. *Assume that the extended standard assumptions hold and that* (3.10) *holds with δ small enough for the conclusions of Theorem* 3.1 *to hold with some $\sigma \in (0,1)$ and $\bar{\delta} > 0$. Then*

$$\lim_{n\to\infty} \phi\left(\frac{E_n s_n}{\|s_n\|_H}\right) = 0$$

for all $\phi \in Y^*$.

Proof. As in §2, set $P_n = P_{s_n}$. Note that

$$(E^A_{n+1})^*\phi = (I - P_n)(E^A_n)^*\phi + P_n(\Delta^A_n)^*\phi.$$

We apply Lemma 2.1 with

$$\psi_n = (E^A_n)^*\phi, \quad \eta_n = \frac{s_n}{\|s_n\|_H}, \quad \text{and} \quad \epsilon_n = P_n(\Delta^A_n)^*\phi.$$

The hypothesis of Lemma 2.1 that

$$\sum \|\epsilon_n\|_H < \infty$$

holds by (3.9) and Theorem 3.1. This completes the proof as the invocation of Lemma 2.1 completed the proof of Lemma 2.3 in §2. □

The main result in this section is that compactness of E^A_0 forces the strong Dennis–Moré condition to hold. The proof is similar to that for the Hilbert space case in [27]. This will in turn imply that the q-superlinear convergence of the sequence $\{u_n\}$ is controlled by the sequence of operator errors $\{E^C_n\}$.

THEOREM 3.3. *Let the assumptions of Lemma 3.2 hold. If* $E^A_0 \in \mathcal{COM}(H, Y)$, *then the family* $\{E^A_n\}$ *is collectively compact and*

$$\lim_{n\to\infty} \frac{\|E^A_n s_n\|_Y}{\|s_n\|_X} = 0. \tag{3.12}$$

Proof. For all $n \geq 0$,

$$E^A_{n+1} = E^A_0 \prod_{k=0}^{n}(I - P_k) + \sum_{k=0}^{n} \Delta^A_k P_k \prod_{l=k+1}^{n} (I - P_l).$$

Since $E^A_0 \in \mathcal{COM}(H, Y)$ by assumption and

$$\left\| \prod_{k=0}^{n}(I - P_k) \right\|_{\mathcal{L}(H)} = 1,$$

it suffices to show that the set

$$\bigcup_{n=0}^{\infty} \sum_{k=0}^{n} \Delta^A_k P_k \mathcal{B}_H(0:1) = \sum_{k=0}^{\infty} \Delta^A_k P_k \mathcal{B}_H(0:1)$$

is precompact in Y.

Recall that $\Delta^A_n P_n$ is a rank-one operator whose range is the span of the vector

$$z_n = \frac{\Delta^A_n s_n}{\|\Delta^A_n s_n\|_Y}$$

and that

$$\|\Delta^A_n P_n\|_{\mathcal{L}(H,Y)} \leq \frac{\gamma_H}{2}(\|e_{n+1}\|_X + \|e_n\|_X) \leq \frac{\gamma_H(1+\sigma)}{2}\sigma^n,$$

where σ is the q-factor of q-linear convergence in the statement of Theorem 3.1.

Therefore the operators $\sum_{k=0}^{n} \Delta_k^A P_k$ are a sequence of finite-rank operators that converge in the operator norm to $\sum_{k=0}^{\infty} \Delta_k^A P_k$. Therefore $\sum_{k=0}^{\infty} \Delta_k^A P_k$ is a compact operator and $\sum_{k=0}^{\infty} \Delta_k^A P_k \mathcal{B}_H(0:1)$ is a compact set. Having the collective compactness of $\{E_n^A\}$, (3.12) holds by an argument exactly the same as that used in the linear case. □

The next result of this section is a combination of Theorem 3.3 and the standard results on equivalence of the Dennis–Moré conditions to q-superlinear convergence.

THEOREM 3.4. *Let the hypotheses of Theorem* 3.3 *hold. Then $u_n \to u^*$ q-superlinearly in the norm of X if and only if*

$$\lim_{n\to\infty} \frac{\|E_n^C s_n\|_Y}{\|s_n\|_X} = 0. \tag{3.13}$$

Examples of situations in which (3.13) holds include exact computation of $F_C'(u_n)$ and computation by differences where the step size in the difference computation tends to zero.

The analysis of the bad Broyden method follows lines similar to those of the standard method such as in the linear case discussed in §2. We replace the extended standard assumptions by Assumption 3.2.

ASSUMPTION 3.2. *The standard assumptions hold. $X = Y$. $F(u)$ is defined and Lipschitz continuously differentiable in X, with Lipschitz constant denoted by γ, $F'(u)$ can be extended to $\mathcal{L}(H,H)$ for each $u \in \mathcal{N}$, and F' is a Lipschitz continuous map from $\mathcal{N}$ to $\mathcal{L}(H,H)$ with Lipschitz constant γ_H. Moreover, there is M_H such that*

$$\|F'(u)w\|_H \geq M_H \|w\|_H \tag{3.14}$$

for all $w \in X$.

The update is

$$B_+^{-1} = B_c^{-1} + \frac{(s - B_c^{-1}y) \otimes y}{\|y\|_H^2}.$$

We have the following theorem.

THEOREM 3.5. *Let Assumption* 3.2 *hold. Let $\sigma \in (0,1)$. Then there is $\delta > 0$ such that if*

$$\|E_0\|_{\mathcal{L}(H,X)} \leq \delta \quad \textit{and} \quad \|e_0\|_X < \delta,$$

the bad Broyden iterates converge q-linearly to u^* in the norm of X with q-factor at most σ. If, in addition, $B_0^{-1} - F'(u^*)^{-1} \in \mathcal{COM}(H,X)$, the bad Broyden iterates converge q-superlinearly to u^* in the norm of X.

Proof. We give those details in the proof that differ from those in the analysis of the standard Broyden method. As in the linear case, we let $\tilde{E} = B^{-1} - F'(u^*)^{-1}$. It is clear that

$$\tilde{E}_+ = \tilde{E}_c + \frac{(s - B_c^{-1}y) \otimes y}{\|y\|_H^2}.$$

Now

$$\begin{aligned} s - B_c^{-1}y &= -E_c y - F'(u^*)^{-1}(F'(u^*)s - y) \\ &= -E_c y + F'(u^*)^{-1}\Delta_c s, \end{aligned}$$

where, recalling (3.7),

$$\Delta_c = \int_0^1 F'(u_c + ts) - F'(u^*)\, dt.$$

Hence

$$\tilde{E}_+ = \tilde{E}_c(I - P_y) + \frac{(F'(u^*)^{-1}\Delta_c s) \otimes y}{\|y\|_H^2}.$$

The proof will follow the lines of that for Theorems 3.1 and 3.3 if we can estimate the sequence of rank-one operators $\{T_n\}$, where

$$T_n = \frac{(F'(u^*)^{-1}\Delta_n s_n) \otimes y_n}{\|y_n\|_H^2}.$$

The operators T_n play the role played by $\Delta_n^A P_n$ in the proof of Theorem 3.3.

Since

$$y = F(u_+) - F(u_c) = \int_0^1 F'(u_c + ts)s\, dt = F'(u^*)s + \int_0^1 (F'(u_c + ts) - F'(u^*))s\, dt,$$

the assumptions imply that

$$\|y\|_H \geq M_H\|s\|_H - \frac{\gamma_H}{2}\|s\|_X\|s\|_H \geq \left(1 - \frac{\gamma_H\delta}{2}\right)\|s\|_H.$$

Therefore, there is C_T such that

$$\|T_n\|_{\mathcal{L}(H,X)} \leq C_T\|s_n\|_X \leq C_T(\|e_n\|_X + \|e_{n+1}\|_X).$$

This estimate plays the role of (3.9) in the proof of q-linear convergence and then can be used to show that $\sum \|T_n\| < \infty$ to conclude that the convergence is superlinear if δ is sufficiently small and if $E_0 \in \mathcal{COM}(H, X)$. □

The analysis of the partitioned form of Broyden's method is similar to the linear case as well.

THEOREM 3.6. *Let the extended standard assumptions and Assumption* 2.1 *hold. Then there is $\delta > 0$ such that if $E_0^A \in \mathcal{E}$, $\|E_0^A\|_{\mathcal{L}(H,Y)} < \delta$, $\|E_n^C\|_{\mathcal{L}(X,Y)} < \delta$ for all n, and $\|e_0\| < \delta$, the iterates given by the update*

$$\Pi_i A_+ = \Pi_i A_c + \Pi_i\left((y^\# - A_c s) \otimes \pi_i s\right)\left(\|\pi_i s\|_H^2\right)^+$$

converge q-linearly to u^*. If, moreover, $E_0 \in \mathcal{COM}(H, Y)$ and (3.13) *holds, then the convergence is q-superlinear.*

We close this section with an application to nonlinear integral equations of the form

$$h(u(x)) - \int_\Omega k(x, \xi, u(\xi))\, d\xi = 0. \tag{3.15}$$

In (3.15) Ω is a closed and bounded subset of R^M for some M. We assume that $h \in C^r(R^N; R^N)$, $k \in C^r \cup L^\infty(\Omega \times \Omega \times R^N; R^N)$, and k_3, the derivative of k with respect to its third argument, is in $C^r \cup L^\infty(\Omega \times \Omega \times R^N; R^{N\times N})$. In [21] a pointwise quasi-Newton method was proposed for systems of this form. In that approach approximations for the matrix-valued functions $h'(u^*)(x)$ and $k_3(x, \xi, u^*(\xi))$ were maintained separately. As is typical with pointwise methods (see, for example, [25] and [22]), local q-linear convergence could be proved. One can also show q-superlinear convergence in the sense that

$$\frac{\|e_{n+1}\|_{L^r(\Omega;R^M)}}{\|e_n\|_{L^s(\Omega;R^M)}} \to 0$$

for $2 \leq r < s \leq \infty$. If one is willing to compute h' analytically and approximate the derivative of the integral operator by Broyden's method, then the results of this section are applicable with $X = Y = C^r(\Omega; R^N)$, $H = L^2(\Omega; R^N)$, $F_C(u) = h(u)$, and

$$F_A(u)(x) = -\int_\Omega k(x, \xi, u(\xi))\, d\xi.$$

$F'_A(u^*) \in \mathcal{COM}(H, Y)$ by the assumption that k_3 is uniformly bounded. The quasi-Newton update given by (3.3) will give local q-superlinear convergence in C^r. If one approximates h' by a forward fixed difference with difference step σ the convergence will be q-linear with a q-factor that is $O(\sigma)$. Note also that any sparsity properties of the kernel k_3 could be preserved by a partitioning approach as described in §2 and that q-superlinear convergence would still take place by Theorem 3.6.

4. Numerical examples.

All of the computations reported in this section were done on the Cray Y-MP at the North Carolina Supercomputing Center under Cray UNICOS v5.1 and the CFT77 v4.0 compiler.

4.1. Linear problems.

This section is divided into sections on linear and nonlinear problems. In this first subsection on linear problems we begin with an integral equation arising in potential theory,

$$u(x) - \lambda^{-1} \int_0^1 k(x + \xi) u(\xi)\, d\xi = f(x), \tag{4.1}$$

where

$$k(x) = \frac{1 - \gamma^2}{1 + \gamma^2 - 2\gamma \cos(2\pi x)}.$$

The parameters λ and γ are given. This equation is an integral equation form of Laplace's equation on an ellipse [19]. It was used as a test problem in the context of fast two grid algorithms in [2].

In the computations reported here we use

$$\lambda = .2, \quad \gamma = .8, \quad \text{and} \quad f(x) = \sin(5\pi x)/\lambda.$$

Note that $k(x) \in C^\infty$ for $|\gamma| < 1$ and hence the integral operator K with kernel $k(x + \xi)$ is a compact operator from L^2 to C^r for any $r \geq 0$. Hence we may apply Theorem 2.4 with $X = Y = C^r$ and $H = L^2$. We could also apply the theorem with $X = L^2$.

In the computations reported in Tables 4.1–4.4, we approximated integrals with the 21 and 401 point composite Simpson's rules. The discrete L^2 inner product and norm were computed using the quadrature rule. We report results for both $X = L^2$ and $X = C^1$ to compare the q-superlinear rates and for both grids to illustrate how

TABLE 4.1
Linear integral equation, $N = 21$, L^2 norm.

i	$\|F\|$	R_F	$\|s\|$	R_S
0	0.3536E+01		0.3536E+01	
1	0.1101E+01	0.3113	0.1355E+01	0.3832
2	0.1314E+01	1.1943	0.1438E+02	10.6153
3	0.6848E+01	5.2101	0.1876E+02	1.3043
4	0.4552E+00	0.0665	0.8001E+00	0.0426
5	0.1309E+00	0.2875	0.2736E+00	0.3420
6	0.1654E−01	0.1264	0.3518E−01	0.1286
7	0.1501E−02	0.0908	0.3461E−02	0.0984
8	0.1800E−04	0.0120	0.4153E−04	0.0120
9	0.1604E−06	0.0089	0.3691E−06	0.0089
10	0.1480E−08	0.0092	0.3420E−08	0.0093
11	0.6529E−11	0.0044		

TABLE 4.2
Linear integral equation, $N = 21$, C^1 norm.

i	$\|F\|$	R_F	$\|s\|$	R_S
0	0.7571E+02		0.7571E+02	
1	0.1249E+02	0.1650	0.1538E+02	0.2031
2	0.3228E+01	0.2585	0.3874E+02	2.5193
3	0.1932E+02	5.9839	0.4424E+02	1.1419
4	0.2658E+01	0.1376	0.3586E+01	0.0811
5	0.9012E+00	0.3391	0.1497E+01	0.4176
6	0.1278E+00	0.1418	0.2168E+00	0.1448
7	0.1176E−01	0.0920	0.2162E−01	0.0997
8	0.1425E−03	0.0121	0.2621E−03	0.0121
9	0.1285E−05	0.0090	0.2355E−05	0.0090
10	0.1193E−07	0.0093	0.2193E−07	0.0093
11	0.5343E−10	0.0045		

the performance does not depend on the level of approximation, but is governed by the properties of the continuous problem.

Initial data was $B_0 = I$ and $u_0 = 0$ in all cases. We tabulate the iteration counter i, and

$$\|F(u_i)\|_X, R_F = \frac{\|F(u_i)\|_X}{\|F(u_{i-1})\|_X},$$

$$\|s_i\|_X, R_S = \frac{\|s_i\|_X}{\|s_{i-1}\|_X}.$$

The discrete C^1 norm was computed as

$$\|u\|_{C^1} = \sup_{1\le j\le N} |u_j| + \sup_{1\le j\le N-1} \frac{|u_{j+1} - u_j|}{|x_{j+1} - x_j|},$$

where N is the number of quadrature points and $\{x_j\}_{j=1}^N$ are the nodes of the quadrature rule. The iteration was terminated when $\|F(u_i)\|_X < 10^{-10}$.

TABLE 4.3
Linear integral equation, $N = 401$, L^2 norm.

i	$\|F\|$	R_F	$\|s\|$	R_S
0	0.3536E+01		0.3536E+01	
1	0.1098E+01	0.3105	0.1351E+01	0.3820
2	0.1311E+01	1.1942	0.1441E+02	10.6663
3	0.6853E+01	5.2274	0.1877E+02	1.3030
4	0.4537E+00	0.0662	0.7969E+00	0.0425
5	0.1306E+00	0.2879	0.2735E+00	0.3432
6	0.1637E−01	0.1253	0.3472E−01	0.1269
7	0.1524E−02	0.0931	0.3520E−02	0.1014
8	0.1506E−04	0.0099	0.3460E−04	0.0098
9	0.1842E−06	0.0122	0.4250E−06	0.0123
10	0.1179E−08	0.0064	0.2723E−08	0.0064
11	0.5414E−11	0.0046		

TABLE 4.4
Linear integral equation, $N = 401$, C^1 norm.

i	$\|F\|$	R_F	$\|s\|$	R_S
0	0.8352E+02		0.8352E+02	
1	0.1270E+02	0.1520	0.1562E+02	0.1871
2	0.3248E+01	0.2558	0.3898E+02	2.4954
3	0.1949E+02	5.9999	0.4438E+02	1.1385
4	0.2667E+01	0.1369	0.3585E+01	0.0808
5	0.9117E+00	0.3418	0.1502E+01	0.4189
6	0.1285E+00	0.1410	0.2161E+00	0.1439
7	0.1210E−01	0.0941	0.2216E−01	0.1026
8	0.1215E−03	0.0100	0.2211E−03	0.0100
9	0.1498E−05	0.0123	0.2736E−05	0.0124
10	0.9674E−08	0.0065	0.1766E−07	0.0065
11	0.6651E−10	0.0069		

Our next example is the two point boundary value problem

$$F(u) = u'' + u' + u - 1 = 0 \quad \text{in } [0,1], \quad u(0) = u(1) = 0. \tag{4.2}$$

For the examples in boundary value problems we will express the compactness criterion for q-superlinear convergence as $B_0^{-1}E_0 \in \mathcal{COM}(H, X)$. As mentioned in the introduction, this is an equivalent formulation, as the Broyden iterates for $F(u) = 0$ and those for $B_0^{-1}F(u) = 0$ are identical. In the process of forming the iterates we will have to understand that derivatives are in the distributional sense. We let

$$B_0 = d^2/dx^2$$

with homogeneous Dirichlet boundary conditions. Since $E_0 = d/dx + I$ is a first-order operator, $B_0^{-1}E_0 \in \mathcal{COM}(L^2[0,1], C[0,1])$ and the theory in §2 is applicable with $X = C[0,1]$ and $H = L^2[0,1]$. In Tables 4.5 and 4.6 we tabulate the history of the iteration for two different mesh sizes, $h = 1/32$ and $h = 1/2048$, corresponding to $N = 31$ and 2047 unknowns. Discretization was by central differences and the initial

TABLE 4.5
Two point boundary value problem, $h = 1/32$, L^∞ norm.

i	$\|F\|$	R_F	$\|s\|$	R_S
0	0.9068E+02		0.9068E+02	
1	0.1420E+02	0.1566	0.1608E+02	0.1774
2	0.2403E+01	0.1692	0.2337E+01	0.1453
3	0.2863E+00	0.1191	0.3190E+00	0.1365
4	0.1001E−01	0.0350	0.1058E−01	0.0332
5	0.4675E−03	0.0467	0.5075E−03	0.0480
6	0.3904E−04	0.0835		

TABLE 4.6
Two point boundary value problem, $h = 1/2048$, L^∞ norm.

i	$\|F\|$	R_F	$\|s\|$	R_S
0	0.9073E+02		0.9073E+02	
1	0.1423E+02	0.1568	0.1611E+02	0.1776
2	0.2418E+01	0.1699	0.2349E+01	0.1458
3	0.2888E+00	0.1194	0.3219E+00	0.1370
4	0.1016E−01	0.0352	0.1071E−01	0.0333
5	0.4769E−03	0.0469	0.5192E−03	0.0485
6	0.4032E−04	0.0845	0.4093E−04	0.0788
7	0.2801E−05	0.0695	0.3060E−05	0.0748
8	0.7255E−07	0.0259	0.7685E−07	0.0251
9	0.2274E−08	0.0313		

iterate was $u_0 = 0$. The iteration was terminated when $\|B_0^{-1}F\|_X < h^2/10$. Note the uniformity in the convergence rates on the two grids. The finite-dimensional theory guarantees convergence in $2N$ steps but convergence to truncation error requires far fewer steps for discretizations of many infinite-dimensional problems.

To close this subsection we consider the partial differential equation

$$F(u) = \nabla^2 u + du_x + u - 1 = 0,$$

on $[0,1] \times [0,1]$ with homogeneous Dirichlet boundary conditions. Letting $B_0 = (\nabla^2)^{-1}$ and $H = L^2$ means that X must be chosen such that the map $B_0^{-1}E_0 \in \mathcal{COM}(H,X)$. For $d = 0$ the choice $X = L^\infty$ and $H = L^2$ is appropriate. For $d \neq 0$, this is not an option [14]. For $H = L^2$ the map $B_0^{-1}E_0 \in \mathcal{COM}(H,X)$ for many choices of X including $X = L^p$, for any $p \in [1,\infty)$ and for $X = H^\alpha$ for any $\alpha \in [0,1)$. For $X = L^\infty$ we must use a stronger inner product than L^2 to obtain the compactness of E_0, and $H = H^\alpha$ for any $\alpha > 0$ would suffice. In both cases we are required to interpret solution as in the weak sense or to consider the preconditioned equation obtained by replacing F by $B_0^{-1}F$, for which the Broyden iterates will be the same.

The situation for $d \neq 0$, requiring an H^α inner product for $\alpha > 1$ is only slightly different from that for $d = 0$, needing an $H^0 = L^2$ inner product, and it is interesting to see how the convergence rates compare. To that end we present two sets of tables: one, Tables 4.7 and 4.8, for $d = 0$, and the other, Tables 4.9 and 4.10, for $d = 10$. In order to facilitate comparison we use the L^∞ norm and L^2 inner product in both sets. We report results from two discretizations for $h = 1/16$ and $h = 1/512$. The

TABLE 4.7
Elliptic boundary value problem, $d = 0$, $h = 1/16$, L^∞ norm.

i	$\|F\|$	R_F	$\|s\|$	R_S
0	0.7345E−01		0.7345E−01	
1	0.4061E−02	0.0553	0.4276E−02	0.0582
2	0.5844E−05	0.0014		

TABLE 4.8
Elliptic boundary value problem, $d = 0$, $h = 1/512$, L^∞ norm.

i	$\|F\|$	R_F	$\|s\|$	R_S
0	0.7367E−01		0.7367E−01	
1	0.4062E−02	0.0551	0.4277E−02	0.0581
2	0.5801E−05	0.0014	0.5875E−05	0.0014
3	0.6112E−07	0.0105		

iterations were terminated when $\|B_0^{-1}F\|_X < h^2/10$. In the computations, $u_0 = 0$ was the initial iterate and central differences were used to approximate derivatives. The action of $(\nabla^2)^{-1}$ on a vector was computed with the Poisson solver from FISHPACK [32].

4.2. Nonlinear problems. The Chandrasekhar H-equation [5]

$$F(u)(x) = u(x) - \mathcal{K}(u)(x) = u(x) - (1 - (Lu)(x))^{-1},$$

where for $x \in [0, 1]$,

$$(Lu)(x) = \frac{c}{2}\int_0^1 \frac{xu(\xi)\, d\xi}{x + \xi}.$$

The equation arises in radiative transfer. The parameter $c \in (0, 1]$; $F'(u^*)$ becomes singular at $c = 1$ and for $c < 1$ but near 1 the problem becomes more difficult. The solution $u^* \in C[0, 1]$ and

$$\mathcal{K}'(u^*)(w)(x) = \frac{(Lw)(x)}{(1 - (Lu^*)(x))^2} = u^*(x)^2(Lw)(x).$$

It is easy to see that the image of the unit ball in $L^2[0, 1]$ under L is a uniformly bounded equicontinuous family of functions and therefore $L \in \mathcal{COM}(L^2[0, 1], C[0, 1])$. As u^* is continuous, this implies that

$$\mathcal{K}'(u^*) \in \mathcal{COM}(L^2[0, 1], C[0, 1]).$$

These considerations motivate our application of Theorem 3.4 with $H = L^2[0, 1]$, $X = C[0, 1]$, $F_A = \mathcal{K}$, and $F_C = I$. As we can expect only local convergence for this nonlinear problem, we must pay attention to the quality of the initial data, u_0 and A_0. This will become apparent from the numerical results.

All computations were done with a 20 point composite Gauss rule with 20 subintervals. In Tables 4.11 and 4.12 we let $c = \frac{1}{2}$, $u_0 = 0$, $C = F_C'(u^*) = I$, and $A_0 = 0$. The iteration was terminated when $\|F(u_i)\|_X < 10^{-8}$. We tabulate the same quantities as in the previous tables. The tables indicate that convergence is q-superlinear for both $X = L^2[0, 1]$ and $X = C[0, 1]$.

TABLE 4.9

Elliptic boundary value problem, $d = 10$, $h = 1/16$, L^∞ norm.

i	$\|F\|$	R_F	$\|s\|$	R_S
0	0.7345E−01		0.7345E−01	
1	0.4005E−01	0.5453	0.4218E−01	0.5743
2	0.6899E−01	1.7225	0.4255E−01	1.0088
3	0.1581E−01	0.2291	0.1387E−01	0.3261
4	0.5077E−02	0.3212	0.4226E−02	0.3046
5	0.3995E−02	0.7867	0.4083E−02	0.9662
6	0.1045E−02	0.2615	0.1036E−02	0.2536
7	0.2136E−03	0.2045		

TABLE 4.10

Elliptic boundary value problem, $d = 10$, $N = 1/512$, L^∞ norm.

i	$\|F\|$	R_F	$\|s\|$	R_S
0	0.7367E−01		0.7367E−01	
1	0.4034E−01	0.5476	0.4248E−01	0.5766
2	0.7019E−01	1.7399	0.4333E−01	1.0200
3	0.1623E−01	0.2312	0.1425E−01	0.3290
4	0.5487E−02	0.3381	0.4475E−02	0.3140
5	0.4422E−02	0.8059	0.4455E−02	0.9954
6	0.1238E−02	0.2800	0.1199E−02	0.2692
7	0.2857E−03	0.2307	0.2061E−03	0.1719
8	0.1293E−03	0.4527	0.1374E−03	0.6668
9	0.4527E−04	0.3501	0.4518E−04	0.3287
10	0.5275E−05	0.1165	0.4292E−05	0.0950
11	0.1819E−05	0.3448	0.1898E−05	0.4422
12	0.6351E−06	0.3492	0.6392E−06	0.3368
13	0.5309E−07	0.0836		

TABLE 4.11

$c = .5$, $N = 400$, L^2 norm.

i	$\|F\|$	R_F	$\|s\|$	R_S
0	0.1000E+01		0.1000E+01	
1	0.1545E+00	0.1545	0.1807E+00	0.1807
2	0.2921E−02	0.0189	0.3146E−02	0.0174
3	0.1067E−03	0.0365	0.1234E−03	0.0392
4	0.4153E−05	0.0389	0.4990E−05	0.0404
5	0.7770E−08	0.0019		

In Tables 4.13 and 4.14 we used the same data as in the computation above, except we set $c = .999$. For that value of c the $F'(u^*)$ is nearly singular. Broyden's method will converge at best q-linearly [6], [20] when $c = 1$. Therefore, one would expect slower convergence than for $c = .5$. The tables indicate, however, that the convergence is still q-superlinear for both choices of X.

TABLE 4.12
$c = .5$, $N = 400$, L^∞ *norm.*

i	$\|F\|$	R_F	$\|s\|$	R_S
0	0.1000E+01		0.1000E+01	
1	0.2096E+00	0.2096	0.2452E+00	0.2452
2	0.5732E−02	0.0273	0.6150E−02	0.0251
3	0.1185E−03	0.0207	0.1414E−03	0.0230
4	0.4977E−05	0.0420	0.6209E−05	0.0439
5	0.9533E−08	0.0019		

TABLE 4.13
$c = .999$, $N = 400$, L^2 *norm.*

i	$\|F\|$	R_F	$\|s\|$	R_S
0	0.1000E+01		0.1000E+01	
1	0.3741E+00	0.3741	0.5739E+00	0.5739
2	0.1312E+00	0.3505	0.2710E+00	0.4722
3	0.4326E−01	0.3299	0.1234E+00	0.4555
4	0.1355E−01	0.3132	0.5731E−01	0.4643
5	0.4602E−02	0.3397	0.3020E−01	0.5270
6	0.1355E−02	0.2945	0.1249E−01	0.4136
7	0.2418E−03	0.1784	0.2648E−02	0.2120
8	0.1312E−04	0.0543	0.1445E−03	0.0546
9	0.4830E−06	0.0368	0.5840E−05	0.0404
10	0.9656E−07	0.1999	0.1430E−05	0.2449
11	0.4757E−08	0.0493		

TABLE 4.14
$c = .999$, $N = 400$, L^∞ *norm.*

i	$\|F\|$	R_F	$\|s\|$	R_S
0	0.1000E+01		0.1000E+01	
1	0.5295E+00	0.5295	0.8122E+00	0.8122
2	0.2553E+00	0.4821	0.4905E+00	0.6039
3	0.9782E−01	0.3832	0.2463E+00	0.5021
4	0.3157E−01	0.3228	0.1155E+00	0.4689
5	0.1057E−01	0.3349	0.6048E−01	0.5238
6	0.3171E−02	0.2999	0.2511E−01	0.4151
7	0.5877E−03	0.1854	0.5357E−02	0.2134
8	0.3410E−04	0.0580	0.2957E−03	0.0552
9	0.8726E−06	0.0256	0.1148E−04	0.0388
10	0.1933E−06	0.2216	0.2836E−05	0.2470
11	0.9520E−08	0.0492		

TABLE 4.15
Modified Bratu problem, $d = 0$, $h = 1/16$, L^∞ norm.

i	$\|F\|$	R_F	$\|s\|$	R_S
0	0.8490E+00		0.8490E+00	
1	0.6914E−01	0.0814	0.7507E−01	0.0884
2	0.1870E−02	0.0271	0.1972E−02	0.0263
3	0.1895E−05	0.0010		

TABLE 4.16
Modified Bratu problem, $d = 0$, $h = 1/512$, L^∞ norm.

i	$\|F\|$	R_F	$\|s\|$	R_S
0	0.8491E+00		0.8491E+00	
1	0.6886E−01	0.0811	0.7475E−01	0.0880
2	0.1855E−02	0.0269	0.1955E−02	0.0262
3	0.1885E−05	0.0010	0.2003E−05	0.0010
4	0.1055E−07	0.0056		

TABLE 4.17
Modified Bratu problem, $d = 1$, $h = 1/16$, L^∞ norm.

i	$\|F\|$	R_F	$\|s\|$	R_S
0	0.8498E+00		0.8498E+00	
1	0.9978E−01	0.1174	0.1090E+00	0.1282
2	0.1132E−01	0.1134	0.1111E−01	0.1020
3	0.7290E−03	0.0644	0.7771E−03	0.0699
4	0.2384E−04	0.0327		

TABLE 4.18
Modified Bratu problem, $d = 1$, $h = 1/512$, L^∞ norm.

i	$\|F\|$	R_F	$\|s\|$	R_S
0	0.8538E+00		0.8538E+00	
1	0.1002E+00	0.1173	0.1094E+00	0.1281
2	0.1151E−01	0.1149	0.1134E−01	0.1037
3	0.7568E−03	0.0658	0.8022E−03	0.0707
4	0.2572E−04	0.0340	0.2639E−04	0.0329
5	0.1646E−05	0.0640	0.1723E−05	0.0653
6	0.9397E−07	0.0571		

Finally, we consider a modified Bratu problem

$$\nabla^2 u + du_x + e^u = 0$$

on $\Omega = (0,1) \times (0,1)$ with homogeneous Dirichlet boundary conditions. As was the case for the linear elliptic boundary value problems discussed above, the functional analytic setting depends on whether $d = 0$ or $d \neq 0$. Moreover, it is crucial that $X \subset L^\infty$ in order that the exponential nonlinearity be defined as a map on X. If $B_0 = \nabla^2$ the compactness condition $B_0^{-1}E_0 \in \mathcal{COM}(H, X)$ is satisfied with $H = L^2$ and $X = L^\infty$ or $X = C(\bar{\Omega})$ if $d = 0$. If $d \neq 0$ one can still take $X = L^\infty$ or $X = C(\bar{\Omega})$ but must use $H = H^\alpha$ for any $\alpha > 0$. As we did for the linear elliptic problem, we present results for two cases: $d = 0$ in Tables 4.15 and 4.16 and $d = 1$ in Tables 4.17 and 4.18. To facilitate comparison we use the discrete L^2 inner product for both tabulations rather than a discrete analog of the H^α inner product in the $d = 1$ case. We report the iterations for $h = 1/16$ and $h = 1/512$. The initial iterate was $u_0 = \sin(\pi x)\sin(\pi y)$.

Acknowledgments. The authors would like to thank Jim Rulla of North Carolina State University and Ekkehard Sachs of the Universität Trier in the Federal Republic of Germany for many very helpful discussions, and the referees for some very useful suggestions.

REFERENCES

[1] P. M. ANSELONE, *Collectively Compact Operator Approximation Theory*, Prentice–Hall, Englewood Cliffs, NJ, 1971.

[2] K. E. ATKINSON, *Iterative variants of the Nyström method for the numerical solution of integral equations*, Numer. Math., 22 (1973), pp. 17–31.

[3] C. G. BROYDEN, J. E. DENNIS, AND J. J. MORÉ, *On the local and superlinear convergence of quasi-Newton methods*, J. Inst. Math. Appl., 12 (1973), pp. 223–246.

[4] W. BURMEISTER, *Zur Konvergenz einiger Verfahren der konjugierten Richtungen*, in Proc. Internationaler Kongreß über Anwendung der Mathematik in dem Ingenieurwissenschaften, Weimar, Germany, 1975.

[5] S. CHANDRASEKHAR, *Radiative Transfer*, Dover, New York, 1960.

[6] D. W. DECKER AND C. T. KELLEY, *Broyden's method for a class of problems having singular Jacobian at the root*, SIAM J. Numer. Anal., 22 (1985), pp. 566–574.

[7] J. E. DENNIS AND J. J. MORÉ, *A characterization of superlinear convergence and its application to quasi-Newton methods*, Math. Comp., 28 (1974), pp. 549–560.

[8] J. E. DENNIS AND R. B. SCHNABEL, *Numerical Methods for Nonlinear Equations and Unconstrained Optimization*, Prentice–Hall, Englewood Cliffs, NJ, 1983.

[9] J. E. DENNIS AND H. F. WALKER, *Convergence theorems for least change secant update methods*, SIAM J. Numer. Anal., 18 (1981), pp. 949–987.

[10] M. ENGELMAN, G. STRANG, AND K. J. BATHE, *The application of quasi-Newton methods in fluid mechanics*, Internat. J. Numer. Methods Engrg., 17 (1981), pp. 707–718.

[11] Z. FORTUNA, *Some convergence properties of the conjugate gradient method in Hilbert space*, SIAM J. Numer. Anal., 16 (1979), pp. 380–384.

[12] D. M. GAY, *Some convergence properties of Broyden's method*, SIAM J. Numer. Anal., 16 (1979), pp. 623–630.

[13] R. R. GERBER AND F. T. LUK, *A generalized Broyden's method for solving simultaneous linear equations*, SIAM J. Numer. Anal., 18 (1981), pp. 882–890.

[14] D. GILBARG AND N. TRUDINGER, *Elliptic Partial Differential Equations of Second Order*, Springer-Verlag, Berlin, 1977.

[15] A. GRIEWANK, *The superlinear convergence of secant methods on mildly nonlinear problems in Hilbert space*, SMU Tech. Report, Department of Mathematics, Southern Methodist University, Dallas, TX, 1974.

[16] ———, *Rates of convergence for secant methods on nonlinear problems in Hilbert space*, in Numerical Analysis, Proceedings Guanajuato, Mexico, 1984, J. P. Hennart, ed., Lecture Notes in Mathematics 1230, Springer-Verlag, Heidelberg, 1986, pp. 138–157.

[17] A. Griewank and P. L. Toint, *Partitioned variable metric methods for large sparse optimization problems*, Numer. Math., 39 (1982), pp. 119–137.

[18] W. A. Gruver and E. Sachs, *Algorithmic Methods In Optimal Control*, Pitman, London, 1980.

[19] L. V. Kantorovich and V. I. Krylov, *Approximate Methods of Higher Analysis*, P. Noordhoff, Groningen, the Netherlands, 1964.

[20] C. T. Kelley, *Operator prolongation methods for nonlinear equations*, in Computational Solution of Nonlinear Systems of Equations, E. L. Allgower and K. Georg, eds., AMS Lectures in Applied Mathematics, Vol. 26, American Mathematical Society, Providence, RI, 1990, pp. 359–388.

[21] C. T. Kelley and J. I. Northrup, *Pointwise quasi-Newton methods and some applications*, in Distributed Parameter Systems, F. Kappel, K. Kunisch, and W. Schappacher, eds., Springer-Verlag, New York, 1987, pp. 167–180.

[22] ———, *A pointwise quasi-Newton method for integral equations*, SIAM J. Numer. Anal., 25 (1988), pp. 1138–1155.

[23] C. T. Kelley and E. W. Sachs, *Broyden's method for approximate solution of nonlinear integral equations*, J. Integral Equations, 9 (1985), pp. 25–44.

[24] ———, *A quasi-Newton method for elliptic boundary value problems*, SIAM J. Numer. Anal., 24 (1987), pp. 516–531.

[25] ———, *A pointwise quasi-Newton method for unconstrained optimal control problems*, Numer. Math., 55 (1989), pp. 159–176.

[26] ———, *Fast algorithms for compact fixed point problems with inexact function evaluations*, SIAM J. Sci. Statist. Comput., 12 (1991), pp. 725–742.

[27] ———, *A new proof of superlinear convergence for Broyden's method in Hilbert space*, SIAM J. Optimization, 1 (1991), pp. 146–150.

[28] J. J. Moré and J. A. Trangenstein, *On the global convergence of Broyden's method*, Math. Comp., 30 (1976), pp. 523–540.

[29] E. Sachs, *Convergence rates of quasi-Newton methods for some nonsmooth optimization problems*, SIAM J. Control Optim., 23 (1985), pp. 401–418.

[30] L. K. Schubert, *Modification of a quasi-Newton method for nonlinear equations with sparse Jacobian*, Math. Comp., 24 (1970), pp. 27–30.

[31] J. Stoer, *Two examples on the convergence of certain rank-2 minimization methods for quadratic functionals in Hilbert space*, Linear Algebra Appl., 28 (1979), pp. 217–222.

[32] P. N. Swarztrauber and R. A. Sweet, *Efficient* FORTRAN *subprograms for the solution of elliptic partial differential equations*, ACM Trans. Math. Software, 5 (1979), pp. 352–364.

SIAM J. OPTIMIZATION
Vol. 2, No. 4, pp. 533–557, November 1992

ON THE BEHAVIOR OF BROYDEN'S CLASS OF QUASI-NEWTON METHODS*

RICHARD H. BYRD[†], DONG C. LIU[‡], AND JORGE NOCEDAL[‡]

Abstract. This paper analyzes algorithms from the Broyden class of quasi-Newton methods for nonlinear unconstrained optimization. This class depends on a parameter ϕ_k, for which the choices $\phi_k = 0$ and $\phi_k = 1$ give the well-known BFGS and DFP methods. This paper examines algorithms that allow for negative values of the parameter ϕ_k. It shows that severe restrictions have to be imposed on the selection of ϕ_k to guarantee q-superlinear convergence. It is argued that negative values of ϕ_k are desirable, and conditions on ϕ_k that guarantee superlinear convergence are given. However, practical algorithms that preserve the excellent properties of the BFGS method are not easy to design.

Key words. quasi-Newton method, Broyden class, variable metric method, global convergence, superlinear convergence

AMS(MOS) subject classifications. 65, 49

1. Introduction.

An important class of quasi-Newton methods for solving the unconstrained optimization problem,

$$\min_{x\in\mathbf{R}^n} f(x), \tag{1.1}$$

was proposed by Broyden (1967). It consists of iterations of the form

$$x_{k+1} = x_k + \alpha_k d_k, \qquad k \ge 1, \tag{1.2}$$

where

$$d_k = -B_k^{-1} g_k. \tag{1.3}$$

Here α_k is a step length parameter and g_k denotes the gradient of f at x_k. The Hessian approximation B_k is updated by means of the formula

$$B_{k+1} = B_k - \frac{B_k s_k s_k^T B_k}{s_k^T B_k s_k} + \frac{y_k y_k^T}{y_k^T s_k} + \phi_k (s_k^T B_k s_k) v_k v_k^T, \tag{1.4}$$

where ϕ_k is a scalar, $y_k = g_{k+1} - g_k$, $s_k = x_{k+1} - x_k$, and

$$v_k = \left[\frac{y_k}{y_k^T s_k} - \frac{B_k s_k}{s_k^T B_k s_k}\right]. \tag{1.5}$$

The choice of the parameter ϕ_k is important, since it can greatly affect the performance of the methods. The BFGS method corresponds to $\phi_k = 0$, and the DFP method

* Received by the editors June 1, 1990; accepted for publication (in revised form) August 12, 1991.

[†] Department of Computer Science, University of Colorado, Boulder, Colorado 80309. The work of this author was supported by National Science Foundation grant CCR-8702403, Air Force Office of Scientific Research grant 85-0251, and Army Research Office contract DAAL03-88-K-0086.

[‡] Department of Electrical Engineering and Computer Science, Northwestern University, Evanston, Illinois 60208. The work of these authors was supported by the Applied Mathematical Sciences subprogram of the Office of Energy Research, U.S. Department of Energy, under grant DE-FG02-87ER2547-A001, and by National Science Foundation grant CCR-8902096.

corresponds to $\phi_k = 1$. The class of methods with $\phi_k \in [0, 1]$ is frequently called the *restricted* Broyden class.

Many theoretical and experimental studies of algorithms belonging to Broyden's class have been published; see Dennis and Moré (1974), (1977); Stoer (1975); Powell (1976), (1986); Schnabel (1978); Werner (1978), (1989); Ritter (1979), (1981); Stachursky (1981); Griewank and Toint (1982); Byrd, Nocedal, and Yuan (1987); and the references therein. Attention has been focused on the DFP and BFGS methods and on the restricted Broyden class. Fletcher (1970) proved that the eigenvalues of the matrix $(\nabla^2 f)^{-1} B_k$ tend monotonically to 1 when $\phi_k \in [0, 1]$ and when f is quadratic. Moreover, Fletcher showed that this property does not necessarily hold for methods outside the restricted class. This result has sometimes been interpreted as indicating that the restricted Broyden class contains the most useful methods within the Broyden family. Most numerical experience to date favors using the BFGS method. In addition, until recently the strongest convergence results could be proved only for the restricted Broyden class excluding DFP (see Powell (1976) and Byrd, Nocedal, and Yuan (1987)).

In this paper we examine a class of update formulas that is larger than the restricted Broyden class in that it allows for negative values of ϕ_k. Some algorithms of this kind have been studied by Zhang and Tewarson (1988), who establish global convergence on convex problems and present encouraging numerical results. Their work, however, leaves some important questions unanswered; in particular, it is not clear whether such strategies are superlinearly convergent. We will see that this is indeed a delicate issue and that for some apparently reasonable choices of $\phi_k \leq 0$, q-superlinear convergence cannot be guaranteed. Nevertheless, we argue that there are good reasons for allowing negative values of ϕ_k, and we therefore investigate conditions that will ensure global and superlinear convergence.

The motivation for this work, as well as many of the techniques used in our analysis, derives from the results of Byrd, Nocedal, and Yuan (1987). We will now review the main ideas of that work, paying particular attention to the effect of negative values of ϕ_k.

Byrd, Nocedal, and Yuan show that global convergence on a convex function is obtained for any starting point x_1 and any symmetric and positive definite starting matrix B_1, if $\phi_k \in [0, 1)$ is bounded away from 1. They assume that the step length satisfies the Wolfe conditions

$$f(x_k + \alpha_k d_k) \leq f(x_k) + \beta_1 \alpha_k g_k^T d_k, \tag{1.6}$$

$$g(x_k + \alpha_k d_k)^T d_k \geq \beta_2 g_k^T d_k, \tag{1.7}$$

where $0 < \beta_1 < \frac{1}{2}$ and $\beta_1 < \beta_2 < 1$. They also show that the rate of convergence is q-superlinear. These results can be extended in various ways: Byrd and Nocedal (1989) and Werner (1978), (1989) show that a larger class of line searches can be used; Byrd and Xie (1990) weaken the restriction that ϕ_k be bounded away from 1, so as to prove global convergence for updates, such as the Hoshino update.

Byrd, Nocedal, and Yuan consider the effect of different choices of $\phi_k \in [0, 1]$ and argue that if ϕ_k is close to 1, the algorithm can be very inefficient because it loses its ability to correct large eigenvalues in the Hessian approximation. Let us be more precise. From (1.4) we have

$$\operatorname{tr}(B_{k+1}) = \operatorname{tr}(B_k) + \frac{\|y_k\|^2}{y_k^T s_k} + \phi_k \frac{\|y_k\|^2}{y_k^T s_k} \frac{s_k^T B_k s_k}{y_k^T s_k} \tag{1.8}$$

$$-(1-\phi_k)\frac{\|B_k s_k\|^2}{s_k^T B_k s_k} - 2\phi_k \frac{y_k^T B_k s_k}{y_k^T s_k},$$

where $\operatorname{tr}(B_k)$ denotes the trace of B_k and $\|\cdot\|$ denotes the l_2 vector norm. If $\phi_k \in [0,1)$, the only term that is guaranteed to be negative is the fourth term on the right-hand side of (1.8). Thus, if B_k has large eigenvalues we must rely on this term to reduce the trace of B_k. When $\phi_k \simeq 1$ this term is weakened and convergence can be very slow if B_k contains excessively large eigenvalues. Note that if ϕ_k is negative, the fourth term in the right-hand side of (1.8) remains negative and increases in magnitude and the third term becomes negative. Thus, when $\phi_k < 0$, the algorithm is more able to correct large eigenvalues.

Negative values of ϕ_k may not be desirable, however, because the Hessian approximation may become indefinite, singular, or nearly singular. Computing the determinant (cf. Pearson (1969)) of (1.4) we obtain

$$\begin{aligned}\det(B_{k+1}) &= \det(B_k)\left[(1-\phi_k)\frac{y_k^T s_k}{s_k^T B_k s_k} + \phi_k \frac{y_k^T B_k^{-1} y_k}{y_k^T s_k}\right] \\ &= \det(B_k)\frac{y_k^T s_k}{s_k^T B_k s_k}[1+\phi_k(\mu_k - 1)], \qquad (1.9)\end{aligned}$$

where

$$\mu_k = \frac{(y_k^T B_k^{-1} y_k)(s_k^T B_k s_k)}{(y_k^T s_k)^2}. \qquad (1.10)$$

We see from (1.9) that $\det(B_{k+1}) = 0$ when ϕ_k has the *critical value*

$$\phi_k^c \equiv \frac{1}{1-\mu_k}. \qquad (1.11)$$

It is well known that if the initial Hessian approximation B_1 is symmetric and positive definite and if at each step $s_k^T y_k > 0$ and $\phi_k > \phi_k^c$, then all the matrices B_k remain symmetric and positive definite (cf. Fletcher (1987)). By applying the Cauchy–Schwarz inequality to (1.10) we see that

$$\mu_k \geq 1,$$

and therefore $\phi_k^c \leq 0$. Moreover, we will show later that if the iterates converge q-superlinearly and $\{B_k^{-1}\}$ is bounded, then $\phi_k^c \to -\infty$. Therefore, there appears to be plenty of room for choosing negative values of ϕ_k.

Clearly, ϕ_k should not be too close to ϕ_k^c to avoid generating nearly singular Hessian approximations. There is, however, another reason for avoiding small values of ϕ_k. We see from (1.9) that when $s_k^T B_k s_k$, the estimated curvature, is small compared to the average curvature, $y_k^T s_k$, the determinant increases, thus increasing some of the eigenvalues of B_k. Thus, when $\phi_k \geq 0$, Broyden's update formula has a strong self-correcting property with respect to the determinant. However, from (1.9) we see that this property is diminished if $[1+\phi_k(\mu_k - 1)]$ is small.

We conclude from the previous discussion that, in order to efficiently correct large eigenvalues, ϕ_k should be as small as possible, but to cope well with small eigenvalues one should ensure that ϕ_k is not too close to ϕ_k^c. This suggests that the best choice, in general, might be $\phi_k = 0$, i.e., the BFGS method. However, results of Powell (1986) and our numerical experience indicate that BFGS suffers more from large eigenvalues than from small ones. We illustrate this in Table 1, which shows the number of iterations needed by the BFGS method to solve Watson's function (Moré, Garbow, and Hillstrom (1981)) for various choices of B_1. The starting point is $x_0 = (0,0,0,0,0)$, the run was stopped when $\|g_k\| \le 10^{-8}$, and the initial matrix is

$$\text{diag }(1,1,1,\lambda^{\frac{1}{2}},\lambda),$$

where λ is a variable parameter.

TABLE 1
Number of iterations required by BFGS.

λ	10^{-6}	10^{-4}	1	10^{4}	10^{6}
Number of iterations	23	23	25	40	50

We therefore ask whether, to deal more efficiently with large eigenvalues, one must allow for negative values of ϕ_k. Zhang and Tewarson (1988) performed numerical tests with fixed negative values of ϕ_k, and their results show a moderate but consistent improvement over the BFGS method. They also prove that, for convex problems, global and linear convergence can be established for negative values of ϕ_k, provided that for all k,

$$(1-\nu)\phi_k^c \le \phi_k \le 0, \tag{1.12}$$

where ν is an arbitrary constant in (0,1).

However, as will be shown in §2, satisfaction of (1.12) will not always give q-superlinear convergence, and therefore other criteria are needed for choosing negative values of ϕ_k. By applying the perturbation results of Griewank and Toint (1982), which use the Frobenius norm, one can show that negative values of ϕ_k exist that will give superlinear convergence (see Zhang and Tewarson (1988, Thm. 3.3)). These values, however, are not computable in practice, and it is not clear how close they must be to zero.

In this paper we use the trace and determinant equations (1.8) and (1.9) to study whether it is possible to design superlinearly convergent algorithms that use negative values of ϕ_k. We first present in §2 some negative results showing that, for a large class of functions of two variables, q-superlinear convergence can occur only if $\liminf \phi_k/(-\phi_k^c) \ge 0$. We also establish some limitations in the case where ϕ_k is a negative constant. In §3 we establish some positive results giving sufficient conditions on ϕ_k for superlinear convergence. These results imply, for any problem, the existence at each iteration of an interval $[\xi_k, 1-\delta]$ of admissible values of ϕ_k for which superlinear convergence occurs and such that $\{\xi_k\}$ is negative and bounded away from zero. The results also imply that convergence is superlinear whenever the sequence $\{\phi_k\}$ satisfies $\sum_{k\ge 1} \phi_k/\phi_k^c < \infty$. Finally, in §4 we present some numerical experiments with two methods that allow for negative values of ϕ_k: the optimally conditioned method proposed by Davidon (1975) and a particular idealized optimal choice of ϕ_k that performs better than the BFGS method.

2. Loss of superlinear convergence. We now investigate what properties of the sequence $\{\phi_k\}$ are necessary for superlinear convergence and show that for some apparently reasonable choices of ϕ_k superlinear convergence may not occur. Since our point of view in this section is essentially negative, it is appropriate to restrict our attention to a special class of problems; thus we consider in detail the case of a two-dimensional quadratic objective function. The analysis of this section is built on the work of Powell (1986), and we will quote several results from that paper.

We consider the function

$$f(x) = \tfrac{1}{2}\left(x_{[1]}^2 + x_{[2]}^2\right), \tag{2.1}$$

where $x_{[i]}$ is the ith component of x. Since quasi-Newton methods from the Broyden class are invariant under a linear change of variables, considering (2.1) is equivalent to considering any strictly convex two-dimensional quadratic. In this section we assume that $\phi_k \le 1$ and write it in the form

$$\phi_k = \phi_k^c(1-\nu_k), \qquad 0 < \nu_k \le 1. \tag{2.2}$$

Superlinear convergence of quasi-Newton methods is usually proved by showing that the search directions approximate the Newton directions, which implies that the step length of one is eventually accepted for all iterates. Since in this section we will establish necessary conditions on ϕ_k for superlinear convergence, we assume that $\alpha_k = 1$ is always chosen by the algorithm, so that the iteration is

$$x_{k+1} = x_k - B_k^{-1}x_k. \tag{2.3}$$

Because of the secant equation,

$$B_{k+1}(x_{k+1}-x_k) = x_{k+1}-x_k, \tag{2.4}$$

we see that B_k has one unit eigenvalue for all $k > 1$, and we let λ_k denote the other eigenvalue. As suggested by Powell, the analysis can be done using only two scalars, λ_k and Θ_k, where

$$\cos\Theta_k = \frac{x_k^T(x_k - x_{k-1})}{\|x_k\|\,\|x_k - x_{k-1}\|}. \tag{2.5}$$

Thus Θ_k is the angle between x_k and the eigenvector of B_k corresponding to the unit eigenvalue. Let $B_k = Q_kD_kQ_k^T$, where $D_k = \operatorname{diag}(1,\lambda_k)$ is the matrix of eigenvalues and Q_k is the orthogonal matrix of eigenvectors. From the definition of Θ_k we have

$$\frac{Q_k^Tx_k}{\|x_k\|} = \begin{pmatrix} \cos\Theta_k \\ \sin\Theta_k \end{pmatrix}. \tag{2.6}$$

We write (2.3) as

$$x_{k+1} = x_k - Q_kD_k^{-1}Q_k^Tx_k, \tag{2.7}$$

and thus

$$\begin{aligned}\frac{Q_k^Tx_{k+1}}{\|x_k\|} &= \frac{Q_k^Tx_k}{\|x_k\|} - \begin{pmatrix} 1 & 0 \\ 0 & \lambda_k^{-1} \end{pmatrix}\frac{Q_k^Tx_k}{\|x_k\|} \\ &= \begin{pmatrix} 0 \\ (1-\lambda_k^{-1})\sin\Theta_k \end{pmatrix}.\end{aligned}$$

Taking norms,

$$\frac{\|x_{k+1}\|^2}{\|x_k\|^2} = \frac{(\lambda_k - 1)^2}{\lambda_k^2}\sin^2\Theta_k. \tag{2.8}$$

Powell (1986) considered only the BFGS and DFP update formulas. However, we note that his equation (2.3), which states that

$$\tan^2 \Theta_{k+1} \tan^2 \Theta_k = \lambda_k^2, \tag{2.9}$$

holds for any update formula that generates symmetric and positive definite matrices. From (2.2), (1.9), and (1.11) we have

$$\det(B_{k+1}) = \det(B_k) \frac{y_k^T s_k}{s_k^T B_k s_k} \nu_k.$$

Since B_k has one unit eigenvalue for all $k > 1$, and noting that $\nu_k = 1$ corresponds to the BFGS method, we have

$$\begin{aligned} \lambda_{k+1} &= \det(B_{k+1}) \\ &= \det(B_{k+1}^{\text{BFGS}})\nu_k \\ &= \lambda_{k+1}^{\text{BFGS}} \nu_k. \end{aligned} \tag{2.10}$$

Here B_{k+1}^{BFGS} is the matrix obtained by updating B_k with the BFGS formula, and $\lambda_{k+1}^{\text{BFGS}}$ is its non-unit eigenvalue. An expression for $\lambda_{k+1}^{\text{BFGS}}$ in terms of Θ_k and λ_k is given by Powell (1986, eq. (2.2)). Substituting this value in (2.10) gives

$$\lambda_{k+1} = \frac{\tan^2 \Theta_k + \lambda_k^2}{\tan^2 \Theta_k + \lambda_k} \nu_k. \tag{2.11}$$

In addition, we can derive a simple expression for the critical value of ϕ in this two-dimensional case. Since by (2.6) and (2.7),

$$s_k = -Q_k \begin{pmatrix} \cos \Theta_k \\ \lambda_k^{-1} \sin \Theta_k \end{pmatrix} \|x_k\|,$$

it follows after some algebraic manipulations using (1.10), (1.11), and the fact that $y_k = s_k$, that the critical value ϕ_k^c is given by

$$\phi_k^c = -\frac{(\lambda_k^2 + \tan^2 \Theta_k)^2}{\lambda_k (1 - \lambda_k)^2 \tan^2 \Theta_k}. \tag{2.12}$$

Our first result involves the ratio of ϕ_k and $-\phi_k^c$ (note that ϕ_k^c is always negative). It shows that if a subsequence of $\{\phi_k/(-\phi_k^c)\}$ is bounded below zero, then the superlinear convergence property is lost (recall that $\phi_k = 0$ gives BFGS). This of course implies that the choice

$$\phi_k = \phi_k^c(1 - \nu), \qquad \nu \in (0, 1),$$

which ensures r-linear convergence on convex functions, will not guarantee q-superlinear convergence.

THEOREM 2.1. *Let algorithm (1.2)–(1.4), with $\alpha_k = 1$ and $\phi_k \le 1$, be applied to a strictly convex two-dimensional quadratic objective function with any initial x_1 and positive definite B_1. Assume that the solution is not obtained in a finite number of steps. Then, the algorithm converges q-superlinearly only if* $\liminf_{k \to \infty} \phi_k/(-\phi_k^c) \ge 0$.

Proof. Suppose that the iterates converge superlinearly to $x_* = 0$. In (2.2) ν_k is defined so that $\phi_k/\phi_k^c = 1 - \nu_k$. Our result follows if we can show that $\limsup(1 - \nu_k) \le$

0, i.e., that $\liminf \nu_k \geq 1$. We will show this by contradiction. Suppose that there is an infinite subsequence

$$\{\nu_{k_j}\} \tag{2.13}$$

of $\{\nu_k\}$ such that

$$\nu_{k_j} \leq \hat{\nu} < 1, \qquad j = 1, 2, \ldots. \tag{2.14}$$

Since the algorithm does not terminate in a finite number of steps, by (2.8) it must be the case that for all k, $\lambda_k \neq 1$. We first show that (2.14) implies that $\lambda_k < 1$ for all sufficiently large k. We will use Fletcher's monotonicity result (Fletcher (1970)), which says that if $\phi_k \in [0, 1]$,

$$\lambda_k \geq 1 \;\Rightarrow\; 1 \leq \lambda_{k+1} \leq \lambda_k \tag{2.15}$$

and

$$\lambda_k \leq 1 \;\Rightarrow\; \lambda_k \leq \lambda_{k+1} \leq 1. \tag{2.16}$$

On the other hand, (2.11) implies that if $\phi_k \leq 0$, so that $\nu_k \leq 1$, then

$$\lambda_k \geq 1 \;\Rightarrow\; \lambda_{k+1} \leq \lambda_k \nu_k \leq \lambda_k \tag{2.17}$$

and

$$\lambda_k < 1 \;\Rightarrow\; \lambda_{k+1} < 1. \tag{2.18}$$

Now consider the possibility that $\lambda_k > 1$ for all k. Then, (2.15) when $\phi_k \in [0, 1]$ and (2.17) when $\phi_k \leq 0$ imply that $\{\lambda_k\}$ is monotone decreasing. Furthermore, for all iterates of the infinite subsequence (2.13) satisfying (2.14) we have that $\phi_k < 0$, and the first inequality of (2.17) implies that $\lambda_{k_j+1} \leq \hat{\nu}\lambda_{k_j}$. Thus, eventually $\lambda_m < 1$ for some m, which contradicts the assumption that $\lambda_k > 1$ for all k. This contradiction implies that $\lambda_k < 1$ for some k, in which case it then follows from (2.18) when $\phi_k \leq 0$ and from (2.16) when $\phi_k \in [0, 1]$ that $\lambda_{k+l} < 1$ for $l = 1, 2, \ldots$. Thus, the subsequence (2.13) can be redefined so that $\lambda_{k_j} < 1$ for all j.

This fact and (2.11) imply that for all j

$$\lambda_{k_j+1} < \nu_{k_j} \leq \hat{\nu}. \tag{2.19}$$

The superlinear convergence assumption, together with (2.8) and (2.19), implies that

$$\sin^2 \Theta_{k_j+1} \to 0 \tag{2.20}$$

and then that

$$\frac{\tan^2 \Theta_{k_j+1}}{{\lambda_{k_j+1}}^2} \to 0, \tag{2.21}$$

$$\frac{\tan^2 \Theta_{k_j+1}}{\lambda_{k_j+1}} \to 0. \tag{2.22}$$

From (2.2), (2.12), and the fact that $\phi_k \leq 1$ it follows that

$$\nu_{k_j+1} \leq 1 + \frac{\lambda_{k_j+1}(1-\lambda_{k_j+1})^2 \tan^2 \Theta_{k_j+1}}{({\lambda_{k_j+1}}^2 + \tan^2 \Theta_{k_j+1})^2},$$

and therefore

$$\nu_{k_j+1}\lambda_{k_j+1} \leq \lambda_{k_j+1} + \frac{\lambda_{k_j+1}^{-2}(1-\lambda_{k_j+1})^2 \tan^2 \Theta_{k_j+1}}{(1+\lambda_{k_j+1}^{-2} \tan^2 \Theta_{k_j+1})^2}. \tag{2.23}$$

By (2.21) the second term on the right side of (2.23) converges to zero and therefore

$$\limsup_{j\to\infty}(\nu_{k_j+1}\lambda_{k_j+1}) \leq \limsup_{j\to\infty} \lambda_{k_j+1}. \tag{2.24}$$

Equation (2.11) may be expressed as

$$\lambda_{k_j+2} = \nu_{k_j+1}\lambda_{k_j+1}\left(\frac{\tan^2\Theta_{k_j+1}}{{\lambda_{k_j+1}}^2}+1\right) \bigg/ \left(\frac{\tan^2\Theta_{k_j+1}}{\lambda_{k_j+1}}+1\right).$$

By (2.21), (2.22), (2.24), and (2.19) this implies that

$$\limsup_{j\to\infty} \lambda_{k_j+2} \leq \limsup_{j\to\infty} \lambda_{k_j+1} \leq \hat{\nu}.$$

From this relation, from the superlinear convergence assumption, and from (2.8) we have

$$\sin^2 \Theta_{k_j+2} \to 0. \tag{2.25}$$

Finally, from (2.8), (2.9), (2.19), (2.20), and (2.25)

$$\begin{aligned}\frac{\|x_{k_j+2}\|^2}{\|x_{k_j+1}\|^2} &= \frac{(\lambda_{k_j+1}-1)^2 \sin^2\Theta_{k_j+1}}{\tan^2\Theta_{k_j+2}\tan^2\Theta_{k_j+1}} \\ &= \frac{(\lambda_{k_j+1}-1)^2 \cos^2\Theta_{k_j+1}}{\tan^2\Theta_{k_j+2}} \to \infty,\end{aligned}$$

which contradicts the assumption of superlinear convergence. □

Note that an immediate consequence of this result is that if ϕ_k is nonpositive for all k, then $\phi_k/\phi_k^c \to 0$. This result is for two-dimensional quadratics; however, since a two-dimensional problem can be embedded in a larger space, it follows that for many larger problems, superlinear convergence with nonpositive ϕ_k requires that $\phi_k/\phi_k^c \to 0$.

It is interesting to consider the choice

$$\phi_k = \phi_k^c(1-\nu), \qquad 0 < \nu < 1, \tag{2.26}$$

where ν is a constant, a bit more closely. Theorem 2.1 shows that this choice cannot give q-superlinear convergence. In fact, for any value of $\nu \in (0,1)$, there exist initial values x_1 and B_1 for which the iterates converge exactly linearly to the solution. Specifically, if we set

$$\lambda_1 = \frac{\nu}{2-\nu}, \qquad \tan\Theta_1 = \sqrt{\lambda_1}, \tag{2.27}$$

then it follows from (2.9) and (2.11) that

$$\lambda_k = \lambda_1 \quad \text{and} \quad \tan\Theta_k = \tan\Theta_1, \tag{2.28}$$

for all $k > 1$. We conclude by (2.8) that, if $\nu > \frac{1}{2}$ the sequence $\{x_k\}$ converges to the solution with a q-linear rate of $(1-\nu)\sqrt{2/\nu}$. (This is assuming that step lengths of one are used, which will happen if β_1 in (1.6) is small enough that $\nu > \frac{1}{2}$ causes (1.6) to be satisfied with step length one. If $\nu \le \frac{1}{2}$ then a line search procedure would cause different behavior.) To further understand the behavior of the iteration, we express it in the form

$$(\lambda_{k+1}, \tan\Theta_{k+1}) = S(\lambda_k, \tan\Theta_k),$$

where $S : \mathbf{R}^2 \to \mathbf{R}^2$ is defined by (2.9) and (2.11) with $\nu_k = \nu$. Differentiating S and evaluating the derivative at the fixed point given by (2.27) yields

$$S'\left(\frac{\nu}{2-\nu}, \sqrt{\frac{\nu}{2-\nu}}\right) = \begin{bmatrix} \nu - \frac{1}{2} & (1-\nu)\sqrt{\frac{\nu}{2-\nu}} \\ \sqrt{\frac{2-\nu}{\nu}} & -1 \end{bmatrix}. \tag{2.29}$$

This matrix has the characteristic polynomial $\chi(\mu) = (\nu - \frac{1}{2} - \mu)(-1-\mu) - (1-\nu)$. Since $\chi(-1) = \nu - 1 < 0$ for $\nu \in (0,1)$, and $\chi(\mu) > 0$ for μ sufficiently negative, it follows that, at the fixed point, S' has an eigenvalue less than -1. (The other eigenvalue is in $(0,1)$.) Thus, $(\nu/(2-\nu), \sqrt{\nu/(2-\nu)})$ is not a stable fixed point, which indicates that for most starting points the iteration will not converge to it. Therefore, the linear convergence mentioned should not occur often. Experiments with this iteration using various values of ν bore this out. When the iteration was started close to this fixed point, or at almost any other starting point, it tended to converge to a two-cycle where for even-numbered iterates

$$\lambda_{2k} \to \nu, \qquad \tan\Theta_{2k} \to 0, \tag{2.30}$$

and for odd-numbered iterates

$$\lambda_{2k+1} \to \nu^2, \qquad \tan\Theta_{2k+1} \to \infty.$$

For such a sequence, (2.8) and (2.30) imply that if step lengths of one are used, convergence is two-step superlinear, that is,

$$\frac{\|x_{k+2}\|}{\|x_k\|} \to 0.$$

This behavior seems to indicate that, although the algorithm does not converge q-superlinearly when ν_k is constant, its performance still might be quite good in many cases.

Another interesting issue is the rate of convergence when ϕ_k is set equal to a negative constant. As mentioned in §2, if superlinear convergence occurs, $\phi_k^c \to -\infty$. Thus setting ϕ_k to a negative constant does not necessarily violate the condition $\phi_k/\phi_k^c \to 0$, which we have shown is necessary for superlinear convergence. Zhang and Tewarson (1988) experimented with such a choice with good results. However, one can show that in the two-dimensional quadratic case, if the constant is less than -1, then there are initial values for which superlinear convergence does not occur even though (1.12) may hold for some value of $\nu \in (0,1)$.

To demonstrate this we suppose $\phi_k = -C$ for all k, where $C > 0$ is a constant, and consider the resulting map given by (2.9) and (2.11). If $C > 1$ a fixed point of this map is given by

$$\lambda_k = \frac{C-1}{C+1}, \qquad \tan\Theta_k = \sqrt{\frac{C-1}{C+1}}. \tag{2.31}$$

To see this, note that if λ_k and $\tan\Theta_k$ are given by (2.31), it follows from (2.12) that $\phi_k^c = -C^2$. By (2.2) this implies that $\nu_k = 1 - 1/C$. Then if λ_k and $\tan\Theta_k$ are given by (2.31), it follows from (2.9) and (2.11) that $\lambda_{k+1} = \lambda_k$ and $\tan\Theta_{k+1} = \tan\Theta_k$, so that we have a fixed point (in fact the same one we considered in the constant ν_k iteration). From (2.8) we see that the sequence $\{x_k\}$ converges linearly with rate $\sqrt{2/C(C-1)}$. This example definitely indicates that superlinear convergence cannot be proved when ϕ_k is a constant less than -1.

3. Obtaining superlinear convergence. Although the results of the previous section are mainly negative, we now show that there are strategies for choosing negative values of ϕ_k that give rise to superlinear convergence. We begin by introducing some notation and by stating the assumptions we make about the objective function.

The matrix of second derivatives of f is denoted by G, the starting point for the algorithm is x_1, and we define the level set $D = \{x \in \mathbf{R}^n : f(x) \leq f(x_1)\}$. Throughout the paper $\|\cdot\|$ denotes the l_2 vector norm or its induced matrix norm.

ASSUMPTIONS 3.1.

(1) The objective function f is twice continuously differentiable.

(2) The level set D is convex, and there exist positive constants m and M such that

$$m\|z\|^2 \leq z^T G(x) z \leq M\|z\|^2 \tag{3.1}$$

for all $z \in \mathbf{R}^n$ and all $x \in D$. Note that this implies that f has a unique minimizer x_* in D.

(3) The Hessian matrix G is Lipschitz continuous at x_*; i.e., there exists a positive constant L such that

$$\|G(x) - G(x_*)\| \leq L\|x - x_*\| \tag{3.2}$$

for all x in a neighborhood of x_*.

Let us first consider the question of global convergence. Powell (1976) showed that the BFGS method is globally convergent for convex functions, and Zhang and Tewarson (1988) extended his result to negative values of ϕ_k that satisfy

$$(1-\nu)\phi_k^c \leq \phi_k \leq 0, \tag{3.3}$$

where ν is a number in (0,1). Zhang and Tewarson also showed that if f is uniformly convex, the rate of convergence is r-linear. Note that (3.3) does not imply that $\{\phi_k\}$ is bounded below because $\{\phi_k^c\}$ may diverge to $-\infty$. In fact, if the algorithm converges superlinearly and B_k^{-1} is bounded, then $\{\phi_k^c\} \to -\infty$. (This will be shown in the proof of Theorem 3.5.)

It is not hard to generalize the result of Zhang and Tewarson slightly so that it allows for both positive and negative values of ϕ_k. In the following result we do this under the assumption that ϕ_k is bounded away from 1 (recall that $\phi_k = 1$ corresponds to the DFP method). Such a restriction is necessary as global convergence for the DFP is still an open question.

THEOREM 3.1. *Let x_1 be a starting point for which f satisfies Assumptions* 3.1. *Then, for any positive definite B_1, algorithm* (1.2)–(1.4) *with a line search satisfying the Wolfe conditions* (1.6) *and* (1.7), *and with*

$$(1-\nu)\phi_k^c \le \phi_k \le 1-\delta, \quad \delta > 0, \quad \nu \in (0,1), \tag{3.4}$$

generates iterates that converge to x_*. Moreover, there is a constant $0 \le c < 1$ such that

$$f_{k+1} - f_* \le c^k[f_1 - f_*] \tag{3.5}$$

for all k and

$$\sum_{k=0}^{\infty} \|x_{k+1} - x_*\| < \infty. \tag{3.6}$$

Proof. It is a simple extension of the analysis given by Byrd, Nocedal, and Yuan (1987). That paper, which we refer to as BNY, considers $\phi_k \in [0, 1-\delta]$. However, from (1.4) it is clear that the trace of B_{k+1} is a monotone function of ϕ_k, and therefore all the inequalities in that paper involving the trace still hold. Specifically, equations (3.2) and (3.7) of BNY hold. For the determinant we see that (3.4), (1.9), and (1.11) imply

$$\det(B_{k+1}) \ge \det(B_k)\frac{y_k^T s_k}{s_k^T B_k s_k}\nu,$$

which differs from equation (3.9) of BNY only in multiplicative constant ν. It is then straightforward to see that the proofs of Lemma 3.2, Theorem 3.1, and Lemmas 4.1 and 4.2 of BNY still go through under the condition (3.4). Equation (3.5) then follows from Lemma 4.2 of BNY. Now note that the uniform convexity assumption (3.1) implies that

$$f_{k+1} - f_* \ge \frac{m}{2}\|x_{k+1} - x_*\|^2.$$

Therefore, by using (3.5),

$$\sum_{k=0}^{\infty} \|x_{k+1} - x_*\| \le \left(\frac{2[f_1 - f_*]}{m}\right)^{\frac{1}{2}} \sum_{k=0}^{\infty} c^{k/2} < \infty. \qquad \Box$$

Since the results of §2 show that (3.4) will not guarantee superlinear convergence, we need to look for more restrictive conditions on ϕ_k. Hereafter we will assume that ϕ_k belongs to the subinterval given by (3.4), and therefore by Theorem 3.1 we can assume that the iterates converge r-linearly to the solution. To simplify the analysis that follows, we define the scaled quantities

$$\tilde{s_k} = G_*^{1/2} s_k, \qquad \tilde{y_k} = G_*^{-1/2} y_k, \tag{3.7}$$

and

$$\tilde{B_k} = G_*^{-1/2} B_k G_*^{-1/2}, \tag{3.8}$$

where $G_* = G(x_*)$. We also define

$$q_k = \frac{\tilde{s}_k{}^T \tilde{B}_k \tilde{s}_k}{\tilde{s}_k{}^T \tilde{s}_k}, \tag{3.9}$$

$$\cos\theta_k = \frac{\tilde{s}_k{}^T \tilde{B}_k \tilde{s}_k}{\|\tilde{s}_k\| \, \|\tilde{B}_k \tilde{s}_k\|}, \tag{3.10}$$

which are used to characterize superlinear convergence, as the following lemma shows.

LEMMA 3.2. *Suppose that the sequence of iterates* $\{x_k\}$ *is generated by* (1.2) *and* (1.3), *using some positive definite sequence* $\{B_k\}$, *and that* $\alpha_k = 1$ *whenever that value satisfies the Wolfe conditions. If* $x_k \to x_*$, *then the following two conditions are equivalent:*

(i) *The step length* $\alpha_k = 1$ *satisfies the Wolfe conditions* (1.6) *and* (1.7) *for all large* k, *and the rate of convergence is superlinear.*

(ii)

$$\lim_{k\to\infty} \cos\theta_k = \lim_{k\to\infty} q_k = 1. \tag{3.11}$$

Proof. From (3.7)–(3.10)

$$\begin{aligned} \frac{\|G_*^{-1/2}(B_k - G_*)s_k\|^2}{\|G_*^{1/2} s_k\|^2} &= \frac{\|(\tilde{B}_k - I)\tilde{s}_k\|^2}{\|\tilde{s}_k\|^2} \\ &= \frac{\|\tilde{B}_k \tilde{s}_k\|^2 - 2\tilde{s}_k^T \tilde{B}_k \tilde{s}_k + \tilde{s}_k^T \tilde{s}_k}{\tilde{s}_k^T \tilde{s}_k} \\ &= \frac{q_k^2}{\cos^2\theta_k} - 2q_k + 1. \end{aligned} \tag{3.12}$$

Suppose that (3.11) holds. Then we conclude from (3.12) that

$$\lim_{k\to\infty} \frac{\|(B_k - G_*)s_k\|}{\|s_k\|} = 0. \tag{3.13}$$

A result of Dennis and Moré (1977) shows that the unit step length is accepted for all large k, and the Dennis and Moré (1974) characterization result shows that the rate of convergence is superlinear. Conversely, if $\alpha_k = 1$ for all large k, and if the rate of convergence is superlinear, the Dennis and Moré characterization implies that (3.13) holds, and thus the right-hand side of (3.12) converges to zero. Since this quantity is greater than or equal to

$$q_k^2 - 2q_k + 1,$$

we see that $q_k \to 1$, which in turn implies $\cos\theta_k \to 1$. □

We now analyze the behavior of the scaled Hessian approximations $\tilde{B}_k$. From (1.4), (3.7), and (3.8) we have

$$\tilde{B}_{k+1} = \tilde{B}_k - \frac{\tilde{B}_k \tilde{s}_k \tilde{s}_k{}^T \tilde{B}_k}{\tilde{s}_k{}^T \tilde{B}_k \tilde{s}_k} + \frac{\tilde{y}_k \tilde{y}_k{}^T}{\tilde{y}_k{}^T \tilde{s}_k} + \phi_k(\tilde{s}_k{}^T \tilde{B}_k \tilde{s}_k)\tilde{v}_k \tilde{v}_k^T, \tag{3.14}$$

where

$$\tilde{v}_k = \left[\frac{\tilde{y}_k}{\tilde{y}_k{}^T \tilde{s}_k} - \frac{\tilde{B}_k \tilde{s}_k}{\tilde{s}_k{}^T \tilde{B}_k \tilde{s}_k} \right]. \tag{3.15}$$

Computing the trace of (3.14) and using (3.9) and (3.10), we obtain

$$\text{(3.16)} \qquad \operatorname{tr}(\tilde{B}_{k+1}) = \operatorname{tr}(\tilde{B}_k) + \frac{\|\tilde{y}_k\|^2}{\tilde{y_k}^T\tilde{s_k}} - \frac{q_k}{\cos^2\theta_k} + \phi_k(\tilde{s}_k^T\tilde{B}_k\tilde{s}_k)\tilde{v}_k^T\tilde{v}_k.$$

To obtain the determinant we first observe that μ_k, defined by (1.10), is unchanged when s_k, y_k, and B_k are replaced by $\tilde{s}_k, \tilde{y}_k$, and $\tilde{B}_k$. Then, using (1.9) and (3.9) we have

$$\text{(3.17)} \qquad \det(\tilde{B}_{k+1}) = \det(\tilde{B}_k)\frac{1}{q_k}\frac{\tilde{y_k}^T\tilde{s_k}}{\tilde{s_k}^T\tilde{s_k}}[1 + \phi_k(\mu_k - 1)].$$

To measure the goodness of the matrix B_k, we will use the matrix function ψ,

$$\text{(3.18)} \qquad \psi(B_k) = \operatorname{tr}(B_k) - \ln\det(B_k),$$

whose properties are discussed in more detail by Byrd and Nocedal (1989); see also Fletcher (1989). We only note here that

$$\text{(3.19)} \qquad \psi(B) = \sum_{i=1}^{n}[\lambda_i(B) - \ln(\lambda_i(B))],$$

where $\lambda_i(B)$ are the eigenvalues of B, and thus $\psi(B_k) > 0$ if B_k is positive definite. From (3.16) and (3.17) we have

$$\begin{aligned} \psi(\tilde{B}_{k+1}) &= \psi(\tilde{B}_k) + \frac{\|\tilde{y}_k\|^2}{\tilde{y_k}^T\tilde{s_k}} - \ln\frac{\tilde{y_k}^T\tilde{s_k}}{\tilde{s_k}^T\tilde{s_k}} - \frac{q_k}{\cos^2\theta_k} + \phi_k(\tilde{s_k}^T\tilde{B}_k\tilde{s}_k)\tilde{v}_k^T\tilde{v}_k \\ \text{(3.20)} \qquad & + \ln q_k - \ln\left[1 + \phi_k(\mu_k - 1)\right]. \end{aligned}$$

By (1.11) and (3.4) we have that

$$\text{(3.21)} \qquad 1 + \phi_k(\mu_k - 1) = 1 - \frac{\phi_k}{\phi_k^c} \geq \nu,$$

and thus the last term in (3.20) is well defined. We now estimate the second and third terms on the right side of (3.20), assuming that x_k is sufficiently near x_*. First note that

$$\begin{aligned} y_k &= \left[\int_0^1 G\left(x_k + \tau s_k\right) d\tau\right] s_k \\ &= G_* s_k + \left[\int_0^1 \left(G(x_k + \tau s_k) - G_*\right) d\tau\right] s_k. \end{aligned}$$

Let us define

$$\text{(3.22)} \qquad \epsilon_k = L\max\{\|x_{k+1} - x_*\|, \|x_k - x_*\|\},$$

where L is the Lipschitz constant given by (3.2). It follows that

$$\begin{aligned} \tilde{y}_k &= \left[G_*^{-1/2}\int_0^1 \left(G(x_k + \tau s_k) - G_*\right) d\tau G_*^{-1/2}\right]\tilde{s_k} + \tilde{s_k} \\ \text{(3.23)} \qquad &\equiv E_k\tilde{s_k} + \tilde{s_k}, \end{aligned}$$

where

$$\|E_k\| \leq \|G_*^{-1}\|\epsilon_k. \tag{3.24}$$

We now have from (3.23) and (3.24)

$$\frac{\|\tilde{y}_k\|^2}{\tilde{y_k}^T\tilde{s_k}} = 1 + O(\epsilon_k), \tag{3.25}$$

$$m_k \equiv \frac{\tilde{y_k}^T\tilde{s_k}}{\tilde{s_k}^T\tilde{s_k}} = 1 + O(\epsilon_k), \tag{3.26}$$

and thus

$$\ln m_k = O(\epsilon_k). \tag{3.27}$$

We also have from (3.23), (3.24), and (3.26) that

$$\begin{aligned}
\frac{\tilde{y}_k^T\tilde{B}_k\tilde{s}_k}{\tilde{y}_k^T\tilde{s}_k} &= \frac{q_k}{m_k} + O\left(\frac{\|\tilde{s}_k\|\|\tilde{B}_k\tilde{s}_k\|}{\tilde{y_k}^T\tilde{s_k}}\epsilon_k\right) \\
&= \frac{q_k}{m_k} + O\left(\frac{q_k\epsilon_k}{\cos\theta_k m_k}\right) \\
&= q_k\left(1 + \frac{O(\epsilon_k)}{\cos\theta_k}\right). \qquad (3.28)
\end{aligned}$$

We are ready to present the first positive result. It gives a fairly complicated sufficient condition for superlinear convergence, which limits ϕ_k when it is negative. Later we will see that it implies other simpler and more intuitive conditions.

LEMMA 3.3. *Let x_1 be a starting point for which f satisfies Assumptions* 3.1, *and let B_1 be any symmetric and positive definite starting matrix. Assume that in the algorithm* (1.2)–(1.4), *$\{\phi_k\}$ satisfies* (3.4) *and in addition,*

$$\sum_{\phi_k \leq 0} \ln \frac{\cos^{2-2r}\theta_k}{[1 + \phi_k(\mu_k - 1)][1 - \phi_k\|\tilde{s}_k\|^2\|\tilde{v}_k\|^2\cos^2\theta_k]^{(1-r)}} < \infty \tag{3.29}$$

for some $r > 0$. Assume also that the line search enforces Wolfe conditions (1.6) *and* (1.7) *and that in so doing, it always tries the step length $\alpha_k = 1$ first. Then $x_k \to x_*$ q-superlinearly, and $\{\|B_k\|\}$ and $\{\|B_k^{-1}\|\}$ are bounded.*

Proof. Since all the conditions of Theorem 3.1 are satisfied, $x_k \to x_*$ and (3.6) holds. From (3.15), (3.25), (3.26), and (3.28) we have that

$$\|\tilde{v}_k\|^2 = \frac{1}{\|\tilde{s}_k\|^2}\left[\frac{1}{m_k}\frac{\|\tilde{y}_k\|^2}{\tilde{y_k}^T\tilde{s_k}} - \frac{2}{q_k}\frac{\tilde{y}_k^T\tilde{B}_k\tilde{s}_k}{\tilde{y}_k^T\tilde{s}_k} + \frac{1}{\cos^2\theta_k}\right] \tag{3.30}$$

$$= \frac{1}{\|\tilde{s}_k\|^2}\left[\frac{1}{\cos^2\theta_k} - 1 + \frac{O(\epsilon_k)}{\cos\theta_k}\right]. \tag{3.31}$$

In order to analyze (3.20) we define

$$t_k \equiv \frac{1 - \phi_k\|\tilde{s}_k\|^2\|\tilde{v}_k\|^2\cos^2\theta_k}{\cos^2\theta_k} \tag{3.32}$$

$$= \frac{1 - \phi_k + \phi_k\cos^2\theta_k + \phi_k\cos\theta_k O(\epsilon_k)}{\cos^2\theta_k}, \tag{3.33}$$

where the last equality follows from (3.31). Note that when $\phi_k \leq 0$, (3.32) implies that $t_k \geq 1$. When $\phi_k > 0$ we see from (3.33) and from the upper bound in (3.4) that

$$t_k \geq \frac{\delta + (1-\delta)\cos\theta_k O(\epsilon_k)}{\cos^2\theta_k},$$

which was obtained by dropping the third term in the numerator of (3.33). Since $\epsilon_k \to 0$, for sufficiently large k, say $k > k_0$, we have $t_k > 0$. Thus $\ln t_k$ is well defined for such k, and by using (3.32) we can write the assumption (3.29) as

$$\sum_{\substack{k=k_0\\ \phi_k \leq 0}}^{\infty} \ln\left(\frac{1}{[1+\phi_k(\mu_k-1)]\, t_k^{(1-r)}}\right) < \infty.$$

The finiteness of this sum implies that there exists a constant c_1 such that

$$-\sum_{\substack{k=k_0\\ \phi_k \leq 0}}^{j} \ln([1+\phi_k(\mu_k-1)]\, t_k) + r\sum_{\substack{k=k_0\\ \phi_k \leq 0}}^{j} \ln t_k \leq c_1 \tag{3.34}$$

for all j. Using (3.25)–(3.27) in (3.20) and then using (3.32), we obtain, for $k > k_0$,

$$\begin{aligned}\psi(\tilde{B}_{k+1}) &= \psi(\tilde{B}_k) + \left[1 - \frac{q_k}{\cos^2\theta_k} + \ln q_k + \phi_k q_k \|\tilde{s}_k\|^2\|\tilde{v}_k\|^2\right]\\ &\quad - \ln[1+\phi_k(\mu_k-1)] + O(\epsilon_k)\\ &= \psi(\tilde{B}_k) + \eta_k - \ln t_k - \ln[1+\phi_k(\mu_k-1)] + O(\epsilon_k),\end{aligned} \tag{3.35}$$

where

$$\eta_k = [1 - q_k t_k + \ln(q_k t_k)]\,. \tag{3.36}$$

Note that $\eta_k \leq 0$ because, for any number $z > 0$, the quantity $1 - z + \ln z \leq 0$.

We now sum (3.35), using (3.34) and ignoring the two nonpositive terms

$$-\ln[1+\phi_k(\mu_k-1)] \quad \text{for } \phi_k > 0$$

and

$$-\ln t_k \quad \text{for } \phi_k \leq 0$$

to obtain

$$\begin{aligned}\psi(\tilde{B}_{j+1}) \leq &\psi(\tilde{B}_{k_0}) + \sum_{k=k_0}^{j} \eta_k + r\sum_{\substack{k=k_0\\ \phi_k \leq 0}}^{j} (-\ln t_k)\\ &- \sum_{\substack{k=k_0\\ \phi_k > 0}}^{j} \ln t_k + \sum_{k=k_0}^{j} O(\epsilon_k) + c_1.\end{aligned} \tag{3.37}$$

The first two sums are nonpositive, and the fourth sum is bounded as $j \to \infty$ by (3.6) and (3.22). We need to show that the third sum is bounded as $j \to \infty$, and to do this we estimate how small t_k can be when $\phi_k > 0$. Consider t_k as a function of $\cos\theta_k$ with $\cos\theta_k \in (0,1]$. Differentiating (3.33) and recalling that $\phi_k \leq 1-\delta$, we see that for ϵ_k sufficiently small t_k is a monotonically decreasing function of $\cos\theta_k$, and thus when $\cos\theta_k = 1$ it takes on its minimum value, $t_k = 1 + O(\epsilon_k)$. Therefore, when $\phi_k > 0$,

$$t_k < 1 \Rightarrow t_k \geq 1 + O(\epsilon_k) \Rightarrow \ln t_k \geq O(\epsilon_k), \tag{3.38}$$

and hence

$$\sum_{\substack{k=k_0 \\ \phi_k>0, t_k<1}}^{\infty} \ln t_k > -\infty, \tag{3.39}$$

due to the condition $\sum \epsilon_k < \infty$. This implies that

$$-\sum_{\substack{k=k_0 \\ \phi_k>0}}^{\infty} \ln t_k < \infty,$$

and by (3.37) we conclude that $\{\psi(\tilde{B}_j)\}$ is bounded. Therefore, using (3.19), we have that the largest eigenvalues of B_k and B_k^{-1}, i.e., $\|B_k\|$ and $\|B_k^{-1}\|$, are uniformly bounded.

To prove the superlinear convergence we consider again (3.37), neglecting negative terms. From (3.39) and the fact that $\psi(\tilde{B}_j) > 0$ for all j, we have that

$$r \sum_{\substack{k=k_0 \\ \phi_k \leq 0}}^{\infty} \ln t_k + \sum_{\substack{k=k_0 \\ \phi_k>0,\, t_k \geq 1}}^{\infty} \ln t_k < \infty. \tag{3.40}$$

Since the terms in both sums are nonnegative, they must converge to zero. From the first sum we have that $t_k \to 1$ for $\phi_k \leq 0$. From the second sum we have $t_k \to 1$ when $\phi_k > 0$ and $t_k \geq 1$. However, from (3.38), $t_k \to 1$ also when $\phi_k > 0$ and $t_k < 1$. We conclude that the whole sequence $\{t_k\}$ converges to 1. Therefore, using (3.33) we have

$$\left[1 - \phi_k + \phi_k \cos^2\theta_k + \phi_k O(\epsilon_k) - \cos^2\theta_k\right] \to 0,$$

and thus by (3.4)

$$\left[(1 - \cos^2\theta_k) + \frac{\phi_k}{1-\phi_k} O(\epsilon_k)\right] \to 0.$$

Since $1 - \phi_k > \delta$, the last term inside the square brackets converges to zero, and we have

$$\cos\theta_k \to 1. \tag{3.41}$$

From the fact that the last three sums of (3.37) are bounded, it also follows that

$$\sum_{k \geq k_0} (-\eta_k) < \infty,$$

which by (3.36) implies that

$$[1 - q_k t_k + \ln(q_k t_k)] \to 0.$$

Since $\{t_k\}$ converges to 1, we conclude that

$$q_k \to 1. \tag{3.42}$$

Lemma 3.2, (3.41) and (3.42) imply superlinear convergence. □

Inequality (3.29) is rather complex but in essence gives a condition on how fast ϕ_k/ϕ_k^c must converge to zero. It is the most general relation we have been able to obtain, but, unfortunately, it involves quantities that depend on G_* and are therefore not available during the course of the computation. Nevertheless, (3.29) allows us to derive some interesting and simpler conditions.

THEOREM 3.4. *Suppose that the conditions of Lemma* 3.3 *hold except that* (3.29) *is replaced by the assumption*

$$\sum_{\phi_k \le 0} \frac{\phi_k}{\phi_k^c} < \infty. \tag{3.43}$$

Then $x_k \to x_*$ q-superlinearly.

Proof. Let $0 < r < 1$ in (3.29). Then from (1.11) we have that, for $\phi_k \le 0$,

$$\begin{aligned}\frac{\cos^{2-2r}\theta_k}{[1+\phi_k(\mu_k-1)][1-\phi_k\|\tilde{s}_k\|^2\|\tilde{v}_k\|^2\cos^2\theta_k]^{(1-r)}} &\le \frac{1}{1+\phi_k(\mu_k-1)}\\ &= \left[1-\frac{\phi_k}{\phi_k^c}\right]^{-1}.\end{aligned}$$

Using the mean value theorem and (3.21), we have

$$\ln\left[1-\frac{\phi_k}{\phi_k^c}\right]^{-1} = -\ln\left[1-\frac{\phi_k}{\phi_k^c}\right] \le \frac{\phi_k/\phi_k^c}{1-\phi_k/\phi_k^c} \le \frac{1}{\nu}\frac{\phi_k}{\phi_k^c}.$$

Therefore, (3.43) implies (3.29) with $0 < r < 1$. □

It is interesting to compare this result with those of §2. Theorem 2.1 indicates that for a large class of problems, if $\phi_k \le 1$, then it is necessary that $\phi_k/\phi_k^c \to 0$ converge to zero for negative ϕ_k to obtain q-superlinear convergence. Theorem 3.4 shows that it is sufficient for q-superlinear convergence that this same sequence be summable.

It is possible to enforce the condition (3.43) in a practical algorithm. For example, one can always choose ϕ_k to satisfy, in addition to (3.4),

$$\frac{\phi_k}{\phi_k^c} \le \gamma_0\|g_k\|, \tag{3.44}$$

where γ_0 is a constant. Since $\|g_k\| = O(\epsilon_k)$, it is clear that this strategy ensures (3.43).

However, enforcing (3.44) may result in an algorithm that is very close to BFGS, and thus it is natural to ask whether it is possible to choose values ϕ_k that are bounded away from zero but still guarantee superlinear convergence. The following result gives an affirmative answer to this question.

THEOREM 3.5. *Consider algorithm* (1.2)–(1.4), *where the line search satisfies the Wolfe conditions* (1.6) *and* (1.7) *and always tries the step length* $\alpha_k = 1$ *first. If Assumptions* 3.1 *are satisfied for* x_1 *and if* B_1 *is symmetric and positive definite, then for each* k *there exists* $\xi_k < 0$ *such that* $\sup\{\xi_k\} < 0$ *and such that if*

$$\phi_k \in [\xi_k, 1-\delta], \qquad \delta \in (0,1) \tag{3.45}$$

for all k, *then the iterates converge to* x_* *q-superlinearly.*

Proof. Suppose that ϕ_k is chosen to satisfy (3.4), so that Theorem 3.1 holds and $\sum \epsilon_k < \infty$. Suppose in addition that, for all large k, ϕ_k satisfies

$$\phi_k \geq [1 - (1-\gamma\epsilon_k)\cos^{2-2r}\theta_k]\phi_k^c, \tag{3.46}$$

where $r \in (0, \frac{1}{2})$ and $\gamma > 0$ are arbitrary constants. By (1.11) this is equivalent to

$$1 + \phi_k(\mu_k - 1) \geq \cos^{2-2r}\theta_k(1-\gamma\epsilon_k).$$

Substituting this into the left side of (3.29) and noting that for $\phi_k \leq 0$

$$[1 - \phi_k\|\tilde{s}_k\|^2\|\tilde{v}_k\|^2\cos^2\theta_k]^{(1-r)} \geq 1,$$

we see that the left-hand side in (3.29) is less than or equal to

$$-\sum_{\phi_k \leq 0} \ln(1-\gamma\epsilon_k).$$

Since this sum is finite, (3.29) holds. Therefore, by Lemma 3.3, $\{x_k\}$ converges q-superlinearly and $\{\|B_k\|\}$ and $\{\|B_k^{-1}\|\}$ are bounded. Also, (3.41) and (3.42) hold.

Note that we are assuming that both (3.4) and (3.46) hold. This can be written in the form (3.45) if we define

$$\xi_k = \max\left\{\frac{1-(1-\gamma\epsilon_k)\cos^{2-2r}\theta_k}{1-\mu_k}, (1-\nu)\phi_k^c\right\}. \tag{3.47}$$

Thus we have proved that superlinear convergence is obtained for this choice of ξ_k. To complete the proof we must show that $\sup\{\xi_k\} < 0$. To show that the first term inside the curly brackets is bounded away from zero, we define

$$w_k = \tilde{B}_k\tilde{s}_k - q_k\tilde{s}_k. \tag{3.48}$$

Note that $w_k^T\tilde{s}_k = 0$, so

$$\|\tilde{B}_k\tilde{s}_k\|^2 = q_k^2\|\tilde{s}_k\|^2 + \|w_k\|^2 \tag{3.49}$$

and

$$\cos^2\theta_k = \frac{q_k^2\|\tilde{s}_k\|^2}{q_k^2\|\tilde{s}_k\|^2 + \|w_k\|^2}. \tag{3.50}$$

Since $\{q_k\}$ and $\{\cos\theta_k\}$ converge to 1, this relation implies

$$\frac{\|w_k\|^2}{\|\tilde{s}_k\|^2} \to 0. \tag{3.51}$$

We now estimate the term inside the square brackets in (3.46). Since $r \in (0, \frac{1}{2})$,

$$\cos^2\theta_k \le \cos^{2-2r}\theta_k \le \cos\theta_k,$$

and thus, using (3.50), we have for large k

$$\begin{aligned} 1-(1-\gamma\epsilon_k)\cos^{2-2r}\theta_k &\ge 1-\cos\theta_k+\gamma\epsilon_k\cos^2\theta_k \\ &\ge \tfrac{1}{2}(1-\cos^2\theta_k)+\tfrac{1}{2}(1-2\cos\theta_k+\cos^2\theta_k)+\tfrac{1}{2}\gamma\epsilon_k\cos^2\theta_k \\ &\ge \tfrac{1}{2}(1-\cos^2\theta_k+\gamma\epsilon_k\cos^2\theta_k) \\ &= \frac{1}{2}\frac{\|w_k^2\|+\gamma\epsilon_k q_k^2\|\tilde{s}_k\|^2}{q_k^2\|\tilde{s}_k\|^2+\|w_k\|^2}. \end{aligned} \tag{3.52}$$

To consider μ_k we note that, since $\{\|B_k^{-1}\|\}$ is bounded, (3.23) and (3.24) give

$$\tilde{y}_k{}^T\tilde{B}_k{}^{-1}\tilde{y}_k = \tilde{s}_k{}^T\tilde{B}_k{}^{-1}\tilde{s}_k + O(\epsilon_k\|\tilde{s}_k\|^2). \tag{3.53}$$

By (3.48), $\tilde{B}_k{}^{-1}\tilde{s}_k = (\tilde{s}_k - \tilde{B}_k{}^{-1}w_k)/q_k$. Applying this relation twice and recalling that $w_k^T\tilde{s}_k = 0$, we obtain

$$\tilde{s}_k{}^T\tilde{B}_k{}^{-1}\tilde{s}_k = \frac{\|\tilde{s}_k\|^2}{q_k} + \frac{w_k^T\tilde{B}_k{}^{-1}w_k}{q_k^2}. \tag{3.54}$$

Therefore, by recalling the scale invariance of μ_k, and using (1.10), (3.26), (3.53), (3.54), and the boundedness of $\{\|B_k\|\}$ and $\{\|B_k^{-1}\|\}$,

$$\begin{aligned} \mu_k - 1 &= \frac{(\tilde{s}_k{}^T\tilde{B}_k\tilde{s}_k)(\tilde{y}_k{}^T\tilde{B}_k{}^{-1}\tilde{y}_k)}{(\tilde{y}_k{}^T\tilde{s}_k)^2} - 1 \\ &= \frac{\tilde{s}_k{}^T\tilde{B}_k\tilde{s}_k}{\|\tilde{s}_k\|^4(1+O(\epsilon_k))}\left(\frac{\|\tilde{s}_k\|^2}{q_k} + \frac{w_k^T\tilde{B}_k{}^{-1}w_k}{q_k^2} + O(\epsilon_k\|\tilde{s}_k\|^2)\right) - 1 \\ &= \frac{w_k^T\tilde{B}_k{}^{-1}w_k}{q_k\|\tilde{s}_k\|^2} + O(\epsilon_k). \end{aligned} \tag{3.55}$$

Using this relation and (3.52), we see that for large k there is a constant γ' such that

$$\begin{aligned} \frac{1-(1-\gamma\epsilon_k)\cos^{2-2r}\theta_k}{\mu_k - 1} &\ge \frac{1}{2}\frac{(\|w_k\|^2+\gamma\epsilon_k q_k^2\|\tilde{s}_k\|^2)q_k\|\tilde{s}_k\|^2}{(q_k^2\|\tilde{s}_k\|^2+\|w_k\|^2)(w_k^T\tilde{B}_k{}^{-1}w_k + q_k\|\tilde{s}_k\|^2\gamma'\epsilon_k)} \\ &\ge \frac{1}{2}\frac{q_k\|w_k\|^2+\gamma\epsilon_k q_k^3\|\tilde{s}_k\|^2}{(q_k^2+\frac{\|w_k\|^2}{\|\tilde{s}_k\|^2})(\|\tilde{B}_k{}^{-1}\|\|w_k\|^2 + q_k\|\tilde{s}_k\|^2\gamma'\epsilon_k)} \\ &\ge \frac{1}{2}\left(q_k^2 + \frac{\|w_k\|^2}{\|\tilde{s}_k\|^2}\right)^{-1}\min\left\{\frac{q_k}{\|\tilde{B}_k{}^{-1}\|+\gamma'}, \frac{\gamma q_k^2}{\|\tilde{B}_k{}^{-1}\|+\gamma'}\right\}, \end{aligned}$$

where in the last step we have considered the cases $\|w_k\|^2 < q_k\|\tilde{s}_k\|^2\epsilon_k$ and $\|w_k\|^2 \ge q_k\|\tilde{s}_k\|^2\epsilon_k$ separately. Using (3.51) and the fact that $q_k \to 1$, we have that

$$\begin{aligned} &\liminf_{k\to\infty}\frac{1-(1-\gamma\epsilon_k)\cos^{2-2r}\theta_k}{\mu_k - 1} \\ &\qquad \ge \liminf_{k\to\infty}\tfrac{1}{2}\min\left\{\frac{1}{\|\tilde{B}_k{}^{-1}\|+\gamma'}, \frac{\gamma}{\|\tilde{B}_k{}^{-1}\|+\gamma'}\right\}. \end{aligned} \tag{3.56}$$

Since $\{\|\tilde{B}_k{}^{-1}\}\|$ is bounded, the right side of (3.56) is positive.

Now note that ξ_k can only be zero when $\epsilon_k = 0$, in which case the algorithm terminates. By (3.51), (3.55), and the boundedness of $\{\|\tilde{B}_k{}^{-1}\|\}$, we have that

$$\phi_k^c \to -\infty.$$

Therefore, by (3.47) and (3.56), $\sup\{\xi_k\} < 0$. □

This result does not tell us how to compute the bounds ξ_k in practice, since ξ_k is a function of the quantities $\|\tilde{B}_k{}^{-1}\|$ and q_k, which depend on the exact Hessian. However, the result does show that it is possible to keep $\phi_k < 0$ away from zero and obtain q-superlinear convergence.

4. Numerical experiments. In the previous two sections we considered negative values of the parameter ϕ_k that allow the good properties of the BFGS method to be preserved. Now we will investigate experimentally whether there are negative choices for this parameter that will yield an improvement over the BFGS.

One possible criterion for choosing ϕ_k involves the matrix function ψ, which, as we saw in §3, is a useful measure of the goodness of the Hessian approximation. It is natural to ask what the value of ϕ_k that optimizes $\psi(\tilde{B}_{k+1})$ is, and whether this value can be used in a practical algorithm. From (3.20) we see that $\psi(\tilde{B}_{k+1})$ is a strictly convex function of ϕ_k and has a unique minimizer. Differentiating (3.20) with respect to ϕ_k and noting from (3.7) and (3.8) that $\tilde{s}_k{}^T\tilde{B}_k\tilde{s}_k = s_k^T B_k s_k$, we see that the minimizer is

$$\phi_k^* = \frac{1}{(s_k^T B_k s_k)\tilde{v}_k^T\tilde{v}_k} + \frac{1}{1-\mu_k}. \tag{4.1}$$

However, knowledge of G_* is required in $\tilde{v}_k$, thus making this formula impossible to use in a practical algorithm. One could estimate $\tilde{v}_k^T\tilde{v}_k = v_k^T G_*^{-1} v_k$ by using information available during the iteration, or one could try to estimate ϕ_k^* directly by balancing the trace and determinant equations. We have experimented with several heuristic formulas along these lines, and the results appear to be rather satisfactory. However, since the convergence results of this paper do not apply to these strategies and since the numerical results are not conclusive, we will not present the results here.

Instead, we will take the view that (4.1) is likely to be superior to any of our heuristics, and we will experiment with it. Even though this is not a practical algorithm, because of the knowledge of G_* that is required, such experiments may indicate how much of an improvement over BFGS can be obtained in the ideal case.

Since the last term in (4.1) is ϕ_k^c, we see that $\phi_k^* > \phi_k^c$. In fact, it can be seen that ϕ_k^* guarantees global and superlinear convergence on uniformly convex functions. This is true since the analysis of the BFGS method by Byrd and Nocedal (1989) is based on upper bounding recursions involving the function ψ and thus applies immediately to any update that, at each iteration, gives a lower value of $\psi(\tilde{B}_{k+1})$ than that given by the BFGS formula. Note that ϕ_k^* can be of either sign.

To implement method (4.1) we replace $\tilde{v}_k^T\tilde{v}_k$ by

$$v_k^T G_k^{-1} v_k, \tag{4.2}$$

where v_k is given by (1.5) and G_k is the Hessian of f at x_k. We therefore asymptotically obtain (4.1). Since this formula will cause difficulties when G_k is not positive definite, we include the following safeguards: (i) if $v_k^T G_k^{-1} v_k \leq 0$, then we set $\phi_k = 0$ (the BFGS value); otherwise, (ii) if $\phi_k < 0.95\phi_k^c$, then we set $\phi_k = 0.95\phi_k^c$.

We have also experimented with the optimally conditioned formula of Davidon (1975). In this method B_{k+1} is chosen to be the member of the Broyden class that minimizes the condition number of $B_k^{-1}B_{k+1}$, subject to preserving positive definiteness. The resulting value of ϕ_k sometimes lies outside [0,1], which makes Davidon's formula relevant to our study. The value of ϕ_k is given by

$$(4.3)\quad \phi_k^{\mathrm{D}} = \begin{cases} \dfrac{y_k^T s_k(y_k^T B_k^{-1} y_k - y_k^T s_k)}{(s_k^T B_k s_k)(y_k^T B_k^{-1} y_k) - (y_k^T s_k)^2} & \text{if } y_k^T s_k \leq \dfrac{2(s_k^T B_k s_k)(y_k^T B_k^{-1} y_k)}{s_k^T B_k s_k + y_k^T B_k^{-1} y_k}, \\ \dfrac{y_k^T s_k}{y_k^T s_k - s_k^T B_k s_k} & \text{otherwise.} \end{cases}$$

Davidon's update sometimes coincides with the symmetric rank-one formula (SR1), but it has the advantage that ϕ_k^{D} is always greater than ϕ_k^c, ensuring positive definiteness of $\{B_k\}$. Several other properties of this method are discussed by Schnabel (1978). We have not been able to apply the convergence results of this paper to Davidon's formula, and we do not know whether it is globally or superlinearly convergent. Our investigation of this method will be entirely numerical.

We now list the methods used in our tests; they differ only in the choice of ϕ_k.

1. *BFGS.* The BFGS method ($\phi_k = 0$).
2. *Davidon.* Davidon's method (4.3).
3. *Method I.* The method given by (4.1), where $\tilde{v}_k^T \tilde{v}_k$ is replaced by (4.2).

In all methods we used a line search routine written by Moré, which enforces a strong form of the Wolfe conditions: In addition to (1.6) it ensures that

$$|g(x_k + \alpha_k d_k)^T d_k| \leq \beta_2 |g_k^T d_k|.$$

We used the values $\beta_1 = 10^{-4}$, $\beta_2 = 0.9$. The algorithms were stopped when $\|g_k\| \leq 10^{-7}$. All the runs were made on a SUN 3/60 in double precision.

A technique suggested by several authors (cf. Luenberger (1984)) is to scale the matrices B_k at every iteration to try to alleviate ill conditioning. Shanno and Phua (1978) recommend scaling only once—at the end of the first iteration. We tested this strategy, in which before updating B_1 we multiply it by

$$\frac{y_1^T B_1^{-1} y_1}{y_1^T s_1}.$$

However, in our experiments this strategy did not improve the performance of the methods for any of the starting matrices. It helped in some problems but was detrimental in others; overall, it performed similarly to the unscaled method. Therefore, we will report only the results without scaling.

The first test problem is an extension to $n = 4$ of the function studied in Byrd, Nocedal, and Yuan (1987). It is given by

$$(4.4)\qquad f(x) = \tfrac{1}{2}x^T x + 0.25\sigma(x^T A x)^2,$$

where σ is a parameter and

$$A = \begin{pmatrix} 5 & 1 & 0 & 0.5 \\ 1 & 4 & 0.5 & 0 \\ 0 & 0.5 & 3 & 0 \\ 0.5 & 0 & 0 & 2 \end{pmatrix}.$$

This is a strictly convex function that allows us to control the deviation from a quadratic by means of the parameter σ. The starting point is

$$x_1 = (\cos 70°, \sin 70°, \cos 70°, \sin 70°)^T.$$

The initial Hessian approximation B_1 was always chosen to be diagonal. We used three types of matrices B_1, with large, small, and moderate eigenvalues, respectively. They are

$$B_1^l = \text{diag}\,(1, \ldots, 1, 10, 10^4), \tag{4.5}$$

$$B_1^s = \text{diag}\,(1, \ldots, 1, 10^{-1}, 10^{-4}), \tag{4.6}$$

$$B_1^m = \text{diag}\,(0.1, 0.5, 2, 10). \tag{4.7}$$

The results, using $\sigma = 1$ in (4.4), are given in Table 2. The first number represents iterations, and the second the number of function evaluations. For Method I we indicate in parentheses the number of times that the safeguard (ii) was employed (safeguard (i) was never active since the function is strictly convex).

TABLE 2
Results for the perturbed quadratic function.

	Starting matrix		
Method	Large	Small	Moderate
BFGS	48/53	28/47	28/36
Davidon	40/45	23/42	23/30
Method I	27/43 (8)	18/33 (0)	16/22 (1)

Similar results were obtained for $\sigma = 0.1$. An examination of the computations showed that in the vast majority of the iterations of Davidon's method and Method I, the value of ϕ_k was negative. For Method I this value was often near ϕ_k^c. Overall, Method I gives a substantial reduction of function evaluations and iterations compared to the BFGS method. Table 2 also shows that Davidon's method performs very well on this problem.

TABLE 3
Large eigenvalues in B_1.

Prob	N	BFGS	Davidon	Method I
1	4	246/322	181/223	107/179 (31)
4	5	50/60	44/55	29/50 (11)
6	3	32/34	27/29	16/22 (2)
7	6	73/83	66/77	55/85 (27)
8	3	3/10	3/10	3/8 (1)
9	2	151/198	182/222	121/143 (3)
10	3	28/36	27/35	18/29 (15)
11	10	42/52	40/50	20/37 (9)
12	2	14/27	12/20	13/27 (2)
15	5	43/46	42/45	25/46 (15)
16	8	173/220	163/186	154/248 (96)
17	4	62/70	49/58	45/66 (10)
18	2	22/31	23/31	12/25 (4)
19	4	79/105	89/109	60/96 (20)

To see if these observations generalize to other objective functions we tried 14 problems from the collection of Moré, Garbow, and Hillstrom (1981). The starting points were obtained by setting factor = 1 in their routines. We first present, in

TABLE 4
Small eigenvalues in B_1.

Prob	N	BFGS	Davidon	Method I
1	4	143/194	156/196	107/201 (40)
4	5	23/43	23/44	18/35 (0)
6	3	25/38	27/39	23/35 (0)
7	6	74/89	58/70	209/371 (84)
8	3	5/10	5/9	4/7 (0)
9	2	F	F	F
10	3	21/30	21/26	12/16 (0)
11	10	23/42	21/38	18/34 (0)
12	2	F	13/21	F
15	5	20/34	19/33	19/34 (1)
16	8	98/144	125/158	84/152 (24)
17	4	46/66	35/55	35/55 (3)
18	2	15/38	15/38	12/29 (0)
19	4	56/89	62/98	83/141 (25)

TABLE 5
Moderate eigenvalues in B_1.

Prob	N	BFGS	Davidon	Method I
1	4	154/205	190/236	122/164 (12)
4	5	25/35	23/33	19/26 (3)
6	3	29/38	31/37	25/35 (4)
7	6	37/48	37/42	33/43 (10)
8	3	3/5	3/5	3/5 (0)
9	2	150/192	185/226	120/136 (0)
10	3	29/37	30/38	16/25 (4)
11	10	19/29	19/29	F
12	2	11/29	12/16	11/29 (1)
15	5	20/22	20/24	17/20 (3)
16	8	84/120	92/111	66/119 (21)
17	4	48/58	32/45	34/44 (2)
18	2	14/19	11/17	11/19 (0)
19	4	53/81	52/74	47/84 (11)

Tables 3 and 4, the results with large and small eigenvalues in the initial matrix, defining B_1 by (4.5) and (4.6). Once more the results are presented in the form (number of iterations)/(number of function evaluations). For Method I we indicate in parentheses the number of times that any of the safeguards was used.

We verified that the three methods converged to the same solution point (problems for which this was not the case were not included in our test set). F denotes a failure, which was caused by (i) a line search error due to a badly scaled search direction or (ii) a failure to obtain the solution to the prescribed accuracy. The BFGS and Davidon matrices never suffered from a loss of positive definiteness and used no safeguarding. Method I used the safeguards described in the paragraph following the formula (4.2). These safeguards were often used; the total number of times that safeguard (i) or (ii) was applied is indicated in parentheses next to the number of function evaluations for Method I. We see from Tables 3 and 4 that Davidon's method performs better than the BFGS method for large eigenvalues but is only slightly better for small eigenvalues. Method I is substantially better than BFGS in the large eigenvalue case, but its advantage is less pronounced for small eigenvalues.

To observe the effect of moderate eigenvalues, we use $B_1 = I$, the identity matrix (as opposed to (4.7)). The results are given in Table 5.

TABLE 6
Percentage of iterations for which $\phi_k < 0$.

Prob	N	Davidon			Method I		
		Large	Small	Moderate	Large	Small	Moderate
1	4	50	43	28	78	68	45
4	5	73	39	65	79	67	68
6	3	70	71	48	6	56	68
7	6	42	60	51	64	67	61
8	3	67	0	67	100	0	67
9	2	37	F	36	46	F	46
10	3	74	38	27	44	75	69
11	10	62	76	84	80	0	F
12	2	25	23	25	54	F	45
15	5	74	53	50	80	79	70
16	8	39	27	37	45	70	71
17	4	69	74	81	89	80	73
18	2	65	67	45	75	75	73
19	4	44	31	36	68	64	74

Once more, Method I appears to be the best, and in this case BFGS and Davidon's method are comparable. In Table 6 we give the percentage of iterations for which $\phi_k < 0$ in Davidon's method and Method I for large, small, and moderate eigenvalues in the initial Hessian approximation. It is clear that negative values of ϕ_k are often used by both methods, especially by Method I.

We conclude from this small set of experiments that Davidon's method is probably superior to BFGS, but its advantage is not great. Method I is clearly better than BFGS, but it is not a practical method. A practical algorithm with performance as good as that of Method I would be a candidate for replacing BFGS as the method of choice for solving small- and medium-size problems. However, such a method would need to be nearly as efficient as Method I to represent a significant improvement over the BFGS method.

REFERENCES

C. G. BROYDEN (1967), *Quasi-Newton methods and their application to function minimization*, Math. Comp., 21, pp. 368–381.

R. H. BYRD, J. NOCEDAL AND Y. YUAN (1987), *Global convergence of a class of quasi-Newton methods on convex problems*, SIAM J. Numer. Anal., 24, pp. 1171–1190.

R. H. BYRD AND J. NOCEDAL (1989), *A tool for the analysis of quasi-Newton methods with application to unconstrained minimization*, SIAM J. Numer. Anal., 26, pp. 727–739.

R. H. BYRD AND Y. XIE (1990), *On Updates from Broyden's Class Including the Hoshino Update*, Report, Computer Science Dept., Univ. of Colorado, Boulder, CO.

W. C. DAVIDON (1975), *Optimally conditioned optimization algorithms without line searches*, Math. Programming, 9, pp. 1–30.

J. E. DENNIS AND J. J. MORÉ (1974), *A characterization of superlinear convergence and its application to quasi-Newton methods*, Math. Comp., 28, pp. 549–560.

——— (1977), *Quasi-Newton methods, motivation and theory*, SIAM Rev., 19, pp. 46–89.

R. FLETCHER (1970), *A new approach to variable metric algorithms*, Comput. J., 13, pp. 317–322.

——— (1987), *Practical Method of Optimization*, John Wiley, New York.

——— (1989), *A New Variational Result for Quasi-Newton Formulae*, Numerical Analysis Report NA/119, Dept. of Mathematics and Computer Science, Univ. of Dundee, Scotland, U.K.

A. GRIEWANK AND PH. L. TOINT (1982), *Local convergence analysis of partitioned quasi-Newton updates*, Numer. Math., 39, pp. 429–448.

D. G. LUENBERGER (1984), *Linear and Nonlinear Programming*, 2nd ed., Addison-Wesley, Reading, MA.

J. J. MORÉ, B. S. GARBOW, AND K. E. HILLSTROM (1981), *Testing unconstrained optimization* *software*, ACM Trans. Math. Software, 7, pp. 17–41.

J. D. PEARSON (1969), *Variable metric methods of minimization*, Comput. J., 12, pp. 171–178.

M. J. D. POWELL (1976), *Some global convergence properties of a variable metric algorithm for minimization without exact line searches*, in Nonlinear Programming, SIAM-AMS Proceedings, Vol. 9, R. W. Cottle and C. E. Lemke, eds., Society for Industrial and Applied Mathematics, Philadelphia, PA.

——— (1986), *How bad are the* BFGS *and* DFP *methods when the objective function is quadratic?*, Math. Programming, 34, pp. 34–47.

K. RITTER (1979), *Local and superlinear convergence of a class of variable metric methods*, Computing, 23, pp. 287–297.

——— (1981), *Global and superlinear convergence of a class of variable metric methods*, Math. Programming Stud., 14, pp. 178–205.

R. B. SCHNABEL (1978), *Optimal conditioning in the convex class of rank two updates*, Math. Programming, 15, pp. 247–260.

D. F. SHANNO AND K. H. PHUA (1978), *Matrix conditioning and nonlinear optimization*, Math. Programming, 14, pp. 149–160.

A. STACHURSKI (1981), *Superlinear convergence of Broyden's class of variable metric methods*, Math. Programming, 20, pp. 196–212.

J. STOER (1979), *On the convergence rate of imperfect minimization algorithms in Broyden's beta-class*, Math. Programming, 9, pp. 313–335.

J. WERNER (1978), *Über die globale konvergenz von Variable-Metric Verfahren mit nichtexakter Schrittweitenbestimmung*, Numer. Math., 31, pp. 321–334.

——— (1989), *Global convergence of quasi-Newton methods with practical line searches*, NAM-Bericht Nr. 67, Institut für Numerische und Angewandte Mathematik der Universität Göttingen, Federal Republic of Germany.

Y. ZHANG AND R.P. TEWARSON (1988), *Quasi-Newton algorithms with updates from the pre-convex part of Broyden's family*, IMA J. Numer. Anal., 8, pp. 487–509.

SIAM J. OPTIMIZATION
Vol. 2, No. 4, pp. 558-574, November 1992

NEW RESULTS ON A CONTINUOUSLY DIFFERENTIABLE EXACT PENALTY FUNCTION*

STEFANO LUCIDI†

Abstract. The main motivation of this paper is to weaken the conditions that imply the correspondence between the solution of a constrained problem and the unconstrained minimization of a continuously differentiable function.

In particular, a new continuously differentiable exact penalty function is proposed for the solution of nonlinear programming problems. Under mild assumptions, a complete equivalence can be established between the solution of the original constrained problem and the unconstrained minimization of this penalty function on a perturbation of the feasible set.

This new penalty function and its exactness properties allow us to define globally and superlinearly convergent algorithms to solve nonlinear programming problems. As an example, a Newton-type algorithm is described which converges locally in one iteration in case of quadratic programming problems.

Key words. exact penalty function, nonlinear programming, constrained optimization

AMS(MOS) subject classifications. primary, 90C30; secondary, 65K05

1. Introduction.

Many research works have been devoted to the study of methods which attempt to solve constrained nonlinear programming problems by means of unconstrained minimizations of continuously differentiable exact penalty functions; see, for example, [1], [2], [3], [4], and [5].

Recently Di Pillo and Grippo [6] have shown that, under suitable compactness and regularity assumptions, it is possible to establish a total equivalence between the solution of a constrained problem and the unconstrained minimization of a differentiable function. In particular, given a point $\tilde{x} \in \mathbb{R}^n$, they consider an open perturbation $\mathcal{D}$ of the feasible set containing both $\tilde{x}$ and the feasible region. Then they define a continuously differentiable exact penalty function which goes to infinity on the boundary of the set $\mathcal{D}$. This feature of the penalty function proposed in [6] and its exactness properties ensure that the original constrained problem can be solved by using any algorithm for the unconstrained minimization of this new penalty function and by employing as a starting point the vector $\tilde{x}$. This correspondence between the constrained and the unconstrained problem is established under some regularity requirements on the problem constraints. More specifically, the penalty function of Di Pillo and Grippo can be defined only if the gradients of the active constraints are linearly independent at every point in $\mathcal{D}$, and its exactness properties can be stated if the extended Mangasarian-Fromovitz constraint qualification (see Proposition 2.5(ii)) holds at every nonfeasible point of $\mathcal{D}$. However, imposing both the linear independence assumption of the gradients of the active constraints and the extended Mangasarian-Fromovitz constraint qualification at every nonfeasible point of $\mathcal{D}$ may limit the applicability of the penalty function introduced by Di Pillo and Grippo, especially when the set $\mathcal{D}$ is much larger than the feasible region. Unfortunately, this situation may occur frequently in practice. In fact, rarely do we have a feasible point; therefore, we often have to use, as a starting point of the unconstrained algorithm, a point $\tilde{x}$ that is far from the feasible set, and this implies that we must choose a very large $\mathcal{D}$. Moreover, if we also have a feasible point, it is better, from the computational point of view, to use a set $\mathcal{D}$ quite different from the feasible set. In fact, if the set $\mathcal{D}$ were a small perturbation of the

* Received by the editors May 23, 1990; accepted for publication (in revised form) August 21, 1991.

† Istituto di Analisi dei Sistemi ed Informatica del CNR, Viale Manzoni 30, 00185 Roma, Italy.

feasible set, then any solution of the constrained problem, which is usually located at the boundary of the feasible region, would be very close to the boundary of $\mathscr{D}$ where the penalty function goes to infinity. This may be the cause of serious numerical instabilities in the minimization algorithm.

In this paper we propose a new continuously differentiable exact penalty function that tries to overcome the limitations presented by the penalty function proposed in [6]. In fact, it allows us to significantly weaken the conditions required outside the feasible set that ensure the total correspondence between the unconstrained minimization of this new penalty function and the solution of the original constrained problem.

Our approach is based on the following considerations. When we want to solve a constrained optimization problem, the situation is very different depending on whether or not we have a feasible point. In the first case the constrained problem is well defined and it is completely characterized by the behaviour of the objective function and the constraints over the feasible set. Therefore, in this case, there is no clear reason to impose a regularity assumption at nonfeasible points and, hence, it should be possible to define a penalty function whose exactness properties could be stated by requiring some regularity assumptions only in the feasible set. In the second case, we do not have any feasible point and, in many real situations, we do not even know whether or not the feasible set is empty. In this case the original constrained problem is composed of two subproblems: the feasibility subproblem and the subproblem of minimizing the objective function. In order to ensure that the first subproblem is well defined, the only actual possibility is to impose a "good behaviour" of the constraints at nonfeasible points. In fact, in this case, if we want to prove that an unconstrained minimization of a penalty function yields a solution of the original constrained problem, we must require some regularity assumptions on the constraints in order to ensure the attainment of feasibility. These assumptions, together with the ones that imply the existence of a minimum point of the penalty function, can be considered as sufficiency conditions for the nonemptiness of the constraint region. Therefore, in the case where we do not have a feasible point, the aim should be to define an exact penalty function which requires assumptions that are as weak as possible, at least for a particular class of constraints.

In this paper we define a new continuously differentiable penalty function which, unlike the penalty function of [6], agrees well with the preceding considerations. In fact, if we have a feasible point, then all exactness properties of this new penalty function can be stated without imposing any assumption outside the feasible set. If a feasible point is not available, then the equivalence between the original constrained problem and the unconstrained minimization of this new penalty function can be established without assuming the linear independence of the gradients of the active constraints outside the feasible set and by requiring, at any infeasible point, only a regularity condition on the constraints, which is much weaker than the extended Mangasarian–Fromovitz constraint qualification. In the general case, this regularity condition is a sufficient condition for the feasible set not to be empty, whereas it is also necessary in the case of compact feasible sets given by convex inequalities. Therefore, at least for this class of systems of constraints, this condition is the weakest possible assumption which ensures that the original constrained problem is well defined, and that, hence, it is possible to state a total correspondence between the constrained problem and the unconstrained minimization of a penalty function.

For the proposed penalty function it is possible to introduce an automatic adjustment rule for the penalty coefficient which allows us to define a general algorithm model for solving nonlinear programming problems. This model is the basis for the construction of implementable Newton-type algorithms which, under suitable assumptions, reconcile global convergence properties with a local superlinear convergence rate. Although the algorithm model proposed in this paper has a structure which derives essentially from the structure of a similar algorithm given in [6], it presents, compared with the preceding one, stronger theoretical properties. In fact, on one hand, its global convergence towards Kuhn-Tucker points of the constrained problem can be established under much weaker regularity assumptions and, on the other hand, following the line of recent papers (see, e.g., [7] and [8]), it is able to give some information about the original problem even when the feasible set is empty or when some regularity assumptions do not hold.

The paper is organized as follows. In § 2 we state the problem and we discuss the assumptions. In § 3 we define the new exact penalty function. In § 4 we establish the exactness properties of this new penalty function. In § 5 we define an automatic adjustment rule for the penalty coefficient and we describe an algorithm which allows us to reconcile the global convergence property with a local superlinear convergence rate.

For the sake of simplicity we consider nonlinear programming problems with inequality constraints; however, the results reported in this paper can be easily extended to nonlinear programming problems with equality and inequality constraints (see [9]). An extensive study of the algorithmic applications of the new penalty function and their computational aspects are beyond the scope of this paper (we refer to [9] for some preliminary results). These arguments will be the subject of future research.

2. Problem formulation and assumptions.

The problem considered is the nonlinear programming problem:

$$\text{(P)} \qquad \text{minimize} \quad f(x) \quad \text{s.t.} \quad g(x) \leqq 0,$$

where $f:\mathbb{R}^n \to \mathbb{R}$ and $g:\mathbb{R}^n \to \mathbb{R}^m$ are twice continuously differentiable functions. We denote by

$$\mathcal{F} := \{x \in \mathbb{R}^n \, g(x) \leqq 0\}$$

the feasible set of (P).

The Lagrangian function associated with (P) is the function $L(x, \lambda) := f(x) + \lambda' g(x)$. A Kuhn-Tucker (K-T) pair for (P) is a pair $(\bar{x}, \bar{\lambda}) \in \mathbb{R}^n \times \mathbb{R}^m$ such that

$$\nabla_x L(\bar{x}, \bar{\lambda}) = 0, \quad G(\bar{x})\bar{\lambda} = 0, \quad \bar{\lambda} \geqq 0, \quad g(\bar{x}) \leqq 0,$$

where $G(x) := \operatorname{diag}(g_i(x))$. Furthermore, we say that the strict complementarity holds at a K-T pair $(\bar{x}, \bar{\lambda})$ if $\bar{\lambda}_i > 0$ for i such that $g_i(\bar{x}) = 0$.

For all $x \in \mathbb{R}^n$ we define the index sets

$$I_\beta(x) := \{i : g_i(x) = 0\}, \quad I_\xi(x) := \{i : g_i(x) < 0\}, \quad I_+(x) := \{i : g_i(x) \geqq 0\}.$$

We denote by $g^+(x)$ the vector with components $g_i^+(x) := \max[0, g_i(x)]$, $i = 1, \ldots, m$.

Let $\alpha, p \in \mathbb{R}$ be given scalars such that $\alpha > 0$ and $p \geqq 2$. In connection with these two scalars we consider an open perturbation of the feasible set $\mathcal{F}$ defined by:

$$\mathcal{A}_{\alpha p} := \left\{ x \in \mathbb{R}^n : \sum_{i=1}^m g_i^+(x)^p < \alpha \right\}$$

and we denote by $\bar{\mathcal{A}}_{\alpha p}$ the closure of $\mathcal{A}_{\alpha p}$. Moreover, we introduce the function

$$a(x) := \alpha - \sum_{i=1}^m g_i^+(x)^p, \tag{2.1}$$

which takes positive values on $\mathcal{A}_{\alpha p}$ and is zero on its boundary. Then we introduce the following regularity condition.

DEFINITION 1. Given the scalars $\alpha > 0$ and $p \geqq 2$, the (α, p)-weak Mangasarian-Fromovitz regularity condition is said to be satisfied at a point $x \in \mathcal{A}_{\alpha p}$ if

$$\sum_{i=1}^{m} \left[1+\frac{p}{2}\frac{\|g^+(x)\|^2}{a(x)} g_i^+(x)^{p-2}\right] g_i^+(x)\nabla g_i(x)=0$$

implies that $g_i^+(x)=0$, for $i=1,\dots,m$.

We remark that if $p=2$, the $(\alpha, 2)$-weak Mangasarian-Fromovitz regularity condition reduces to the fact that $\sum_{i=1}^{m} g_i^+(x)\nabla g_i(x)=0$ implies that $g_i^+(x)=0$, for $i=1,\dots,m$.

In the sequel we shall make use of the following hypotheses.

ASSUMPTION A1. The set $\bar{\mathcal{A}}_{\alpha p}$ is compact.

ASSUMPTION A2. For every $x \in \mathcal{F}$ the gradients $\nabla g_i(x)$, $i \in I_\beta(x)$, are linearly independent.

ASSUMPTION A3. The (α, p)-weak Mangasarian-Fromovitz regularity condition holds at every point $x \in \mathcal{A}_{\alpha p}$.

Assumption A1 is a mild requirement on the constraint functions; in fact, in the following proposition, we give as an example some conditions under which Assumption A1 is satisfied for every $\alpha > 0$ and every $p \geqq 2$.

PROPOSITION 2.1. *Assume that one of the following conditions is satisfied*:

(i) *all the problem variables are bounded*;

(ii) *there exists a function* $g_i(x)$ *such that* $\lim_{\|x\|\to\infty} g_i(x)=\infty$;

(iii) *there exists an index set J such that the functions* $g_i(x)$, $i \in J$, *are convex and the set* $\{x \in \mathbb{R}^n \colon g_i(x) \leqq 0,\ i \in J\}$ *is compact.*

Then Assumption A1 *is satisfied for every* $\alpha > 0$ *and every* $p \geqq 2$.

Proof. If either (i) or (ii) holds the assertion follows immediately. Therefore, assume that (iii) is satisfied. First, it is easy to show that the function $\eta(x)=\sum_{i\in J} g_i^+(x)^p$ is convex for all $p \geqq 2$. Then we recall that, by (iii), the set $\{x \in \mathbb{R}^n \colon \eta(x) \leqq 0\}$ is nonempty and compact and, since it is given by a convex inequality, we can apply Theorem 24 of [10] to also establish that the set $\{x \in \mathbb{R}^n \colon \eta(x) \leqq \alpha\}$ is a compact set. □

We observe that all the globally convergent algorithms for the solution of constrained problems need an assumption similar to Assumption A1. In particular, Assumption A1 is often substituted by the almost equivalent requirement that the sequence of points produced by the algorithm be bounded.

Regarding Assumption A2, we note that it is closely related to the fact that we want to define an exact penalty function that is continuously differentiable. In fact, the key element for the construction of every continuously differentiable exact penalty function proposed in the literature is the definition of a continuously differentiable multiplier function that yields an estimate of the multiplier vector associated with (P) as a function of the variable x. Such a multiplier function requires an assumption like Assumption A2. However, we remark that Assumption A2 appears much weaker than the corresponding assumptions considered in all papers dealing with global convergence of exact penalty function algorithms (see, e.g., [2], [3], and [6]) where the linear independence of the gradients of the active constraints is assumed also outside of the feasible set $\mathcal{F}$.

Now consider Assumption A3; this assumption involves the behaviour of the constraint functions outside the feasible set and, as we said before, it is connected to the feasibility of the original problem. The following propositions clearly show this connection.

PROPOSITION 2.2. *Let $\alpha > 0$ and $p \geqq 2$ be such that the set $\mathcal{A}_{\alpha p}$ is not empty. If Assumptions* A1 *and* A3 *hold on the set $\mathcal{A}_{\alpha p}$, then the feasible set is not empty.*

Proof. Let us introduce the following function $\phi_{\alpha p}(x) := \|g^+(x)\|^p / a(x)$, whose gradient is

$$\nabla \phi_{\alpha p}(x) = \frac{1}{a(x)} \sum_{i=1}^{m} \left[1 + \frac{p}{2} \frac{\|g^+(x)\|^2}{a(x)} g_i^+(x)^{p-2} \right] g_i^+(x) \nabla g_i(x).$$

By assumption, the $\bar{\mathcal{A}}_{\alpha p}$ is not empty and by Assumption A1 it is compact. Then, since $\phi_{\alpha p}(x) \to \infty$ for x converging to any point of $\partial \mathcal{A}$, the function $\phi_{\alpha p}$ admits a global minimum point on $\mathcal{A}_{\alpha p}$ and, hence, it has at least a stationary point on $\mathcal{A}_{\alpha p}$. Now, Assumption A3 implies that any stationary point of $\phi_{\alpha p}$ is a feasible point and hence the proposition follows. □

Obviously it is always possible to find $\alpha > 0$ and $p \geqq 2$ such that the set $\Omega_{\alpha p}$ is not empty; in fact, given a point $\tilde{x} \in \mathbb{R}^n$, it is sufficient to choose $\alpha > \sum_{i=1}^{m} g_i^+(\tilde{x})^p$.

PROPOSITION 2.3. *Assume that the feasible set $\mathcal{F}$ is not empty and that $g_1, \ldots, g_m$ are convex functions. Then Assumption* A3 *holds for all $\alpha > 0$ and $p \geqq 2$.*

Proof. Let x be any point in $\mathbb{R}^n$ and let $\tilde{x} \in \mathcal{F}$. By the convexity assumption we have for all $i = 1, \ldots, m$:

$$0 \geqq g_i(\tilde{x}) \geqq g_i(x) + \nabla g_i(x)'(\tilde{x} - x).$$

This implies that, if $g_i(x) > 0$, we must have $\nabla g_i(x)'(\tilde{x} - x) < 0$. Therefore, letting $z = \tilde{x} - x$, we obtain

$$\nabla g_i(x)'z < 0, \qquad i \in \{i : g_i(x) > 0\}.$$

Now, by using Gordan's theorem of the alternative (see, e.g., [11]) and by taking into account that x was arbitrary, we prove the proposition. □

Then, from Proposition 2.1(iii), Proposition 2.2, and Proposition 2.3 we have the following.

COROLLARY 2.4. *Assume that the feasible set $\mathcal{F}$ is a compact set and that $g_1, \ldots, g_m$ are convex functions. Let $\alpha > 0$ and $p \geqq 2$ be such that the set $\mathcal{A}_{\alpha p}$ is not empty. Then the feasible set $\mathcal{F}$ is not empty if and only if Assumption* A3 *holds on $\mathcal{A}_{\alpha p}$.*

The preceding corollary shows that Assumption A3 is the weakest assumption, which implies the consistency of a compact set given by convex inequalities. Regarding more general feasible sets we can state the following result.

PROPOSITION 2.5. *Assume that at $x \in \mathcal{A}_{\alpha p}$ one of the following conditions is satisfied*:

(i) *the following set*:

$$\hat{\mathcal{F}}(x) := \{z \in \mathbb{R}^n : \nabla g_i(x)'z + g_i(x) \leqq 0, \ i \in I_+(x)\}$$

is not empty;

(ii) *the Mangasarian-Fromovitz regularity condition holds; namely, there exists a $z \in \mathbb{R}^n$ such that*:

$$\nabla g_i(x)'z < 0, \qquad i \in I_+(x).$$

Then we have that the (α, p)-weak Mangasarian-Fromovitz regularity condition holds at x for every $\alpha > 0$ and every $p \geqq 2$.

Proof. If (i) holds there exists a $z \in \mathbb{R}$ such that

$$\nabla g_i(x)'z \leqq -g_i(x) < 0, \qquad i \in \hat{I}_+(x),$$

where $\hat{I}_+(x) := \{i : g_i(x) > 0\}$. Then Gordan's theorem of the alternative ensures that

$$\sum_{i \in \hat{I}_+(x)} v_i \nabla g_i(x) = 0,$$

with $v_i \geqq 0$ for all $i \in \hat{I}_+(x)$, and implies that $v_i = 0$ for all $i \in \hat{I}_+(x)$. Thus the assertion of the proposition follows.

If (ii) is satisfied, then the assertion is a direct consequence of Gordan's theorem of the alternative. □

Proposition 2.5(i) is the typical assumption used in sequential quadratic programming methods (see, e.g. [12], [13], and [14]), whereas point (ii) is the assumption used in [6]. Therefore, by Corollary 2.4 and Proposition 2.5, we can conclude that Assumption A3 is a very mild condition and that, in particular, it is implied by the assumptions used by many of the methods proposed to solve nonlinear programming problems.

In the sequel, we will assume that Assumptions A1 and A2 hold. Assumption A3 will be invoked explicitly when needed.

3. The penalty function. In this section we describe the new continuously differentiable penalty function. First, we introduce a new continuously differentiable multiplier function $\lambda(x)$, which yields an estimate of the multiplier vector associated with (P) as a function of the variable x. This multiplier function is a generalization of the function proposed by Glad and Polak in [3] and its distinguishing property is that it can be defined by assuming that the gradients of the active constraints are linearly independent only on the feasible set $\mathcal{F}$ (Assumption A2) and without requiring any assumption outside of $\mathcal{F}$ (as needed by Glad and Polak's multiplier function).

PROPOSITION 3.1. *For any $x \in \mathbb{R}^n$, $\gamma_1 > 0$, $\gamma_2 > 0$, and $p \geqq 2$, we have*:

(i) *there exists a unique minimizer $\lambda(x)$ of the quadratic function in λ,*

$$\Phi(\lambda; x) := \|\nabla_x L(x, \lambda)\|^2 + \gamma_1^2 \|G(x)\lambda\|^2 + \gamma_2^2 \sum_{i=1}^{m} g_i^+(x)^p \|\lambda\|^2$$

over $\mathbb{R}^m$, given by:

$$\lambda(x) = -M^{-1}(x) \nabla g(x)' \nabla f(x), \tag{3.1}$$

where $M(x)$ is the $m \times m$ matrix defined by

$$M(x) = \nabla g(x)' \nabla g(x) + \gamma_1^2 G^2(x) + \gamma_2^2 \sum_{i=1}^{m} g_i^+(x)^p I_m, \tag{3.2}$$

and I_m indicates the $m \times m$ identity matrix;

(ii) *if $(\bar{x}, \bar{\lambda}) \in \mathbb{R}^n \times \mathbb{R}^m$ is a K-T pair for* (P), *we have $\lambda(\bar{x}) = \bar{\lambda}$*;

(iii) *the Jacobian matrix of $\lambda(x)$ is given by*:

$$\nabla \lambda(x)' = -M^{-1}(x) \Bigg[\nabla g(x)' \nabla_x^2 L(x, \lambda(x)) + \sum_{i=1}^{m} e_i^m \nabla_x L(x, \lambda(x))' \nabla^2 g_i(x)$$
$$+ 2\gamma_1^2 \Lambda(x) G(x) \nabla g(x)' + p\gamma_2^2 \lambda(x) \sum_{i=1}^{m} g_i^+(x)^{p-1} \nabla g_i(x)' \Bigg],$$

where $\Lambda(x) := \operatorname{diag}(\lambda_i(x))$ and e_i^m denote the ith column of the $m \times m$ identity matrix.

Proof. (i) First we consider the $m \times (n+2m)$ matrix:

$$N(x) := \left[\nabla g(x)' \quad \gamma_1 G(x) \quad \gamma_2 \left(\sum_{i=1}^{m} g_i^+(x)^p \right)^{1/2} I_m \right].$$

By Assumption A2 we have that rank $[N(x)] = m$, so that the matrix $M(x) = N(x)N(x)'$ is nonsingular and positive definite. This implies that the vector $\lambda(x)$ is the unique minimizer of the quadratic function $\Phi(\lambda; x)$.

(ii) If $(\bar{x}, \bar{\lambda}) \in \mathbb{R}^n \times \mathbb{R}^m$ is a K-T pair for (P), we have $\Phi(\bar{\lambda}; \bar{x}) = 0$, from which $\lambda(\bar{x}) = \bar{\lambda}$.

(iii) By the assumptions made on the problem functions and the fact that the functions $g_i^+(x)^p$, $i=1,\dots,m$, for $p \geqq 2$ are continuously differentiable, we have that the multiplier function $\lambda(x)$ is also continuously differentiable. The expressions of the gradient $\nabla\lambda(x)$ can easily be derived by differentiating the following equations, which are satisfied by $\lambda(x)$:

$$\nabla g(x)'\nabla_x L(x, \lambda(x)) + \gamma_1^2 G(x)^2 \lambda(x) + \gamma_2^2 \sum_{i=1}^m g_i^+(x)^p \lambda(x) = 0. \qquad \square$$

Now, by using the functions $\lambda(x)$ and $a(x)$, given by (3.1) and (2.1), respectively, we can define the following exact penalty function:

$$Z(x;\, \varepsilon) := f(x) + \lambda(x)'c(x;\, \varepsilon) + \frac{1}{\varepsilon a(x)} \|c(x;\, \varepsilon)\|^2, \tag{3.3}$$

where

$$\begin{aligned} & c(x;\, \varepsilon) := g(x) + Y(x;\, \varepsilon) y(x;\, \varepsilon), \\ & y_i(x;\, \varepsilon) := \left\{ -\min\left[0,\, g_i(x) + \frac{\varepsilon a(x)}{2} \lambda_i(x)\right]\right\}^{1/2}, \qquad i = 1, \dots, m, \\ & Y(x;\, \varepsilon) := \operatorname{diag}(y_i(x;\, \varepsilon)). \end{aligned} \tag{3.4}$$

The expression of the function Z can be derived by repeating the same arguments that led to the expression of the penalty function of [6] (see also [4]). The peculiar features of the function Z are the different structure of the multiplier function $\lambda(x)$ and the particular form of the barrier term $1/a(x)$ on the perturbation $\mathcal{A}_{\alpha p}$ of the feasible set. These differences from the function of Di Pillo and Grippo allow us to weaken the conditions that imply the correspondence between the solution of a constrained problem and the unconstrained minimization of the proposed penalty function.

Given a point $\tilde{x} \in \mathcal{A}_{\alpha p}$ we can define the following level set:

$$\Omega_{\alpha p}(\tilde{x};\, \varepsilon) := \{x \in \mathcal{A}_{\alpha p} \colon Z(x;\, \varepsilon) \leqq Z(\tilde{x};\, \varepsilon)\}.$$

Some preliminary properties of the function $Z(x;\, \varepsilon)$, which are an immediate consequence of its definition, are pointed out in the following proposition.

PROPOSITION 3.2. *For any $\varepsilon > 0$:*

(i) *$Z(x;\, \varepsilon)$ is continuously differentiable for all $x \in \mathcal{A}_{\alpha p}$, with gradient*

$$\begin{aligned} \nabla Z(x;\, \varepsilon) &= \nabla f(x) + \nabla g(x)\lambda(x) + \nabla\lambda(x) c(x;\, \varepsilon) \\ &\quad + \frac{2}{\varepsilon a(x)} \nabla g(x) c(x;\, \varepsilon) + p \frac{\|c(x;\, \varepsilon)\|^2}{\varepsilon a(x)^2} \sum_{i=1}^m \nabla g_i(x) g_i^+(x)^{p-1}; \end{aligned} \tag{3.5}$$

(ii) *$Z(x;\, \varepsilon) \leqq f(x)$ for all $x \in \mathcal{F}$;*

(iii) *$Z(x;\, \varepsilon)$ admits a global minimum point on $\Omega_{\alpha p}(\tilde{x};\, \varepsilon)$;*

(iv) *if f and g are three times continuously differentiable and $p \geqq 3$, then $Z(x;\, \varepsilon)$ is twice continuously differentiable for all $x \in \mathcal{A}_{\alpha p}$ except at the points where $g_i(x) + \varepsilon a(x)\lambda_i(x)/2 = 0$ for some i.*

Proof. Parts (i) and (iv) directly follow from the definition of the function Z.

(ii) By (3.3) we have

$$Z(x;\, \varepsilon) - f(x) = \sum_{i=1}^m \left[\lambda_i(x) c_i(x;\, \varepsilon) + \frac{c_i(x;\, \varepsilon)^2}{\varepsilon a(x)}\right]. \tag{3.6}$$

Now we consider the ith term of the summation in (3.6). If the index i is such that $y_i(x;\,\varepsilon)^2=0$, we can write (taking into account that $g_i(x)\leqq 0$)

$$0\leqq g_i(x)+\frac{\varepsilon a(x)}{2}\lambda_i(x)\leqq\frac{g_i(x)}{2}+\frac{\varepsilon a(x)}{2}\lambda_i(x),$$

$$g_i(x)^2+\varepsilon a(x)\lambda_i(x)g_i(x)\leqq 0,$$

which yields

$$\lambda_i(x)c_i(x;\,\varepsilon)+\frac{c_i(x;\,\varepsilon)^2}{\varepsilon a(x)}=\frac{1}{\varepsilon a(x)}[g_i(x)^2+\varepsilon a(x)\lambda_i(x)g_i(x)]\leqq 0. \tag{3.7}$$

If, on the contrary, the index i is such that $y_i(x;\,\varepsilon)^2>0$, we have

$$\begin{aligned}\lambda_i(x)c_i(x;\,\varepsilon)+\frac{c_i(x;\,\varepsilon)^2}{\varepsilon a(x)}&=\lambda_i(x)\left(-\frac{\varepsilon}{2}a(x)\lambda_i(x)\right)+\frac{1}{\varepsilon a(x)}\left(-\frac{\varepsilon}{2}a(x)\lambda_i(x)\right)^2\\&=-\frac{\varepsilon a(x)}{4}\lambda_i(x)^2\leqq 0.\end{aligned}\tag{3.8}$$

Therefore, by using (3.7) and (3.8) we have that any term of the summation in (3.6) is nonpositive, so that we can conclude that $Z(x;\,\varepsilon)\leqq f(x)$ for any $x\in\mathcal{F}$.

(iii) We prove this point by showing that the set $\Omega_{\alpha p}(\tilde{x};\,\varepsilon)$ is compact. By Assumption A1 we have that $\Omega_{\alpha p}(\tilde{x};\,\varepsilon)$ is bounded. In order to prove that it is also closed we show that every limit point $\bar{x}$ of every sequence $\{x_k\}$ of points in $\Omega_{\alpha p}(\tilde{x};\,\varepsilon)$ still belongs to $\Omega_{\alpha p}(\tilde{x};\,\varepsilon)$. Suppose by contradiction, that $\bar{x}\notin\Omega_{\alpha p}(\tilde{x};\,\varepsilon)$; then we should have that $\bar{x}\in\partial\Omega_{\alpha p}(\tilde{x};\,\varepsilon)$ and, hence, $\lim_{k\to\infty}a(x_k)=a(\bar{x})=0$. Then, by recalling that $x_k\in\Omega_{\alpha p}(\tilde{x};\,\varepsilon)$ for all k, we should obtain:

$$0\geqq\lim_{k\to\infty}a(x_k)[Z(x_k;\,\varepsilon)-Z(\tilde{x};\,\varepsilon)]=\|c(\bar{x};\,\varepsilon)\|^2,$$

which would imply that $\bar{x}\in\mathcal{F}$ and hence that $a(\bar{x})=\alpha>0$. □

4. Exactness properties of the penalty function. In this section we describe the exactness properties of the function $Z(x;\,\varepsilon)$.

By repeating the proofs of Theorem 1 and Proposition 4 of [6], we can state the following proposition.

PROPOSITION 4.1. (i) *Let* $(\bar{x},\bar{\lambda})$ *be a K-T pair for* (P). *Then, for any* $\varepsilon>0$, *we have* $c(\bar{x};\,\varepsilon)=0$, $Z(\bar{x};\,\varepsilon)=f(\bar{x})$, *and* $\nabla Z(\bar{x};\,\varepsilon)=0$.

(ii) *Let* $\bar{x}\in\mathcal{A}_{\alpha p}$ *be a stationary point of* $Z(x;\,\varepsilon)$ *and assume that* $c(\bar{x};\,\varepsilon)=0$. *Then* $(\bar{x},\lambda(\bar{x}))$ *is a K-T pair for* (P).

In order to establish a converse theorem we need some intermediate results.

PROPOSITION 4.2. *Let* $\hat{x}\in\mathcal{F}$. *Then there exist numbers* $\varepsilon(\hat{x})>0$, $\sigma(\hat{x})>0$ *and* $\rho(\hat{x})>0$ *such that, for all* $\varepsilon\in(0,\,\varepsilon(\hat{x})]$ *and for all* $x\in\mathcal{A}_{\alpha p}$ *satisfying* $\|x-\hat{x}\|\leqq\sigma(\hat{x})$, *the following formula holds*:

$$\varepsilon^2\|\nabla g(x)'\nabla Z(x;\,\varepsilon)\|^2\geqq\rho(\hat{x})\|c(x;\,\varepsilon)\|^2. \tag{4.1}$$

Proof. By definition of $y(x;\,\varepsilon)$, we have

$$Y^2(x;\,\varepsilon)\lambda(x)=-\frac{2}{\varepsilon a(x)}Y^2(x,\,\varepsilon)c(x;\,\varepsilon). \tag{4.2}$$

Now, by definition of $\lambda(x)$ and by using (4.2), we can write:

$$\begin{aligned}\nabla g(x)'\nabla_x L(x, \lambda(x)) &= -\gamma_1^2 G(x)^2\lambda(x) - \gamma_2^2 \sum_{i=1}^m g_i^+(x)^p\lambda(x)\\ &= -\gamma_1^2 G(x)\left(\Lambda(x) + \frac{2}{\varepsilon a(x)} Y^2(x;\varepsilon)\right)c(x;\varepsilon)\\ &\quad -\gamma_2^2 \sum_{i=1}^m g_i^+(x)^p\lambda(x).\end{aligned} \tag{4.3}$$

Now, by using (3.5) and (4.3) we get

$$\begin{aligned}\varepsilon\nabla g(x)'\nabla Z(x;\varepsilon) &= \varepsilon\nabla g(x)'\nabla_x L(x;\lambda(x)) + \nabla g(x)'\left(\frac{2}{a(x)}\nabla g(x) + \varepsilon\nabla\lambda(x)\right)c(x;\varepsilon)\\ &\quad + p\frac{\|c(x;\varepsilon)\|^2}{a(x)^2}\sum_{i=1}^m \nabla g(x)'\nabla g_i(x)g_i^+(x)^{p-1}\\ &= K(x;\varepsilon)c(x;\varepsilon) - \varepsilon\gamma_2^2\sum_{i=1}^m g_i^+(x)^p\lambda(x),\end{aligned} \tag{4.4}$$

where

$$\begin{aligned}K(x;\varepsilon) := &\frac{2}{a(x)}(\nabla g(x)'\nabla g(x) - \gamma_1^2 G(x)Y(x;\varepsilon)^2)\\ &+ \varepsilon(\nabla g(x)'\nabla\lambda(x) - \gamma_1^2 G(x)\Lambda(x))\\ &+ \frac{p}{a(x)^2}\sum_{i=1}^m \nabla g(x)'\nabla g_i(x)g_i^+(x)^{p-1}c(x;\varepsilon)'.\end{aligned} \tag{4.5}$$

Now we recall that, for any two vectors $v, u \in \mathbb{R}^m$, we have $2\|v+u\|^2 \geqq \|v\|^2 - 2\|u\|^2$, and that, by using the equivalence of the norms on $\mathbb{R}^m$, there exists a constant $\chi > 0$ such that $\gamma_2^2\sum_{i=1}^m g_i^+(x)^p \leqq \chi\|g^+(x)\|^p$. Therefore, by (4.4) we obtain

$$\varepsilon^2\|\nabla g(x)'\nabla Z(x;\varepsilon)\|^2 \geqq \tfrac{1}{2}\sigma_m(K(x;\varepsilon)^2)\|c(x;\varepsilon)\|^2 - \varepsilon^2\chi^2\|\lambda(x)\|^2\|g^+(x)\|^{2p}, \tag{4.6}$$

where $\sigma_m(K(x;\varepsilon)^2)$ is the smallest eigenvalue of $K(x;\varepsilon)^2$.

Then we can observe that

$$c_i(x;\varepsilon)^2 = \left(\max\left[g_i(x), -\frac{\varepsilon}{2}a(x)\lambda_i(x)\right]\right)^2 \geqq g_i^+(x)^2,$$

so that $\|c(x;\varepsilon)\|^2 \geqq \|g^+(x)\|^2$. Therefore, by (4.6) we get:

$$\varepsilon^2\|\nabla g(x)'\nabla Z(x;\varepsilon)\|^2 \geqq [\tfrac{1}{2}\sigma_m(K(x;\varepsilon)^2) - \varepsilon^2\chi^2\|\lambda(x)\|^2\|g^+(x)\|^{2(p-1)}]\|c(x;\varepsilon)\|^2. \tag{4.7}$$

Now we can note that, if $\hat{x} \in \mathcal{F}$, we have

$$K(\hat{x};0) = \frac{2}{a(\hat{x})}[\nabla g(\hat{x})'\nabla g(\hat{x}) + \gamma_1^2 G(\hat{x})^2] = \frac{2}{a(\hat{x})}\tilde{N}(\hat{x})\tilde{N}(\hat{x})', \tag{4.8}$$

where $\tilde{N}(\hat{x}) := [\nabla g(\hat{x}) \quad \gamma_1 G(\hat{x})]$.

By again using Assumption A2 we obtain that $\operatorname{rank}[\tilde{N}(\hat{x})] = m$ and hence the matrix $K(\hat{x};0)$ is nonsingular and positive definite. Therefore, we can find numbers

$\varepsilon(\hat{x})>0$, $\sigma(\hat{x})>0$, and $\rho(\hat{x})>0$ such that, for all $\varepsilon\in(0, \varepsilon(\hat{x})]$ and for all $x\in\mathcal{A}_{\alpha p}$ satisfying $\|x-\hat{x}\|\leqq\sigma(\hat{x})$, we have

$$\tfrac{1}{2}\sigma_m(K(x;\,\varepsilon)^2)-\varepsilon^2\chi^2\|\lambda(x)\|^2\|g^+(x)\|^{2(p-1)}\geqq\rho(\hat{x})>0, \tag{4.9}$$

and hence the proof of the proposition follows from (4.7) and (4.9). □

LEMMA 4.3. *Let $\{\varepsilon_k\}$ be a sequence of positive numbers converging to zero and let $\{x_k\}$ be a sequence of points such that $x_k\in\Omega_{\alpha p}(\tilde{x};\,\varepsilon_k)$; then $\{x_k\}$ admits a limit point $\hat{x}\in\mathcal{A}_{\alpha p}$.*

Moreover, if we assume that $\tilde{x}\in\mathcal{F}$ or, alternatively, that Assumption A3 *holds and*

$$\lim_{k\to\infty}\varepsilon_k\nabla Z(x_k;\,\varepsilon_k)=0, \tag{4.10}$$

then we have that $\hat{x}\in\mathcal{F}$.

Proof. By Assumption A1 we have that $\bar{\mathcal{A}}_{\alpha p}$ is compact; therefore, there exists a convergent subsequence (relabel it again $\{x_k\}$) such that $\lim_{k\to\infty}x_k=\hat{x}\in\bar{\mathcal{A}}_{\alpha p}$.

If $\hat{x}\in\partial\mathcal{A}_{\alpha p}$ then $\lim_{k\to\infty}a(x_k)=a(\hat{x})=0$. Since $x_k\in\Omega_{\alpha p}(\tilde{x};\,\varepsilon_k)$ for any k, we have

$$0\geqq\lim_{k\to\infty}\varepsilon_k a(x_k)[Z(x_k;\,\varepsilon_k)-Z(\tilde{x};\,\varepsilon_k)]=\|c(\hat{x};\,0)\|^2=\|g^+(\hat{x})\|^2, \tag{4.11}$$

which contradicts the statement $a(\hat{x})=0$ and hence implies that $\hat{x}\in\mathcal{A}_{\alpha p}$ and $a(\hat{x})>0$.

If we assume that $\tilde{x}\in\mathcal{F}$, by Proposition 3.2(ii) we obtain $Z(\tilde{x};\,\varepsilon_k)\leqq f(\tilde{x})$ for any k and, by the continuity assumptions, this yields

$$\limsup_{k\to\infty}Z(x_k;\,\varepsilon_k)=f(\hat{x})+\lambda(\hat{x})'c(\hat{x};\,0)+\limsup_{k\to\infty}\frac{\|c(x_k;\,\varepsilon_k)\|^2}{\varepsilon_k a(x_k)}\leqq f(\tilde{x}),$$

which implies $c(\hat{x};\,0)=0$, so that we have $\hat{x}\in\mathcal{F}$.

Now we suppose that Assumption A3 and (4.10) hold. Then, recalling (3.5) and taking the limit of $\varepsilon_k\nabla Z(x_k;\,\varepsilon_k)$ over the subsequence converging to $\hat{x}$, we obtain:

$$0=\lim_{k\to\infty}\varepsilon_k\nabla Z(x_k;\,\varepsilon_k)=\frac{1}{a(\hat{x})}\sum_{i=1}^{m}\left[1+\frac{p}{2}\frac{\|g^+(\hat{x})\|^2 g_i^+(\hat{x})^{p-2}}{a(\hat{x})}\right]g_i^+(\hat{x})\nabla g_i(\hat{x}),$$

which, by Assumption A3, yields that $\hat{x}\in\mathcal{F}$. □

Now we can establish the following result which, together with Proposition 4.1(i), completes the correspondence between stationary points of $Z(x;\,\varepsilon)$ and the K–T pair for (P).

THEOREM 4.4. *Assume that either $\tilde{x}\in\mathcal{F}$ or Assumption* A3 *holds. Then there exists an $\varepsilon^*>0$ such that for all $\varepsilon\in(0, \varepsilon^*]$, if $x_\varepsilon\in\Omega_{\alpha p}(\tilde{x};\,\varepsilon)$ is a stationary point of $Z(x;\,\varepsilon)$, the pair $(x_\varepsilon, \lambda(x_\varepsilon))$ is a K–T pair for* (P).

Proof. The proof is by contradiction. Namely, whether $\tilde{x}\in\mathcal{F}$ or Assumption A3 holds, we assume that, for any integer k, there exists an $\varepsilon_k\leqq 1/k$ and a point $x_k\in\Omega_{\alpha p}(\tilde{x};\,\varepsilon_k)$ such that $\nabla Z(x_k;\,\varepsilon_k)=0$, but $(x_k, \lambda(x_k))$ is not a K–T pair for (P).

Now Lemma 4.3 ensures that the sequence $\{x_k\}$ admits a limit point $\hat{x}\in\mathcal{F}$. Therefore, taking into account that $\varepsilon_k\to 0$ and that every x_k is a stationary point of Z, we have, by Proposition 4.2, that $c(x_k;\,\varepsilon_k)=0$ for sufficiently large values of k. Then, Proposition 4.1 (ii) ensures that, for sufficiently large values of k, the pair $(x_k, \lambda(x_k))$ is a K–T point for (P) and this establishes the expected contradiction. □

Now we are ready to describe the correspondence between local or global solutions of (P) and local or global unconstrained minimum points of Z.

THEOREM 4.5. *Suppose that* (P) *is well defined, namely, the feasible set $\mathcal{F}$ is not empty. Then there exists an ε^* such that for all $\varepsilon\in(0, \varepsilon^*]$, any global minimum point of* (P) *is a global minimum point of $Z(x;\,\varepsilon)$ on $\mathcal{A}_{\alpha p}$ and conversely.*

Proof. Let ε^* be the threshold value of the penalty parameter introduced in Theorem 4.4. If x_ε is a global minima of $Z(x; \varepsilon)$ on $\mathscr{A}_{\alpha p}$ then we have that $\nabla Z(x_\varepsilon; \varepsilon) = 0$ and $x_\varepsilon \in \Omega_{\alpha p}(\tilde{x}; \varepsilon)$ where $\tilde{x} \in \mathscr{F}$. If $\varepsilon \in (0, \varepsilon^*]$, Theorem 4.4 implies that $(x_\varepsilon, \lambda(x_\varepsilon))$ is a K-T pair for (P) and hence Proposition 4.1 yields $Z(x_\varepsilon; \varepsilon) = f(x_\varepsilon)$. Now, if x^* is a global minimum point for (P), by again using Proposition 4.1 we have $f(x^*) = Z(x^*; \varepsilon)$. Therefore, we can conclude that, for any $\varepsilon \in (0, \varepsilon^*]$, the functions f and Z take the same value in correspondence at any point in $\mathscr{A}_{\alpha p}$ that is either a global minimum point for (P) or a global minimizer of Z. This proves the theorem. □

The proofs of the next results follow, with minor modifications, from those of the corresponding results given in [6] (see [6] or [10] for the definition of the isolated compact set of local minima that appears in Theorem 4.6 (i)).

THEOREM 4.6. (i) *Let $C(f^*)$ be an isolated compact set of local minima of* (P), *corresponding to the local minimum value f^*; then there exists an ε^* such that for all $\varepsilon \in (0, \varepsilon^*]$, $x^* \in C(f^*)$ implies that x^* is a local unconstrained minimum point of $Z(x; \varepsilon)$.*

(ii) *Assume that either $\tilde{x} \in \mathscr{F}$ or Assumption* A3 *holds. Then there exists an ε^* such that for all $\varepsilon \in (0, \varepsilon^*]$, if $x^* \in \Omega_{\alpha p}(\tilde{x}; \varepsilon)$ is a local unconstrained minimum point of $Z(x; \varepsilon)$, x^* is a local minimum point of* (P) *and $\lambda(x)$ is the associated K-T multiplier.*

The next proposition concerns the second-order optimality results and it requires that the f and g are three times continuously differentiable, and that $p \geqq 3$.

PROPOSITION 4.7. (i) *Let (x^*, λ^*) be a K-T pair for* (P) *and assume that*

(a) *the strict complementarity holds at (x^*, λ^*);*

(b) *x^* is an isolated local minimum point for* (P) *and satisfying the second-order sufficiency condition:*

$$z'\nabla_x^2 L(x^*, \lambda^*)z > 0, \quad \textit{for all } z\colon \nabla g_\beta(x^*)z = 0,\ z \neq 0.$$

Then, there exists an ε^* such that for all $\varepsilon \in (0, \varepsilon^*]$, x^* is an isolated local minimum point for $Z(x; \varepsilon)$ and the Hessian matrix $\nabla^2 Z(x^*; \varepsilon)$ is positive definite;

(ii) *Suppose that strict complementarity holds at any K-T pair (x^*, λ^*) of* (P) *and assume that either $\tilde{x} \in \mathscr{F}$ or Assumption* A3 *holds. Then, there exists an $\varepsilon^* > 0$ such that, for all $\varepsilon \in (0, \varepsilon^*]$, if $x^* \in \Omega_{\alpha p}(\tilde{x}; \varepsilon)$ is a local unconstrained minimum point of $Z(x; \varepsilon)$, with positive definite Hessian $\nabla^2 Z(x^*; \varepsilon)$, x^* is an isolated local minimum point of* (P), *satisfying the second-order sufficiency condition.*

5. The algorithm model. As in [6], we can define an automatic adjustment rule for the penalty coefficient that appears in the function $Z(x; \varepsilon)$. This rule allows us to propose an implementable algorithm that can be proved to be globally convergent towards K-T points of (P). In the algorithm we make use of an iteration map $A : \mathscr{A}_{\alpha p} \to 2^{\mathscr{A}_{\alpha p}}$ that satisfies the following assumption.

ASSUMPTION A4. For every fixed value of ε and every starting point $x_0 \in \mathscr{A}_{\alpha p}$, all the points x_k produced by A belong to the level set $\Omega_{\alpha p}(x_0; \varepsilon)$ and all the limit points of the sequence produced by A are stationary points of $Z(x; \varepsilon)$.

These requirements on the map A can be easily satisfied by every globally convergent algorithm for the unconstrained minimization of Z. In fact we can always ensure, by simple device, that the trial points produced (along the search direction) remain in $\Omega_{\alpha p}(x_0; \varepsilon)$.

```
ALGORITHM EPS.
Data: $\tilde{x} \in \mathbb{R}^n$, $\varepsilon_0 > 0$ and $\delta > 0$.
Step 0: Choose $\alpha > 0$ and $p \geqq 2$ such that $\tilde{x} \in \mathscr{A}_{\alpha p}$, set $j = 0$ and $z_0 = \tilde{x}$.
Step 1: Set $k = 0$. If $Z(\tilde{x}; \varepsilon_j) \leqq Z(z_j; \varepsilon_j)$ set $x_0 = \tilde{x}$; else set $x_0 = z_j$.
Step 2: If $\nabla Z(x_k; \varepsilon_j) = 0$ go to Step 3; else go to Step 4.
```

Step 3: If $c(x_k;\ \varepsilon_j)=0$ stop; else go to Step 6.
Step 4: If

$$\|\nabla Z(x_k;\ \varepsilon_j)\|^2+\|\nabla g(x_k)'\nabla Z(x_k;\ \varepsilon_j)\|^2 \geqq \delta\|c(x_k;\ \varepsilon_j)\|^2,$$

go to step 5; else go to Step 6.
Step 5: Compute $x_{k+1}\in A[x_k]$, set $k=k+1$, and go to Step 2.
Step 6: Choose $\varepsilon_{j+1}\in(0,\varepsilon_j)$, set $z_{j+1}=x_k$, $j=j+1$, and go to Step 1.

Algorithm EPS differs from the algorithm of [6] only in Step 1. However, although little, this difference is very important because it allows us to take advantage of all the potentialities of the new function Z so that Algorithm EPS presents theoretical properties that are better than those of the algorithm in [6]. In fact, its global convergence can be stated under the same regularity assumptions used to prove the exactness of the new penalty function Z, whereas the convergence of the algorithm given in [6] is limited by the stronger assumptions required by the penalty function of Di Pillo and Grippo. Furthermore, another distinguishing feature of the algorithm proposed here is its capability to extract some information about the original problem even when Assumption A3 does not hold or when the feasible set is empty.

Now we show the convergence properties of Algorithm EPS. If the algorithm produces a finite sequence of points, by applying Proposition 4.1 (ii) directly we obtain this first result.

PROPOSITION 5.1. *If Algorithm* EPS *terminates at some* $x_\nu\in\mathcal{A}_{\alpha p}$, *then* $(x_\nu,\lambda(x_\nu))$ *is a* K-T *pair for* (P).

In what follows we assume that the algorithm produces an infinite sequence of points $\{x_k\}$.

Remark. Algorithm EPS produces a sequence of points x_k which belong to the set $\mathcal{A}_{\alpha p}$ and hence, by Assumption 1, we have that the sequence $\{x_k\}$ is bounded and it admits at least a limit point.

As we said before, under the same assumptions required for the exactness of Z, we can state the global convergence of the algorithm and, in particular, we can show that the penalty parameter ε is updated finitely many times.

THEOREM 5.2. *Suppose that either* $\tilde{x}\in\mathcal{F}$ *or Assumption* A3 *holds. Then the sequence* $\{\varepsilon_j\}$ *produced at Step* 6 *is finite and every limit point* x^* *of the sequence* $\{x_k\}\subseteq\mathcal{A}_{\alpha p}$ *produced by Algorithm* EPS *yields a* K-T *pair* $(x^*,\lambda(x^*))$ *for* (P).

Proof. We note that the hypotheses made on the iteration map A, and the instructions at Step 1 ensure that

$$Z(x_k;\ \varepsilon_j)\leqq Z(\tilde{x};\ \varepsilon_j) \tag{5.1}$$

for all k and for all j.

First we have to show that the sequences $\{z_j\}$ and ε_j produced at Step 6 are finite.

By using Theorem 4.4 and Proposition 4.1 (i) we have that there exists a $j^*>0$ such that the algorithm cannot construct any point z_j with $j\geqq j^*$ on account of a failure to satisfy the test in Step 3.

It follows that the point z_j, $j\geqq j^*$, should have been produced because of a failure to satisfy the test at Step 4; namely, for $j\geqq j^*$ we should have

$$\|\nabla Z(z_{j+1};\ \varepsilon_j)\|^2+\|\nabla g(z_{j+1})'\nabla Z(z_{j+1};\ \varepsilon_j)\|^2<\delta\|c(z_{j+1};\ \varepsilon_j)\|^2. \tag{5.2}$$

By (5.2) and the instructions at Step 1 we have:

$$z_k\in\mathcal{A}_{\alpha p},\qquad \lim_{j\to\infty}\varepsilon_j\|\nabla Z(z_{j+1};\ \varepsilon_j)\|=0, \tag{5.3}$$

and hence Lemma 4.3 yields that the sequence $\{z_j\}$ admits a limit point $\tilde{z} \in \mathscr{F}$. By recalling (4.1) of Proposition 4.2 we have

$$(5.4) \qquad (\varepsilon_j)^2 \|\nabla g(z_{j+1})' \nabla Z(z_{j+1}; \varepsilon_j)\|^2 \geqq \theta(z_{j+1}; \varepsilon_j) \|c(z_{j+1}; \varepsilon_j)\|^2,$$

where

$$\theta(z_{j+1}; \varepsilon_j) = [\tfrac{1}{2}\sigma_m(\tilde{K}(z_{j+1}; \varepsilon_j)) - (\varepsilon_j)^2 \chi^2 \|\lambda(z_{j+1})\|^2 \|g^+(z_{j+1})\|^{2(p-1)}].$$

Since the matrix $\tilde{K}$ is positive definite in a neighbourhood of $\tilde{z}$ for sufficiently small values of ε and since $z_j \to \tilde{z}$ and $\varepsilon_j \to 0$, we have, for sufficiently large values of j, $\theta(z_{j+1}; \varepsilon_j) > (\varepsilon_j)^2$, which contradicts (5.2).

Therefore, we can conclude that the sequence $\{z_j\}$ is finite. Then the algorithm produces an infinite sequence $\{x_k\} \subset \mathscr{A}_{\alpha p}$ and by Assumption A4 every limit point $\bar{x}$ of this sequence is a stationary point of Z in $\mathscr{A}_{\alpha p}$. By Step 4 we also have $c(\bar{x}; \varepsilon_j) = 0$ so that, again by Proposition 4.1 (ii), the theorem is proved. □

The next proposition follows, in some sense, the line (proposed by the recent results given in [7] and [8]) of investigating the behaviour of Algorithm EPS in the absence of the regularity assumptions of Theorem 5.2 (namely, the knowledge of a feasible point or Assumption A3). In particular, this case occurs when the feasible set is empty.

PROPOSITION 5.3. *Let* $\{\varepsilon_j\}$, $\{z_j\} \subseteq \mathscr{A}_{\alpha p}$ *and* $\{x_k\} \subseteq \mathscr{A}_{\alpha p}$ *be the sequences produced by Algorithm* EPS. *Then*

(i) *if the sequence* $\{\varepsilon_j\}$ *is finite, every limit point* x^* *of the sequence* $\{x_k\}$ *yields a* K-T *pair* $(x^*, \lambda(x^*))$ *for* (P);

(ii) *if the sequence* $\{\varepsilon_j\}$ *is infinite, every limit point* $\bar{z} \in \Omega_{\alpha p}$ *of the sequence* $\{z_j\}$ *is such that* $\bar{z} \notin \mathscr{F}$ *and*

$$(5.5) \qquad \sum_{i=1}^{m} \left[1 + \frac{p}{2} \frac{\|g^+(\bar{z})\|^2}{a(\bar{z})} g_i^+(\bar{z})^{p-2}\right] g_i^+(\bar{z}) \nabla g_i(\bar{z}) = 0.$$

Proof. (i) Let $\varepsilon_{\bar{j}}$ be the last element of the sequence $\{\varepsilon_j\}$ and let x^* be any limit point of the sequence $\{x_k\}$. By Assumption A4 and by the test at Step 4 we have

$$\nabla Z(x^*; \varepsilon_{\bar{j}}) = 0, \qquad c(x^*; \varepsilon_{\bar{j}}) = 0,$$

so that Proposition 4.1 (ii) proves Proposition 5.3 (i).

(ii) Consider the sequence $\{z_j\}$ (which is a subsequence of $\{x_k\}$) produced at Step 6. The points z_j are produced because of a failure to satisfy either the test at Step 3, namely,

$$(5.6) \qquad \nabla Z(z_{j+1}; \varepsilon_j) = 0, \qquad c(z_{j+1}; \varepsilon_j) \neq 0,$$

or the test at Step 4, namely,

$$(5.7) \qquad \|\nabla Z(z_{j+1}; \varepsilon_j)\|^2 + \|\nabla g(z_{j+1})' \nabla Z(z_{j+1}; \varepsilon_j)\|^2 < \delta \|c(z_{j+1}; \varepsilon_j)\|^2.$$

Now let $\bar{z}$ be any limit point of $\{z_j\}$; therefore, there exists a subsequence that we relabel $\{z_j\}$ such that $\lim_{j\to\infty} z_j = \bar{z}$.

First we show that $\bar{z} \notin \mathscr{F}$. In fact, if $\bar{z} \in \mathscr{F}$ then, for sufficiently large values of j, both (5.6) and (5.7) would contradict Proposition 4.2.

Now, recalling that $z_{j+1} \in \Omega_{\alpha p}(\tilde{x}; \varepsilon_j)$ for all j, we have by Lemma 4.3 that $\bar{z} \in \mathscr{A}_{\alpha p}$ and $a(\bar{z}) > 0$. Then we can note that both (5.6) and (5.7) imply

$$\lim_{j\to\infty} \varepsilon_j \nabla Z(z_{j+1}; \varepsilon_j) = 0,$$

which yields

$$0 = \lim_{j\to\infty} \varepsilon_j \nabla Z(z_{j+1}; \varepsilon_j) = \frac{1}{a(\bar{z})} \sum_{i=1}^{m} \left[1 + \frac{p}{2} \frac{\|g^+(\bar{z})\|^2}{a(\bar{z})} g_i^+(\bar{z})^{p-2}\right] g_i^+(\bar{z}) \nabla g_i(\bar{z}). \qquad \square$$

We remark that the points that satisfy (5.5) can be viewed as nonfeasible Fritz and John points, in fact, (5.5) shows that the zero vector can be expressed with a linear combination of the gradients of the constraints with coefficients nonnegative and not all equal to zero. Furthermore, we note also that these points are the stationary points of the function $\phi_{\alpha p}(x) = \|g^+(x)\|^2/a(x)$, where $a(x)$ is given by (2.1), and this function can be interpreted, loosely speaking, as a weighted measure of the violation of the constraints. These last features of the points that satisfy (5.5) are pointed out more clearly by setting $p=2$ in Algorithm EPS. In particular, we can state the following result.

PROPOSITION 5.4. *Let $p=2$ in Algorithm* EPS. *If the sequences $\{\varepsilon_j\}$ and $\{z_j\} \subseteq \mathcal{A}_{\alpha 2}$ produced by the algorithm are infinite, then every limit point $\bar{z} \in \mathcal{A}_{\alpha 2}$ of the sequence $\{z_j\}$ is a stationary point of the distance function*

$$\hat{\phi}(x) = \operatorname{dist}[g(x)\,|\,\mathbb{R}^m_-] := \inf_y \{\|g(x)-y\|_2;\ y_i \leqq 0,\ i=1,\dots,m\}. \tag{5.8}$$

Proof. By Proposition 5.3 (and recalling that $p=2$) we have that every limit point $\bar{z}$ of the sequence $\{z_j\}$ satisfies the relation

$$\sum_{i=1}^m g_i^+(\bar{z})\nabla g_i(\bar{z}) = 0,$$

which implies that $\bar{z}$ is a stationary point of the function $\tilde{\phi}(x) = \|g^+(x)\|^2$. Then we note that, at every nonfeasible point, the stationary points of $\tilde{\phi}(x)$ coincide with the stationary points of the function $\hat{\phi}(x) = \|g^+(x)\|$. Now, the thesis of the proposition follows from the fact that (see [7, § 7])

$$\operatorname{dist}[g(x)\,|\,\mathbb{R}^m_-] := \inf_y \{\|g(x)-y\|_2;\ y_i \leqq 0,\ i=1,\dots,m\} = \|g^+(x)\|. \qquad \square$$

When the feasible set $\mathcal{F}$ is convex, we have Corollary 5.5, which follows directly from Theorem 5.2, Corollary 2.4, and Proposition 5.4.

COROLLARY 5.5. *Let $p=2$ in Algorithm* EPS. *Assume that the feasible set $\mathcal{F}$ is a compact set and that $g_1,\dots,g_m$ are convex functions. If the sequences $\{\varepsilon_j\}$ and $\{z_j\} \subseteq \mathcal{A}_{\alpha 2}$ produced by the algorithm are infinite, then the feasible set $\mathcal{F}$ is empty and every limit point $\bar{z} \in \mathcal{A}_{\alpha 2}$ of the sequence $\{z_j\}$ is a minimum point of the distance function* (5.8).

Therefore, Proposition 5.3 and Corollary 5.5 ensure that, when the feasible set $\mathcal{F}$ is bounded and is given by convex inequalities, Algorithm EPS yields a KKT point for (P) if this problem is feasible. However, it provides a point that is as close to feasibility as possible if (P) is nonfeasible.

Proposition 5.4 and Corollary 5.5 are quite similar to some results reported in [7] and [8]. The main differences between our approach and that of [7] and [8] are that, on one hand, Algorithm EPA has the advantage of using a continuously differentiable merit function and, on the other hand, it has the disadvantage of requiring Assumption A2.

In order to complete the description of Algorithm EPS we must specify the iteration map A. In principle, any method for the unconstrained minimization of the function Z can be easily modified to satisfy Assumption A4 and hence it can be employed as iteration map A in Algorithm EPS. In particular, if the map A is a globally convergent modification of a Newton-type method, then, under the assumptions of Theorem 5.2, Algorothm EPS allows us to solve (P) and to conciliate the global convergence property with an ultimate superlinear convergence rate. Although the Hessian matrix $\nabla^2 Z$, where it exists, requires the third-order derivatives of the problem functions, it is possible to define some Newton-type algorithms based on consistent approximation of the Newton direction of the penalty function Z that employ only the first- and second-order derivatives of the problem functions. Here, as an example, we describe one of these algorithms and we refer the reader to [9] for some other Newton-type algorithms for the function Z.

First we need some additional notation. For every $x\in\mathcal{A}_{\alpha p}$ and every $\varepsilon>0$, we introduce the following index sets:

$$I_A(x;\varepsilon)=\left\{i\colon g_i(x)+\frac{\varepsilon}{2}a(x)\lambda_i(x)\geqq 0\right\},\qquad I_N(x;\varepsilon)=\left\{i\colon g_i(x)+\frac{\varepsilon}{2}a(x)\lambda_i(x)<0\right\}.$$

Then, given a matrix B with columns B_i, $i=1,\dots,m$, we denote by B_A and B_β the submatrices of B consisting of columns B_i, $i\in I_A(x;\varepsilon)$ and B_i, $i\in I_\beta(x)$, respectively. Given an m vector h we denote by h_A, h_N, h_β, and h_ξ the subvectors of h with components h_i, $i\in I_A(x;\varepsilon)$, h_i, $i\in I_N(x;\varepsilon)$, h_i, $i\in I_\beta(x)$, and h_i, $i\in I_\xi(x)$, respectively.

ALGORITHM NT.

$$x_{k+1}=x_k+\alpha_k d_k,$$

where d_k is computed by solving the system

$$\begin{bmatrix}\nabla^2L(x_k,\lambda(x_k)) & \nabla g_A(x_k)\\ \nabla g_A(x_k)' & 0\end{bmatrix}\begin{bmatrix}d_k\\ z_k\end{bmatrix}=-\begin{bmatrix}\nabla f(x_k)\\ g_A(x_k)\end{bmatrix}\tag{5.9}$$

and α_k is the stepsize along the search direction and it can be computed by means of some line search procedure.

Now we can state the following proposition, which describes the local behaviour of Algorithm NT.

PROPOSITION 5.6. *Let (x^*,λ^*) be a K-T pair for (P) satisfying the strict complementarity assumption and let d be the solution of system (5.9). Then*

(i) *for any given $\varepsilon>0$, there exists a neighbourhood $\mathcal{B}$ of x^* and a continuous matrix $H(x;\varepsilon)$ such that, for all $x\in\mathcal{B}$, we have*

$$H(x;\varepsilon)d=-\nabla Z(x;\varepsilon)$$

and in particular we have $H(x^*;\varepsilon)=\nabla^2Z(x^*;\varepsilon)$;

(ii) *if (P) is a quadratic programming problem and the second-order sufficiency conditions for (P) hold at (x^*,λ^*), then, for any given $\varepsilon>0$, there exists a neighbourhood $\mathcal{B}^*$ of x^* such that, for any $x\in\mathcal{B}^*$, we have*

$$x^*=x+d.$$

Proof. The proof of part (i) is quite cumbersome and, for the sake of brevity, we omit it; however, it can be derived by following essentially the same steps as the proof of Proposition 8 of [15].

(ii) In this case we have $f(x)=x'Qx/2+c'x$ and $g(x)=D'x-b$. Under the assumption stated, the pair (x^*,λ^*) is the unique solution of the system in (x,λ)

$$Qx+c+D\lambda=0,\quad D'_\beta x-b_\beta=0,\quad \lambda_\xi=0.\tag{5.10}$$

Let $x\in\mathcal{A}_{\alpha p}$. We note that, if the vector (d,z) solves system (5.9), then the vector (d,u), where $u_A=z-\lambda(x)_A$ and $u_N=-\lambda(x)_N$, solves the system

$$\begin{aligned}&Qd+Du=-(Qx+c+D\lambda(x)),\\ &D'_Ad=-(D'_Ax-b_A),\\ &u_N=-\lambda(x)_N.\end{aligned}\tag{5.11}$$

Then by Proposition 3.1 and the definition of the index sets I_A, I_N, I_β, and I_ξ, we have that if (x^*, λ^*) is a K–T pair for (P) satisfying the strict complementarity assumption, then, for any given $\varepsilon > 0$, there exists a neighbourhood $\mathscr{B}^*$ of x^* such that for all $x \in \mathscr{B}^*$ we have

$$I_A(x;\, \varepsilon) = I_\beta(x^*), \qquad I_N(x;\, \varepsilon) = I_\xi(x^*).$$

Therefore, if $x \in \mathscr{B}^*$ we obtain from (5.11):

$$\begin{aligned} &Q(x+d) + D(u+\lambda(x)) + c = 0,\\ &D'_\beta(x+d) - b_\beta = 0,\\ &u_\xi + \lambda(x)_\xi = 0. \end{aligned}$$

Therefore, the pair $(x+d,\, u+\lambda(x))$ solves the system (5.10) and, hence, coincides with (x^*, λ^*). □

Proposition 5.6 (i) and Proposition 4.7 ensure that the direction d is a consistent approximation of the Newton's direction and that, in a neighbourhood of a local minimum point of (P), it is a descent direction. By using these results, a global and superlinearly convergent algorithm can be defined by using any stabilization technique (see, e.g., [16], [17], and [18]). Part (iii) shows that Algorithm NT converges locally in one iteration if the problem considered is a quadratic programming problem and hence, in this case, it takes advantage of the simple structure of this problem. This property is not peculiar to the new penalty function Z; however, it also holds for the penalty functions proposed in [4] and [6], but it seems that the authors of these papers did not notice this feature.

REFERENCES

[1] R. E. Fletcher (1970), *A class of methods for nonlinear programming with termination and convergence properties*, in Integer and Nonlinear Programming, J. Abadie, ed., North-Holland, Amsterdam, pp. 157–173.

[2] D. P. Bertsekas (1982), *Constrained Optimization and Lagrange Multiplier Methods*, Academic Press, New York.

[3] T. Glad and E. Polak (1979), *A multiplier method with automatic limitation of penalty growth*, Math. Programming, 17, pp. 140–155.

[4] G. Di Pillo and L. Grippo (1985), *A continuously differentiable exact penalty function for nonlinear programming problems with inequality constraints*, SIAM J. Control Optim., 23, pp. 72–84.

[5] S.-P. Han and O. L. Mangasarian (1983), *A dual differentiable exact penalty function*, Math. Programming, 25, pp. 293–306.

[6] G. Di Pillo and L. Grippo (1986), *An exact penalty method with global convergence properties for nonlinear programming problems*, Math. Programming, 36, pp. 1–18.

[7] J. V. Burke and S.-P. Han (1989), *A robust sequential quadratic programming method*, Math. Programming, 43, pp. 277–303.

[8] J. V. Burke (1989), *A sequential quadratic method for potentially infeasible mathematical programs*, J. Math. Anal. Appl., 139, pp. 319–351.

[9] S. Lucidi (1989), *New results on a continuously differentiable penalty function*, Tech. Rep. R.277, Istituto di Analisi dei Sistemi ed Informatica del CNR, Roma, Italy.

[10] A. V. Fiacco and G. P. McCormick (1968), *Nonlinear Programming: Sequential Unconstrained Minimization Techniques*, John Wiley, New York.

[11] O. L. Mangasarian (1969), *Nonlinear Programming*, Prentice-Hall, Englewood Cliffs, NJ.

[12] U. M. Garcia and O. L. Mangasarian (1976), *Superlinear convergent quasi-Newton algorithms for nonlinearly constrained optimization problems*, Math. Programming, 11, pp. 1–13.

[13] S.-P. Han (1977), *A globally convergent method for nonlinear programming*, J. Optim. Theory Appl., 22, pp. 297–309.

[14] M. J. D. POWELL (1978), *Algorithms for nonlinear constraints that use Lagrangian functions*, Math. Programming, 14, pp. 224-248.

[15] G. DI PILLO, L. GRIPPO, AND S. LUCIDI (1986), *Globally convergent exact penalty algorithms for constrained optimization*, in System Modelling and Optimization, A. Prepoka, J. Szelezsan, and B. Strazicky, eds., Springer-Verlag, Berlin, New York.

[16] J. J. MORÈ AND D. C. SORENSEN (1984), *Newton's Method*, G. H. Golub, ed., The Mathematical Association of America, Washington, DC, pp. 29-82.

[17] J. E. DENNIS AND R. B. SCHNABEL (1983), *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*, Prentice-Hall, Englewood Cliffs, NJ.

[18] L. GRIPPO, F. LAMPARIELLO AND S. LUCIDI (1986), *A nonmonotone line search technique for Newton's method*, SIAM J. Numer. Anal., 23, pp. 707-716.

SIAM J. OPTIMIZATION
Vol. 2, No. 4, pp. 575-601, November 1992

003

ON THE IMPLEMENTATION OF A PRIMAL-DUAL INTERIOR POINT METHOD*

SANJAY MEHROTRA†

Abstract. This paper gives an approach to implementing a second-order primal-dual interior point method. It uses a Taylor polynomial of second order to approximate a primal-dual trajectory. The computations for the second derivative are combined with the computations for the centering direction. Computations in this approach do not require that primal and dual solutions be feasible. Expressions are given to compute all the higher-order derivatives of the trajectory of interest. The implementation ensures that a suitable potential function is reduced by a constant amount at each iteration.

There are several salient features of this approach. An adaptive heuristic for estimating the centering parameter is given. The approach used to compute the step length is also adaptive. A new practical approach to compute the starting point is given. This approach treats primal and dual problems symmetrically.

Computational results on a subset of problems available from *netlib* are given. On mutually tested problems the results show that the proposed method requires approximately 40 percent fewer iterations than the implementation proposed in Lustig, Marsten, and Shanno [*Tech. Rep.* TR J-89-11, Georgia Inst. of Technology, Atlanta, 1989]. It requires approximately 50 percent fewer iterations than the dual affine scaling method in Adler, Karmarkar, Resende, and Veiga [*Math. Programming*, 44 (1989), pp. 297-336], and 35 percent fewer iterations than the second-order dual affine scaling method in the same paper. The new approach for estimating the centering parameter and finding the step length and the starting point have contributed to the reduction in the number of iterations. However, the contribution due to the use of second derivative is most significant.

On the tested problems, on the average the implementation shown was found to be approximately two times faster than OB1 (version 02/90) described in Lustig, Marsten, and Shanno and 2.5 times faster than MINOS 5.3 described in Murtagh and Saunders [*Tech. Rep.* SOL 83-20, Dept. of Operations Research, Stanford Univ., Stanford, CA, 1983].

Key words. linear programming, interior point methods, primal-dual methods, power series methods, predictor-corrector methods

AMS(MOS) subject classifications. 90C05, 90C06, 90C20, 49M15, 49M35

1. Introduction.

This paper considers interior point algorithms for simultaneously solving the primal linear program:

$$(P)\qquad \begin{array}{rl} \text{minimize} & c^Tx \\ \text{s.t.} & Ax=b, \\ & x\geqq 0, \end{array}$$

and its dual

$$(D)\qquad \begin{array}{rl} \text{maximize} & b^T\pi \\ \text{s.t.} & A^T\pi+s=c, \\ & s\geqq 0, \end{array}$$

where $c, x, s \in \Re^n$, $\pi, b \in \Re^m$, and $A \in \Re^{m\times n}$. It is assumed that A has full row rank. This can be ensured by removing the linearly dependent rows in the beginning. The

* Received by the editors June 18, 1990; accepted for publication (in revised form) August 30, 1991. This paper was presented at the Second Asilomar Workshop on Progress in Mathematical Programming, Monterey, CA, February 5-7, 1990. This research was supported in part by National Science Foundation research initiation grant CCR-8810107 and by a grant from the GTE Laboratories.

† Department of Industrial Engineering and Management Sciences, Northwestern University, Evanston, Illinois 60208 (mehrotra@iems.nwu.edu).

primal-dual methods, which use solutions of both (P) and (D) in the scaling matrix, are of primary interest.

The primal-dual algorithms have their roots in Megiddo [21]. These were further developed and analyzed by Kojima, Mizuno, and Yoshise [15] and Monteiro and Adler [27]. They showed that the central trajectory can be followed to the optimal solution in $O(\sqrt{n}L)$ iterations by taking "short steps." Kojima, Mizuno, and Yoshise [16] showed that the primal-dual potential function [31], which is a variant of Karmarkar's potential function [13], can also be reduced by a constant amount at each iteration and therefore, they developed a primal-dual large step potential reduction algorithm.

McShane, Monma, and Shanno [20] were the first to develop an implementation of this method. They found it to be a viable alternative to the then popular dual affine scaling method [1], [26] for solving large sparse problems. They also found that this method typically takes fewer iterations than the dual affine scaling method. However, it was found to be only competitive with the dual affine scaling method because of the additional computations that their implementation had to perform. This implementation created some artificial problems (by adding artificial variables/constraints to the original problem) and maintained primal and dual feasible solutions of these problems. Further developments on this implementation were reported in Choi, Monma, and Shanno [4].

Lustig, Marsten, and Shanno [18] implemented a variant of the primal-dual method which is based on an earlier work of Lustig [17]. This method is developed by considering the Newton direction on the optimality conditions for the logarithmic barrier problem. An important feature of their approach is that it did not explicitly require that feasible solutions for the primal or the dual problem be available. They showed that the resulting direction is a particular combination of primal-dual affine scaling direction, feasibility direction, and centering direction. In support of their method they reported success in solving all the problems in the *netlib* [7] test set.

This paper builds on the work of Lustig, Marsten, and Shanno [18]. In doing so it makes use of the work of Monteiro, Adler, and Resende [28] and Karmarkar, Lagarias, Slutsman, and Wang [14]. The discussion in this paper assumes that direct methods are preferred over iterative methods for solving linear equations arising at each iteration of the algorithm. In our view, the following accomplishments are reported in this paper:

- It gives an algorithm and describes its implementation by using first and second derivatives of the primal-dual affine scaling trajectory Taylor polynomial and by effectively combining the second derivative with a centering direction.
- A comparison with the results reported in the literature (on mutually tested problems) shows that the method developed in this paper takes approximately 50 percent fewer iterations than the dual affine scaling method as implemented by Adler, Karmarkar, Resende, and Veiga [1], 40 percent fewer iterations than the primal-dual method implemented in Lustig, Marsten, and Shanno [18], and 55 percent fewer iterations than the logarithmic barrier function method implemented in Gill, Murray, and Saunders [8]. It requires 35 percent fewer iterations than the second-order dual affine scaling method implemented in Adler, Karmarkar, Resende, and Veiga [1] and 20 percent fewer iterations than the "optimal three-dimensional method" implemented by Domich, Boggs, Donaldson, and Witzgall [5].
- An efficient preliminary implementation of the proposed approach was developed. On average, it was found to be two times faster than OB1 (version 02/1990) [18]. On average, it was also found to be 2.5 times faster than MINOS 5.3.

- While developing our implementation we ensure that a suitable potential function is reduced by a constant amount. This is accomplished by taking a two tier approach. Most of the work is performed at the first level, which uses extensive heuristic arguments. The second level ensures the robustness of the implementation.
- It gives expressions for computing first and higher derivatives of a primal-dual trajectory. This trajectory starts from any positive point and goes to the optimum.
- It gives an adaptive approach to computing the centering parameter.
- It gives a modified heuristic for computing step length at each iteration. The approach allows us to adaptively take steps much closer to the boundary.
- It gives an approach to generating primal and dual starting points, which treat these problems symmetrically.

We find it convenient to outline the proposed approach first. This is done in the next section. The organization of this paper is given in that section. The following notation and terminology is used throughout this paper.

Notation and terminology. x^k, π^k, and s^k represent the estimate of solutions of (P) and (D) at the beginning of iteration k. X^k and S^k are used to represent diagonal matrices whose elements are $x_1^k, x_2^k, \ldots, x_n^k$ and $s_1^k, s_2^k, \ldots, s_n^k$, respectively. $\xi_x^k = Ax^k - b$, $\xi_s^k = A^T\pi^k + s^k - c$, ξ_x, and ξ_s are referred to as *error vectors.* D^2 represents matrix $(S^k)^{-1}X^k$. e is used to represent a vector of all ones. e_i represents column i of an identity matrix. $\| \|$ is used to represent the Euclidean norm of a vector.

The term *search direction in primal space* is used for a direction p_x, which is constructed from a combination of directions (p_x1, p_x2). The term *primal blocking variable* is used for a variable that will first become negative when moving in a direction. The term *step factor* represents the fraction of step length that makes the blocking variable zero. Similar terminology is used for directions in the dual space.

Central trajectory is the set of feasible points in (P) and (D) satisfying $x_i(\mu)s_i(\mu) = \mu$, for $i = 1, \ldots, n$. In all references to central trajectory we assume that $x(\mu)$, $s(\mu)$ exists for all μ.

2. Implementation of an interior point method.

This section outlines the approach we take to implement an interior point method. We do this to provide a complete picture of this paper and to fix certain additional notations used throughout. Various steps in the development of this implementation are discussed in more detail in §§ 3-7. The procedure is outlined in Exhibit 2.1 and it is called AIPM (an interior point method). We now discuss the procedure.

```
Procedure AIPM
Input: Let $x^0 > 0$ and $s^0 > 0$, $\pi^0$ be the given starting points.
For $k = 0, 1, \ldots$ until a stopping criterion is satisfied do:
    Step 0
        $\xi_s^k := A^T\pi^k + s^k - c,$
        $\xi_x^k := Ax^k - b,$
        $D^2 := S^{k^{-1}}X^k.$
    Step 1
c find the first derivative of primal-dual affine scaling trajectory.
        $p_\pi 1 := -(AD^2A^T)^{-1}(b - AD^2\xi_s^k),$
        $p_s 1 := \xi_s^k - A^Tp_\pi 1,$
        $p_x 1 := x^k - D^2p_s 1.$
```

EXHIBIT 2.1. *A pseudo-code for implementing a second-order primal-dual method.*

```
    Step 2
c compute centering parameter $\mu^k$.
                  CALL CENPAR $(x^k, p_x1, s^k, p_s1, \mu^k)$.
c compute the second derivative of primal-dual trajectory with the centering direction.
    Step 3
            $v_i := -2 * ((p_x1)_i * (p_s1)_i - \mu^k)/s_i^k$ for $i = 1, 2 \ldots n$,
            $p_\pi 2 := (AD^2A^T)^{-1}Av$,
            $p_s2 := A^T p_\pi 2$,
            $p_x := v - D^2 p_s 2$.
c construct a Taylor polynomial and find maximum steps $(\varepsilon_x, \varepsilon_s)$ using this polynomial.
    Step 4
            CALL SFSOP $(x^k, p_x1, p_x2, \varepsilon_x, s^k, p_s1, p_s2, \varepsilon_s)$.
c construct a search direction.
    Step 5
                  $p_s := \varepsilon_s * p_s1 - .5 * \varepsilon_s^2 * p_s2$,
                  $p_\pi := \varepsilon_s * p_\pi 1 - .5 * \varepsilon_s^2 * p_s2$,
                  $p_x := \varepsilon_x * p_x1 - .5 * \varepsilon_x^2 * p_x2$.
c compute step factors $(f_x, f_s)$.
    Step 6
                  CALL GTSF $(x^k, p_x, s^k, p_s, f_x, f_s)$.
c generate trial points.
    Step 7
                  $\hat{x} := x^k - f_x * p_x$,
                  $\hat{s} := s^k - f_s * p_s$,
                  $\hat{\pi} := \pi^k - f_s * p_\pi$.
c test if the trial point is acceptable.
    Step 8
        If an appropriate potential function is reduced, then
                  $x^{k+1} := \hat{x}$,
                  $s^{k+1} := \hat{y}$,
                  $\pi^{k+1} := \hat{\pi}$,
        else
            perform a line search/if necessary compute
            additional vectors and ensure reduction in the potential function.
        endif
```

EXHIBIT 2.1 (*continued*).

The approach used to generate a starting point is discussed in § 7.

Given x^k, π^k, and s^k, Step 0 computes error vectors ξ_x^k and ξ_s^k representing the amount by which primal and dual constraints are violated. D^2 has the primal-dual scaling matrix.

Step 1 computes direction p_x1 in primal and $p_\pi 1$, $p_s 1$ in dual spaces. These directions are tangent to a primal-dual trajectory. This trajectory is discussed in § 4. Expressions to compute all derivatives of this trajectory at a point are also developed in § 4.

The primal and dual directions computed at Step 1 are used in procedure CENPAR to estimate the centering parameter μ^k. Our approach to estimating the centering parameter is given in Exhibit 5.1. This approach is discussed in § 5.

Step 3 computes the second derivative of the primal-dual trajectory and the centering direction. These directions could be computed separately. However, in the current implementation we prefer to combine their computation in order to save a forward and a back solve. We use the tangent direction and the direction in Step 3 to construct a Taylor polynomial and to find a maximum step to the boundary (in primal and dual spaces separately) using this polynomial. This is done in Procedure SFSOP given in Exhibit 4.1. The computations performed in this procedure are also discussed in § 4.

In Step 5 we use the maximum step in a Taylor polynomial to generate search directions p_x, p_π, and p_s. In Procedure GTSF (Exhibit 6.1) we compute a fraction (f_x, f_s) of the total step to the boundary in the search direction. This is discussed further in § 6. Using the search directions and the step factors, trial point $\hat{x}$, $\hat{\pi}$, $\hat{s}$ is generated in primal and dual spaces.

Step 8 ensures the robustness of the overall procedure. It is loosely defined here. It depends on the choice of the function used to measure the progress of the algorithm and the best possible theoretical results that could be proved for this function. The potential function we used to ensure the progress is developed in the next section (§ 3), and our motivations for using it are discussed there.

If the potential function is not reduced by the desired amount at the trial points, we may perform a line search and, if necessary, compute additional directions to ensure a reduction in this function. This actually happened for the potential function we discuss in the next section. If this happens, we generate an additional three trial points by using $\varepsilon_x := \varepsilon_s := \min(\varepsilon_x, \varepsilon_s)$; $\varepsilon_x := \varepsilon_x$, $\varepsilon_s := 0$; and $\varepsilon_x := 0$, $\varepsilon_s := \varepsilon_s$ in Step 5 to compute p_x, p_π, p_s. The potential function was always reduced by the desired amount at one of the new trial points. Therefore, on the tested problems, additional vectors were never computed and explicit line searches were never performed.

3. A potential function. In our view the interior point methods generate one or more interesting search directions at each iteration and effectively combine these directions to ensure that sufficient progress in a suitable convergence function is made. Various proposed methods differ in the directions they compute, in how they combine these directions (implicitly or explicitly), and in the convergence function they use to measure the progress [12]. Unfortunately, to our knowledge, the current theoretical understanding of these methods has not reached a point where a clear superiority of one method is established. Hence, practical implementations [1], [5], [18], [19], [20], [22], [26] rely on heuristic arguments and empirical evidence obtained from performing experiments on a set of real problems.

Use of a suitable potential function is frequently ignored while developing fast implementations. In our experience, an important reason, among others, is that the cost of performing line searches in one- or higher-dimensional subspaces is significant on sparse problems, and it is frequently not justified by the return.

In our opinion use of heuristic arguments is justified, but not at the cost of the robustness of the solution procedures. Hence, even though in this paper several different heuristics are proposed and their use justified solely on the basis of empirical evidence, in the actual implementation we recommend the use of a potential function.

We now develop the potential function that was used to measure progress in our implementation. We find it instructive to go through some construction to motivate this function. Some steps used in this construction appeared in Karmarkar [13] and others in Goldfarb and Mehrotra [9] and Todd and Ye [31].

Let $x^0 > 0$, π^0, $s^0 > 0$, be any given point. Let $\xi_x^0 = Ax^0 - b$ and $\xi_s^0 = A^T\pi^0 + s^0 - c$. In order to solve (P) and (D), it is enough to find an optimal solution of

$$\begin{aligned} &\text{minimize} \quad \lambda \\ &\quad\text{s.t.} \quad Ax - \lambda\xi_x^0 = b, \\ (3.1) \qquad &\qquad A^T\pi + s - \lambda\xi_s^0 = c, \\ &\qquad c^Tx - b^T\pi + \lambda(b^T\pi^0 - c^Tx^0) = 0, \\ &\qquad x_i, s_i, \quad \lambda \geqq 0, \quad i = 1, 2, \ldots, n. \end{aligned}$$

$(x^0, \pi^0, s^0, 1)$ is a feasible interior solution of (3.1). Let Z be a matrix whose columns are the basis for the null space of A. Multiplying the second set of equations in (3.1) with $[A^T : Z]^T$ and solving for the free variables π results in

$$(3.2) \qquad \pi = (AA^T)^{-1}(Ac - As + \lambda A\xi_s^0);$$

therefore, solving (3.1) is the same as:

$$\begin{aligned} &\text{minimize} \quad \lambda \\ &\quad\text{s.t.} \quad Ax - \lambda\xi_x^0 = b, \\ (PD0) \qquad &\qquad Z^Ts - \lambda Z^T\xi_s^0 = Z^Tc, \\ &\qquad c^Tx + b^T(AA^T)^{-1}As + \lambda\xi_a = b^T(AA^T)^{-1}Ac, \\ &\qquad x_i, s_i, \lambda \geqq 0, \end{aligned}$$

where $\xi_a = (b^T\pi^0 - c^Tx^0 - b^T(AA^T)^{-1}A\xi_s^0)$. Consider the potential function

$$(3.3) \qquad F(x, s, \lambda) = \rho \ln \lambda - \sum_{i=1}^{n} \ln x_i s_i$$

for $\rho = 2n + \sqrt{2n+1}$. In the Appendix we show that $F(x, s, \lambda)$ can be reduced by a constant amount (.25) at any feasible solution of $(PD0)$.

Let $\xi^k = ((b - Ax^k)^T, (c - s^k)^TZ, b^T(AA^T)^{-1}A(c - s^k) - c^Tx^k)^T$. Note that ξ^k is λ^k times the last column in $(PD0)$. If $x^k, \pi^k, s^k, \lambda^k$ and $x^{k+1}, \pi^{k+1}, s^{k+1}, \lambda^{k+1}$ are feasible solutions of $(PD0)$, then

$$\begin{aligned} F(x^{k+1}, s^{k+1}, \lambda^{k+1}) - F(x^k, s^k, \lambda^k) &= \rho \ln \frac{\lambda^{k+1}}{\lambda^k} - \sum_{i=1}^{n} \ln \frac{x_i^{k+1}s_i^{k+1}}{x_i^k s_i^k} \\ &= \rho \ln \frac{\|Q\xi^{k+1}\|}{\|Q\xi^k\|} - \sum_{i=1}^{n} \ln \frac{x_i^{k+1}s_i^{k+1}}{x_i^k s_i^k} \end{aligned}$$

for any nonsingular matrix $Q \in \Re^{(n+1)\times(n+1)}$. An important consequence of this observation is that it ensures that the potential function

$$E(x, s, \xi, Q) = \rho \ln \|Q\xi\| - \sum_{i=1}^{n} \ln x_i s_i \tag{3.4}$$

can be reduced by a constant amount at each iteration. We use the following potential function

$$E(x, s, \xi) = \rho \ln \|(\kappa_x \xi_x^T, \kappa_s \xi_s^T, \xi_a)^T\| - \sum_{i=1}^{n} \ln x_i s_i, \tag{3.5}$$

where κ_x and κ_s are some prespecified constants. The potential function (3.5) is used for the following reasons: (i) We think that a potential function of the form (3.5) is superior for developing implementations because it allows for numerical errors [10]. (ii) It is easily computable without having to know Z. (iii) It allows us to separately update primal and dual solutions and the corresponding error vectors. (iv) There is no unknown that has to be determined during the algorithm. (v) Finally, it is possible to compute directions which ensure that (3.3), and therefore (3.5), is reduced by a constant amount.

The potential function (3.5) is, however, dependent on the scaling of rows (in general, the choice of Q in (3.4)). Because of this, a search direction that may be acceptable while using one scaling matrix may become unacceptable for a different choice. However, in our implementation we use it to our advantage. We think that the construction of directions p_x1, $p_\pi 1$, p_s1 and p_x2, $p_\pi 2$, p_s2 discussed in the next section is inherently biased towards finding (nearly) feasible solutions first. The values of κ_x and κ_s are chosen so that they emphasize primal and dual feasibility over the feasibility of the last equality constraint in ($PD0$). As a consequence of this, search directions that reduce ξ_x and/or ξ_s significantly, and do not reduce (or possibly increase) the error in the last equality constraint of (3.1) become acceptable.

$\kappa_x = 100 * \max_i \{s_i^0\}$ and $\kappa_s = 100 * \max_i \{x_i^0\}$ were used for all the problems in our implementation. The construction of x^0 and s^0 is described in § 7.

A reduction by constant amount in (3.5) at each iteration ensures convergence to an optimal solution provided that $\sum_{i=1}^n \ln x_i s_i$ remain bounded. On the other hand, if (3.5) cannot be reduced by a constant amount at some iteration, then either (P) or (D) or both do not have a feasible solution. We may introduce a constraint providing an upper bound on x and s if we detect (through some tests) that the method is not converging.

4. Derivatives of a primal-dual trajectory. This section provides motivation for using directions p_x1, $p_\pi 1$, p_s1, p_x2, $p_\pi 2$, p_s2 in Procedure AIPM. It was mentioned that these directions use first and second derivative information of a primal-dual trajectory at a given point. This section defines the trajectory of interest and also shows how to compute all of its derivatives at a given point. While we used the potential function (3.5) to measure the progress, derivatives of the primal-dual trajectory being considered are used because they are easily computed and found to be effective in practice.

The results in Monteiro, Adler, and Resende [28] are used frequently to develop these expressions. Monteiro, Adler, and Resende [28] assume that feasible solutions are available. We do not assume this here. The expressions are given in the context of linear programming problems. Extensions to convex quadratic programming are straightforward.

Assume that $x^k > 0$, π^k, $s^k > 0$ is the current point. Consider the following system of nonlinear equations:

$$\begin{aligned} X(\alpha)s(\alpha) &= \alpha X^k s^k, \\ Ax(\alpha) &= b + \alpha\xi_x^k, \\ A^T\pi(\alpha) + s(\alpha) &= c + \alpha\xi_s^k, \\ x(\alpha) \geqq 0, &\quad s(\alpha) \geqq 0 \end{aligned} \tag{4.1}$$

for $\alpha \in [0, 1]$. Let $w(\alpha) \equiv (x(\alpha), \pi(\alpha), s(\alpha))$ represent the solutions of (4.1) for a given α.

PROPOSITION 4.1. *If the system of equations* (4.1) *has a solution for* $\alpha = 0$, *then it has a solution for all* $\alpha \in [0, 1]$. *Furthermore, the solution is unique for* $\alpha \in (0, 1]$.

Proof. For $\alpha \in (0, 1]$ (4.1) gives the optimality conditions for the weighted logarithmic barrier problems

$$(P_\alpha) \qquad \begin{aligned} \text{minimize} \quad & B(x, \alpha) \equiv (c + \alpha\xi_s^k)^T x - \alpha \sum_{i=1}^n x_i^k s_i^k \ln x_i \\ \text{s.t.} \quad & Ax = b + \alpha\xi_x^k, \\ & x > 0, \end{aligned}$$

and

$$(D_\alpha) \qquad \begin{aligned} \text{maximize} \quad & (b + \alpha\xi_x^k)^T \pi + \alpha \sum_{i=1}^n x_i^k s_i^k \ln s_i \\ \text{s.t.} \quad & A^T\pi + s = c + \alpha\xi_s^k, \\ & s > 0. \end{aligned}$$

Let $x(0)$, $\pi(0)$, $s(0)$ represent a solution of (4.1) for $\alpha = 0$. For a fixed $\alpha \in [0, 1]$, $\tilde{x}(\alpha) = (1-\alpha)x(0) + \alpha x^k$ is a feasible solution for (P_α) and $\tilde{\pi}(\alpha) = (1-\alpha)\pi(0) + \alpha\pi^k$, $\tilde{s}(\alpha) = (1-\alpha)s(0) + \alpha s^k$ is a feasible solution for (D_α). If the feasible set of (P_α) is bounded, then obviously (P_α) has a solution.

We now consider the case when the feasible set of (P_α) is unbounded. Since $\tilde{\pi}(\alpha)$, $\tilde{s}(\alpha)$ is a feasible solution for (D_α), it can be shown that the set $\{d \mid Ad = 0,\ d \geqq 0,\ d \neq 0,\ (c + \alpha\xi_s^k)^T d \leqq 0\}$ is empty. Hence, the set $P_t \equiv \{x \mid Ax = b + \alpha\xi_x^k, x > 0, (c + \alpha\xi_s^k)^T x = t\}$ is bounded for all values of $t < \infty$. Furthermore, since the feasible region of (P_α) is nonempty and unbounded, P_t is nonempty for all values of $t > t^*$, where t^* is the minimum value of $(c + \alpha\xi_s^k)^T x$ subject to $Ax = b + \alpha\xi_x^k$, $x \geqq 0$. Clearly, $B(x, \alpha)$ is bounded over P_t for all values of t. Now, to complete the proof for the existence of the solution, note that in $B(x, \alpha)$, $(c + \alpha\xi_s^k)^T x$ increases linearly in t, while $\max \sum_{i=1}^n (x_i^k s_i^k) \ln x_i$ subject to $\{x \mid Ax = b + \alpha\xi_x^k, x > 0, (c + \alpha\xi_s^k)^T x = t\}$ increases only logarithmically in t.

The proof for the uniqueness of solution follows from the strict convexity of $B(x, \alpha)$. □

Let $d^j w(\alpha)/d\alpha^j$ be the jth derivative of $w(\alpha)$. It is now shown that $d^j w(\alpha)/d\alpha^j$ can be computed for all j at $\alpha = 1$. Differentiating (4.1) for the first time gives

$$\begin{aligned} S(\alpha)\frac{dx(\alpha)}{d\alpha} + X(\alpha)\frac{ds(\alpha)}{d\alpha} &= X^k s^k, \\ A\frac{dx(\alpha)}{d\alpha} &= \xi_x^k, \\ A^T\frac{d\pi(\alpha)}{d\alpha} + \frac{ds(\alpha)}{d\alpha} &= \xi_s^k. \end{aligned} \tag{4.2}$$

The solution of (4.2) at $\alpha = 1$ is given by

$$\begin{aligned}
&\frac{d\pi(1)}{d\alpha} = -(AD^2A^T)^{-1}(b - AD^2\xi_s^k),\\
(4.3)\qquad &\frac{ds(1)}{d\alpha} = \xi_s^k - A^T\frac{d\pi(1)}{d\alpha},\\
&\frac{dx(1)}{d\alpha} = x^k - D^2\frac{ds(1)}{d\alpha}.
\end{aligned}$$

Further differentiating (4.2) gives

$$\begin{aligned}
&\sum_{l=0}^{j}\binom{j}{l}\frac{d^l x_i(1)}{d\alpha^l}\,\frac{d^{(j-l)}s_i(1)}{d\alpha^{(j-l)}} = 0, \quad j \geqq 2, \quad i = 1, \dots, n,\\
(4.4)\qquad &A\frac{d^j x(1)}{d\alpha^j} = 0,\\
&A^T\frac{d^j\pi(1)}{d\alpha^j} + \frac{d^j s(1)}{d\alpha^j} = 0.
\end{aligned}$$

From (4.4) it is clear that $d^jw(\alpha)/d\alpha^j$ can be computed recursively. The derivatives can be computed explicitly from

$$\begin{aligned}
&\frac{d^j\pi(1)}{d\alpha^j} = -(AX^kS^{k^{-1}}A^T)^{-1}AS^{k^{-1}}u,\\
&\frac{d^j s(1)}{d\alpha^j} = -A^T\frac{d^j\pi(1)}{d\alpha^j},\\
(4.5)\qquad &\frac{d^j x(1)}{d\alpha^j} = S^{k^{-1}}u + S^{k^{-1}}X^k\frac{d^j s(1)}{d\alpha^j},\\
&u_i = -j!\sum_{l=1}^{j-1}\left(\frac{d^l x(1)}{d\alpha^l}\right)i\left(\frac{d^{(j-1)}s(1)}{d\alpha^{(j-1)}}\right)i, \quad i = 1, \dots, n; \quad j \geqq 2.
\end{aligned}$$

The recursion (4.5) is the same as the recursion given in Monteiro, Adler, and Resende [28] and Karmarkar et al. [14] (see also Megiddo [21] and Bayer and Lagarias [2]). The derivatives resulting from the computations would be the same if $\xi_x^k = 0$ and $\xi_s^k = 0$ is assumed, and in the latter paper if no centering and reparameterization is done.

The point $w(1-\varepsilon)$ for $1 > \varepsilon > 0$ can be approximated by using the rth-order Taylor polynomial

$$(4.6)\qquad w((1-\varepsilon), r) \equiv w(1) + \sum_{j=1}^{r}\frac{(-\varepsilon)^j}{j!}\,\frac{d^j w(1)}{d\alpha^j}.$$

The Taylor polynomial is considered for the following two reasons.

(1) In the special case Monteiro, Adler, and Resende [28] established powerful results (near the central path) for this approximation.

(2) Computational results indicate that near the optimal solution the first- and second-order approximations result in nearly unit steps. Our results also indicate that, asymptotically, $w(\alpha)$ is well represented by the Taylor polynomial of a lower order. In this context Megiddo [21] has argued that for problems with unique optimal solution, if we start close to an optimal solution, the primal-dual paths take us approximately in a straight line to the optimal solution. Most of the tested problems do not satisfy this assumption, however they still show this property.

In addition to other things, practical implementations that compute more than one direction must offset the cost of doing extra work. Adler et al. [1] were the first to show that in the dual affine scaling method the information from a second derivative can be used to significantly reduce the number of iterations. However, on problems in the *netlib* test set they found that reduction in the number of iterations did not always translate into reduction in cpu time on sparse problems. Computational results were also given in Karmarkar et al. [14] on a small set of "representative problems" using methods implemented in the AT&T KORBX system [3].

This paper restricts itself to using the second-order Taylor polynomial. In fact, we compute only two directions. The tangent direction $d^j w(\alpha)/d\alpha$ is computed at Step 1 of Procedure AIPM. Step 3 combines the computation of a second derivative with that of a centering direction. This saves a forward and a back solve. We must compute two directions at each iteration in order to use the adaptive approach for computing the centering parameter (§ 5).

Our strategy of combining the computations for second derivative and the centering direction seem to work in practice for the following reasons. (i) The performance of interior point methods in practice weakly depends on the choice of centering parameter. (ii) If we view the computations in constructing the Taylor polynomial as that of finding a search direction, then in practice it appears that a wide range of ε can be used without adversely affecting the performance of the implementation. To illustrate this, we would like the reader to compare the iteration counts reported in Mehrotra [24], [25] for a predictor-corrector method with those in Table 8.2. The predictor-corrector method results if we take $\varepsilon = 1$ at each iteration.

We find that taking different steps in primal and dual spaces generally results in superior performance. This is similar to the experience of Choi, Monma, and Shanno [4] for their method. We construct different polynomials

$$x(\varepsilon_2, 2) \equiv x^k - \varepsilon_x p_x 1 + \varepsilon_x^2 p_x 2, \tag{4.7}$$

$$s(\varepsilon_s, 2) \equiv s^k - \varepsilon_s p_s 1 + \varepsilon_s^2 p_s 2 \tag{4.8}$$

in primal and dual spaces. The computations for ε_x and ε_s that use (4.7)-(4.8) are described in Procedure SFSOP (step from second-order polynomial) of Exhibit 4.1. A procedure that finds the root of a quadratic equation is used to implement SFSOP.

```
Procedure SFSOP ($x^k, p_x 1, p_x 2, \varepsilon_x, s^k, p_s 1, p_s 2, \varepsilon_s$)
    Find maximum $0 \leqq \varepsilon_x \leqq 1$ such that $x(\varepsilon_x, 2)$ is feasible.
    Find maximum $0 \leqq \varepsilon_s \leqq 1$ such that $s(\varepsilon_s, 2)$ is feasible.
```

EXHIBIT 4.1. *Computations for step size using the Taylor polynomial.*

Before concluding this section, we point out that if there are reasons to believe that at the current iterate it is better to target a solution satisfying $XS = W$ for some positive diagonal matrix W, then expressions for derivatives of a trajectory taking us to such a point can be obtained in a similar manner. The only difference would be to replace "$\alpha X^k s^k$" in (4.1) with "$\alpha X^k s^k + (1-\alpha)We$." In particular, we may use the Heuristic CENPAR to compute the centering parameter (given in the next section) in order to decide a target point on the central path, then go back and find desired derivatives of a trajectory going to this point.

5. Centering. In § 4 expressions were developed to construct a Taylor polynomial at a given point in order to approximate a path going to an optimal solution. Obviously, the performance of the algorithm depends to a great extent on how well a "small-order" Taylor polynomial approximates this path at the current point, and on the domain in which the Taylor polynomial results in good approximations.

The results in Monteiro, Adler, and Resende [28] and the convergence results of large step polynomial time algorithms proved by Freund [6], Gonzaga and Todd [11], and Ye [32] implicitly or explicitly use the properties of the central path. The projected gradient of the potential function used for analysis in these papers encourages centering. On the other hand, it is not clear if the central path (with equal weights) is the best path to follow, particularly since it is affected by the presence of redundant constraints [30]. Furthermore, the points on (or near) the central path are only intermediate to solving the linear programming problem. It is only the limit point on this path that is of interest to us.

In view of this, we make our implementation weakly dependent on centering. The centering direction is obtained by solving the equations

$$\tilde{p}_\pi = \mu^k (AD^2A^T)^{-1} A S^{k^{-1}} e, \quad \tilde{p}_s = A^T \tilde{p}_\pi, \quad \tilde{p}_x = \mu^k S^{k^{-1}} e - D^2 \tilde{p}_s.$$

μ^k is called the centering parameter. It was mentioned in § 4 that the computation for the centering direction is combined with computations for the second derivative to save an extra forward and backward solve.

Heuristic CENPAR given in Exhibit 5.1 was used to compute μ^k. In the description of this heuristic we assume that the direction tangent to the primal-dual affine scaling trajectory has been computed. The heuristic is adaptive. It attempts to generate a value of μ^k, depending on the progress that could be made by moving in the tangent direction.

Heuristic CENPAR $(x^k, p_x 1, s^k, p_s 1, \mu^k)$

Step 1. Let $\varepsilon_x^1, \varepsilon_s^1$ be computed as follows:

$$\varepsilon_x 1 = \min\left(\frac{x_{lx}^k}{(p_x 1)_{lx}}, 1\right),$$

$$lx = \operatorname{argmin}\left\{\frac{x_i^k}{(p_x 1)_i} \,\middle|\, (p_x 1)_i > 0\right\},$$

$$\varepsilon_s 1 = \min\left(\frac{s_{ls}^k}{(p_s 1)_{ls}}, 1\right),$$

$$ls = \operatorname{argmin}\left\{\frac{s_i}{(p_s 1)_i} \,\middle|\, (p_s 1)_i > 0\right\}.$$

Step 2. Let $mdg = (x - \varepsilon_x 1 p_x 1)^T (s - \varepsilon_s 1 p_s 1)$.

Step 3. Let $\mu^k = \dfrac{x^T s}{n}\left(\dfrac{mdg}{x^T s}\right)^\nu$.

Step 4. Let $ef = \dfrac{(p_x 1)^T D^{-2} (p_x 1) + (p_s 1)^T D^2 (p_s 1)}{x^T s}$.

Step 5. If $(ef > 1.1) \mu^k = \mu^k / \min(\varepsilon_x^1, \varepsilon_s^1)$.

EXHIBIT 5.1. *A heuristic to compute centering parameter.*

The motivation behind various steps in Heuristic CENPAR are now discussed. For the moment assume that $\xi_x = 0$ and $\xi_s = 0$. If this is the case, then $x^T s$ is the current duality gap and mdg is the minimum duality gap that one can achieve by moving in directions $p_x 1$ and $p_s 1$ in primal and dual spaces, respectively. $\varepsilon_x 1$ and $\varepsilon_s 1$ are always taken smaller than one, because at this value the computations for $p_x 1$ and $p_s 1$ ensure that $\xi_x = 0$ and $\xi_s = 0$ if no numerical error is present. Hence, for $\nu = 1$ the choice of μ is such that it targets the point on the central path at which the duality gap is mdg.

The ratio mdg$/x^T s$ provides us "some indication" of how well the primal-dual affine scaling trajectory is being approximated locally. A value of ratio mdg$/x^T s$ near 1 means that the local approximations are not good, whereas mdg$/x^T s$ near zero indicates that the approximations of the trajectory are good.

Table 5.1 gives the number of iterations required to solve the problems for choices of $\nu = 1, 2, 3, 4$. All other parameters were the same as those for results in Table 8.2. The last column of this table gives the number of iterations required to solve the problem if no centering was done. The results in Table 5.1 on the test problems show only a moderate variation in the number of iterations for values of ν between two and four.

The discussion on the computation of the centering parameter thus far assumed that $\xi_x = 0$ and $\xi_s = 0$. If this is not the case, then ef (error factor) is used as an indicator for their contribution to the search direction. If $\xi_x = 0$ and $\xi_s = 0$, then it is easy to see that

$$Dp_s 1 = [DA^T(AD^2A^T)^{-1}AD](X^k S^k)^{1/2} e,$$

$$D^{-1} p_x 1 = [I - DA^T(AD^2A^T)^{-1}AD](X^k S^k)^{1/2} e;$$

hence ef as defined in Step 4 of Exhibit 5.1 is equal to 1. If ef is smaller than 1, it indicates that the presence of ξ_x and ξ_s is probably reducing the norm of the search direction and, therefore, it is expected to allow for larger steps in primal and/or dual spaces. Since ξ_x^k and ξ_s^k reduce linearly in step size when moving in directions $p_x 1$ and $p_s 1$, respectively, larger steps result in greater reduction in the error vectors. Hence, the value of ef smaller than one is not likely to hurt the performance of the implementation.

Now if ef>1, then empirical results indicate that the presence of ξ_x and/or ξ_s results in a reduction in the step length, which adversely affects the improvement in the duality gap as well as the reduction in the error vectors. Therefore, it might be indicating trouble ahead. If this happens in practice, we seem to quickly get out of the trouble spots by placing more emphasis on centering. In the current implementation this is accomplished in Step 5.

6. Step length. Standard practice [1], [5], [18], [19], [20], [26] has been to move a certain fixed distance (step factor) to the boundary to avoid one-dimensional line searches. The step factor in the case of primal-dual methods has typically been .995 or .9995 [18].

Although the performance of step factor =.995 or .9995 appears satisfactory in practice, in our view, it has a major drawback as a heuristic: it limits the asymptotic rate of convergence of the algorithm. Furthermore, during the earlier phase of the algorithm it is overly aggressive. A modified approach to computing step factor is given in Exhibit 6.1. It adaptively allows for larger (and smaller) step factors. In an extreme case it may allow a full step to the boundary and generate a point with zero duality gap.

TABLE 5.1
Performance of implementation for different choices of ν. $+$*Stopped after* 100 *iterations with one digit of accuracy in objective function.*

	ν				
Problem	1	2	3	4	nc
afiro	9	8	7	7	8
adlittle	14	11	10	10	11
scagr7	17	14	13	13	15
stochforl	16	16	16	16	21
sc205	14	12	11	11	11
share2b	14	12	12	13	14
sharelb	23	22	22	20	25
scorpion	13	12	12	12	12
scagr25	19	17	16	17	18
sctapl	17	15	15	15	17
brandy	21	20	20	19	19
scsdl	9	8	8	8	8
israel	30	25	24	24	26
bandm	20	17	17	19	17
scfxml	21	19	18	18	19
e226	25	21	20	20	21
agg	27	22	25	27	56
scrs8	24	21	21	21	22
beaconfd	11	8	7	7	7
scsd6	13	10	10	10	10
ship04s	15	12	13	13	15
agg2	23	21	24	23	26
agg3	22	19	21	21	23
scfxm2	22	18	19	19	20
ship041	14	12	12	12	12
fffff800	36	38	38	38	100^+
ship08s	16	13	13	13	13
sctap2	15	13	12	12	13
scfxm3	25	20	20	20	19
shipl2s	18	16	16	17	19
scsd8	12	10	9	9	9
czprob	39	33	35	35	38
ship081	18	14	14	14	14
shipl21	19	16	16	17	17
25fv47	32	27	26	25	27

Procedure GTSF $(x^k, p_x, s^k, p_s, f_x, f_s)$
Let

$$lx = \operatorname{argmin}\left\{\frac{x_i}{(p_x)_i} \mid (p_x)_i > 0\right\}$$

and

$$ls = \operatorname{argmin}\left\{\frac{s_i}{(p_s)_i} \mid (p_s)_i > 0\right\}.$$

Compute f_x such that

$$(x^k_{lx} - f_x * (p_x)_{lx})(s^k_{lx} - (p_s)_{lx}) = \frac{(x^k - p_x)^T(s^k - p_s)}{n * \gamma_a}, \tag{6.1}$$

$$f_x := \max(f_x, \gamma_f)$$

and f_s such that

$$(s^k_{lx} - f_s * (p_s)_{ls})(x^k_{ls} - (p_x)_{ls}) = \frac{(x^k - p_x)^T(s^k - p_s)}{n * \gamma_a}, \tag{6.2}$$

$$f_s = \max(f_s, \gamma_f)$$

EXHIBIT 6.1. *Computation of step factor.*

The step factor in the primal space is chosen so that the product of primal blocking variable (lx) and the corresponding dual slack is nearly equal to the value their product would take at the point on the central trajectory at which the duality gap is equal to $(x - p_x)^T(s - p_s)/(n * \gamma_a)$. Note that if $\xi_x = 0$ and $\xi_s = 0$, then $(x - p_x)^T(s - p_s)$ is the duality gap at the point obtained after moving full step. The parameter $\gamma_a > 1$ should be used. The parameter $0 < \gamma_f \leqq 1$ is used to safeguard against very small or negative steps. The explanation for computation of step factor for the dual variables is similar. In essence, the choice of f_x and f_s is guessing the minimizer of potential function (3.5) in directions p_x and p_s while implicitly assuming that $\xi_x = 0$ and $\xi_s = 0$.

Provided that the computations for the search directions were performed with sufficient accuracy, the computational experience on the tested problems shows that the number of iterations required to solve the problems is relatively insensitive to the choice of γ_a and γ_f in a large range. We experimented with values of $\gamma_f = .5, .75, .9, .99, .999$ and $\gamma_a = 1/(1 - \gamma_f)$. In our experience we found that on most problems the number of iterations were fewer for a larger choice of γ_f, but the difference was small for γ_f in the range .75 to .999.

However, we observed a very interesting phenomenon. The implementation showed signs of instability for larger values of γ_f for problems brandy, scfxm1, scfxm2, and scfxm3. For these problems to obtain eight digits of accuracy in the solutions at the last two iterations of the algorithm, the conjugate gradient method was needed to improve the accuracy in the search direction. These problems were successfully solved to the desired accuracy for $\gamma_f = .5, .75$ and $\gamma_f = .9$.

An examination of problem data of brandy, scfxm1, scfxm2, and scfxm3 shows that for these problems the set of optimal primal solutions is unbounded. An examination of various stages of our implementation (with different choices of parameters) revealed that allowing for a larger step factor may result in premature convergence of dual slacks corresponding to primal variables unbounded in the optimal set. At a later iteration, this further causes the primal unbounded variables to become very large.

Hence, x_i/s_i corresponding to these variables become disproportionately large. This results in cancellation of large numbers when computing the Cholesky factor, causing computations to become less stable.

The reader is referred to Mehrotra [23] for a more elaborate discussion of the precision of computations in the context of interior point methods and to Mehrotra [22] for a discussion of issues involved in developing implementations based on the preconditioned conjugate gradient method, which we may want to use to improve the numerical accuracy.

7. Initial point. In all of our implementations the initial point is generated as follows. We first compute

$$\tilde{\pi} = (AA^T)^{-1}Ac; \quad \tilde{s} = c - A^T\pi; \quad \tilde{x} = A^T(AA^T)^{-1}b, \tag{7.1}$$

and $\delta_x = \max(-1.5 * \min\{\tilde{x}_i\}, 0)$ and $\delta_s = \max(-1.5 * \min\{\tilde{s}_i\}, 0)$. We then obtain

$$\tilde{\delta}_x = \delta_x + .5 * \frac{(\tilde{x} + \delta_x e)^T(\tilde{s} + \delta_s e)}{\sum_{i=1}^n (\tilde{s}_i + \delta_s)}, \tag{7.2}$$

$$\tilde{\delta}_s = \delta_s + .5 * \frac{(\tilde{x} + \delta_x e)^T(\tilde{s} + \delta_s e)}{\sum_{i=1}^n (\tilde{x}_i + \delta_x)} \tag{7.3}$$

and generate $\pi^0 = \tilde{\pi}$ and $s_i^0 = \tilde{s}_i + \tilde{\delta}_s$, $i = 1, \ldots, n$ and $x_i^0 = \tilde{x}_i + \tilde{\delta}_x$, $i = 1, \ldots, n$ as an initial point.

We first discuss the validity of the above approach in generating $x^0 > 0$ and $s^0 > 0$. In the cases in which it fails to produce such a point, either the problems reduce to that of finding a feasible solution of (P) or (D), or an optimal solution is generated.

From the definition of δ_x and δ_s we know that $x^0 \geqq 0$ and $s^0 \geqq 0$. A positive point is always generated, if $\delta_x > 0$ and $\delta_s > 0$. Furthermore, to show that $x^0 > 0$ and $s^0 > 0$, it is sufficient to show that $\tilde{\delta}_x > 0$ and $\tilde{\delta}_x > 0$.

First consider the case when $\delta_x = 0$ and $\delta_s = 0$. Clearly, in this case $\tilde{x}$ is a feasible solution for (P) and $\tilde{\pi}$, $\tilde{s}$ is a feasible solution for (D). If $\tilde{x}^T\tilde{s} = 0$, then these solutions are optimal for the respective problems. Otherwise, $\tilde{x}^T\tilde{s} > 0$ and hence $\tilde{\delta}_x > 0$ and $\tilde{\delta}_s > 0$. Now consider the case when $\delta_x = 0$ and $\delta_s > 0$. In this case if $\tilde{x}_i \neq 0$ for all i, then obviously $\tilde{\delta}_x > 0$ and $\tilde{\delta}_s > 0$. On the other hand, if $\tilde{x}_i = 0$ for all i, then $b = 0$ and the problem reduces to that of finding a feasible solution of (D). This problem can then be solved separately or by generating a perturbed problem for which the right-hand side is δAe for any positive δ. δe can be used as a feasible interior solution of the perturbed problem, and (7.2)-(7.3) can be used to generate a feasible point of the perturbed problem. Finally, the case when $\delta_x > 0$ and $\delta_s = 0$ can be argued in a similar manner.

We now discuss some properties of the proposed approach.

7.1. Desirable properties of the proposed approach.

Shift in origin. The approach is independent of the shift in origin. To explain what we mean by this, consider the dual problem

$$(D(\Delta)) \qquad \begin{aligned} \text{maximize} \quad & b^T(\pi + \Delta\pi) \\ \text{s.t.} \quad & A^T(\pi + \Delta\pi) + s = c, \\ & s \geqq 0 \end{aligned}$$

for any fixed choice of $\Delta\pi$. Clearly, the polytope defined by the constraints in $(D(\Delta))$ is the same as the polytope defined by the constraints in (D) except for a shift of origin. It is desirable that an initial point be the same in relation to the respective polytopes. Note that $\tilde{s}$ in (7.1) is independent of the choice of $\Delta\pi$.

To demonstrate how similar arguments hold for (P), we consider an equivalent formulation of this problem. Let Z be a matrix whose columns form the basis for the null space of A, and let x_o be any point satisfying $Ax_o = b$. It is easy to see that (P) is equivalent to

$$(P(\Delta)) \qquad \begin{aligned} \text{minimize} \quad & (c^TZ)(y+\Delta y) \\ \text{s.t.} \quad & Z(y+\Delta y) \geqq -x_o \end{aligned}$$

for $y \in \Re^{n-m}$ and any (fixed) choice of Δy. An approach analogous to that of finding $\tilde{s}$ computes

$$\tilde{y} = -(Z^TZ)^{-1}Z^Tx_o.$$

The slacks in the constraint of $(P(Z))$ are given by

$$x_o - Z(Z^TZ)^{-1}Z^Tx_o = [I - Z(Z^TZ)^{-1}Z^T]x_o = A^T(AA^T)^{-1}Ax_o = A^T(AA^T)^{-1}b = \tilde{x}.$$

Note that $\tilde{x}$ is the orthogonal projection of any vector satisfying $Ax = b$ onto the range space of A. Since $\tilde{x}$ and $\tilde{s}$ are independent of Δy and $\Delta\pi$, respectively, it is obvious that x^0 and s^0 are also independent of this.

Simple scaling. The initial point is not affected if all the constraints in (P) are scaled by a constant or if c is scaled.

7.2. Undesirable properties of the proposed approach.

Column scaling. The initial point is affected by the scaling of columns of A. To illustrate this, in Table 7.1 we give a number of iterations required to solve the problems after problems were scaled by using subroutine M5SCAL from MINOS. All other details of the implementation were kept the same as those for results in Table 8.2.

The results in Table 7.1 indicate that scaling of columns may effect the performance of the implementation. On most problems, use of M5SCAL improved the number of iterations required to solve the problem or the number of iterations did not change significantly. The notable exception was problem ffffff800. For this problem, scaling increased the number of iterations by about 40 percent.

We point out that (P) and (D) can be solved in one iteration if we know the correct scaling of columns of A. Hence, the problem of finding the best scaling of columns appears to be as difficult as solving the linear programming problem itself.

Presence of redundant constraint. The initial point generated by using this approach is affected by the presence of redundant constraints. This can be a problem, for example, if redundant dual constraints with large slacks (primal variables with huge costs) are present. In this regard we point out that the central trajectory itself is affected by the presence of such constraints.

A possible way to alleviate this problem would be to ask the user to give relative importance to various primal variables (dual constraints) in solving the problem. A clearly redundant variable could be assigned zero weight, and therefore it could be removed while generating an initial point. In general, the possibility of solving weighted least squares problems to generate initial points in the context of "warm start" should be explored further.

All the approaches [1], [5], [18], [26] used for practical implementations that are reported in the literature are dependent on column scaling and the presence of redundant constraints. Furthermore, they lack one of the desirable properties that the proposed approach has.

TABLE 7.1
Effect of scaling on the performance of the algorithm.

Problem	No scaling	Scaling
afiro	7	6
adlittle	10	10
scagr7	13	14
stochfor1	16	8
sc205	11	11
share2b	12	14
share1b	22	24
scorpion	12	12
scagr25	16	17
sctap1	15	15
brandy	20	17
scsd1	8	7
israel	24	17
bandm	17	15
scfxm1	18	17
e226	20	16
agg	25	16
scrs8	21	18
beaconfd	7	8
scsd6	10	10
ship04s	13	12
agg2	24	20
agg3	21	19
scfxm2	19	20
ship04l	12	12
fffff800	38	52
ship08s	13	14
sctap2	12	12
scfxm3	20	20
ship12s	17	14
scsd8	9	9
czprob	35	30
ship08l	14	15
ship12l	16	15
25fv47	26	23

The choice of constants "1.5" and ".5" in computing $\tilde{\delta}_x$ and $\tilde{\delta}_s$ is arbitrary. The arguments about the validity of the approach (and its properties) do not change if we replace 1.5 with any constant larger than one and .5 with any constant larger than zero. We may do a one-dimensional line search (possibly in direction e) on the potential function (3.5) to generate these constants.

8. Computational performance. The basic ideas presented in this paper were implemented in a FORTRAN code. This section discusses our computational experience with this implementation and compares it with the results documented in the literature.

All testing was performed on a SUN 4/110 work station. The code was compiled with SUN Fortran version 1.0 compiler option "−O3." All the cpu times were obtained by using utility *etime*. The test problems were obtained from *netlib* [7]. The test set includes small and medium size problems. The problem names and additional information on these problems are given in Table 8.1.

All problems were cleaned by using the procedure outlined in Mehrotra [22]. An in-house implementation of the minimum degree heuristic was used to permute the rows. Complete Cholesky factor was computed at each iteration to solve the linear equations. The procedure and the associated data structure, which we used to compute the Cholesky factor, were described in Mehrotra [23]. All the linear algebra subroutines were written by the author.

The algorithm was terminated when the relative duality gap satisfied

$$\frac{c^Tx-b^T\pi}{1+|b^T\pi|}\leqq \varepsilon_{\text{exit}}. \tag{8.1}$$

In the actual implementation ξ_x and ξ_s were computed afresh at each iteration. However, $(\xi_s)_i$ was set to zero and absorbed in s_i if it satisfied $|(\xi_s)_i|/s_i<.001$. This occasionally saved some computational efforts.

The following parameters were set to obtain the results reported in Table 8.2.

$$\begin{aligned}
&\varepsilon_{\text{exit}}=10^{-8} && \text{(in (8.1))},\\
&\nu=3 && \text{(in Procedure CENPAR)},\\
&\gamma_f=.9,\ \gamma_a=10 && \text{(in Procedure GFSF)},\\
&\kappa_x=100*\max\{s_i^0\} && \text{(in (3.5))},\\
&\kappa_s=100*\max\{x_i^0\} && \text{(in (3.5))}.
\end{aligned}$$

The number of iterations required to solve the problems is given in the second column of Table 8.2. The primal objective value recorded at termination is given in column 3. The relative duality gap (8.1) is given in column 4. The primal infeasibility,

$$\|Ax-b\|/(1+\|x\|), \tag{8.2}$$

and the dual infeasibility,

$$\|A^T\pi+s-c\|/(1+\|s\|), \tag{8.3}$$

recorded at termination are given in columns 5 and 6, respectively. This information on relative duality gap and primal and dual feasibility is the same as that given in Lustig, Marsten, and Shanno [18]. We use the same stopping criterion and provide similar information in order to be consistent while making comparisons.

All the problems were accurately solved to eight digits. A comparison with the results in Lustig, Marsten, and Shanno [18] show that on many problems in our implementation the accuracy in the objective value at termination was better. This is primarily due to our approach for computing the step factor.

In Table 8.3 we compare the number of iterations required to solve the test problems. Column 2 of this table gives the number of iterations taken by our implementation. Column 3 gives the number of iterations taken by the dual affine scaling method, as reported by Adler et al. [1]. Column 4 gives the number of iterations required by the second-order dual affine scaling method in Adler et al. [1]. Column 5 gives the number of iterations reported in Lustig, Marsten, and Shanno [18] for a primal-dual method. Column 6 gives the number of iterations reported in Gill, Murray, and Saunders [8] for a logarithmic barrier function method. Column 7 gives the number

TABLE 8.1
Problem statistics.

Problem	Rows	Columns	Nonzeros
afiro	28	32	88
adlittle	57	97	465
scagr7	130	140	553
stochfor1	118	111	474
sc205	206	203	552
share2b	97	79	730
share1b	118	225	1,182
scorpion	389	358	1,708
scagr25	472	500	2,029
sctap1	301	480	2,052
brandy	221	249	2,150
scsd1	78	760	3,148
israel	175	142	2,358
bandm	306	472	2,659
scfxm1	331	457	2,612
e226	224	282	2,767
agg	489	163	2,541
scrs8	491	1,169	4,029
beaconfd	174	262	3,476
scsd6	148	1,350	5,666
ship04s	403	1,458	5,810
agg2	517	302	4,515
agg3	517	302	4,531
scfxm2	661	914	5,229
ship04l	403	2,118	8,450
fffff800	525	854	6,235
ship08s	779	2,387	9,501
sctap2	1,091	1,880	8,124
scfxm3	991	1,371	7,846
ship12s	1,152	2,763	10,941
scsd8	398	2,750	11,334
czprob	930	3,523	14,173
ship08l	779	4,283	17,085
ship12l	1,152	5,427	21,597
25fv47	822	1,571	11,127

TABLE 8.2
Computational performance of the implemented algorithm.

Problem	itn	c^Tx^k	$b^T\pi^k$	$\frac{\lvert c^Tx^k-b^T\pi^k\rvert}{1+\lvert b^T\pi^k\rvert}$	$\frac{\|Ax^k-b\|}{1+\|x^k\|}$	$\frac{\|A^T\pi^k+s^k-c\|}{1+\|s^k\rvert}$	$\frac{\mu^k}{1+\lvert b^T\pi^k\rvert}$
afiro	7	−464.75314272968	−464.75314309228	7e−10	5e−15	0	4e−13
adlittle	10	225,494.96330271	225,494.96311956	8e−10	1e−11	5e−14	4e−13
scagr7	13	−2,331,389.8242996	−2,331,389.8243996	4e−11	2e−13	0	2e−13
stochfor	16	−41,131.976219436	−41,131.976219346	3e−15	4e−15	0	3e−26
sc205	11	−52.202061191133	−52.202061364254	3e−9	4e−12	0	7e−14
share2b	12	−415.73224070502	−415.73224074419	9e−11	8e−12	8e−18	1e−16
share1b	22	−76,589.318579203	−76,589.318579231	3e−13	5e−13	0	1e−15
scorpion	12	1,878.1248227381	1,878.1248227369	5e−13	6e−14	7e−16	1e−17
scagr25	16	−14,753,443.031527	−14,753,433.097009	4e−9	1e−11	1e−16	5e−12
sctap1	15	1,412.2500000000	1,412.2500000000	1e−14	2e−14	0	9e−28
brandy	20	1,518.5098972369	1,518.5098964881	4e−10	1e−09	1e−17	9e−20
scsd1	8	8.6666666743334	8.6666666743334	7e−15	5e−15	0	2e−27
israel	24	−896,644.82032330	−896,644.82213083	2e−10	5e−14	0	9e−14
bandm	17	−158.62801845009	−158.62801845012	2e−13	3e−14	1e−16	3e−18
scfxm1	18	18,416.759026662	18,416.759028090	7e−11	2e−10	4e−16	2e−14
e226	20	−18.751929066371	−18.751929066371	1e−14	6e−14	2e−17	1e−20
agg	25	−35,991,767.286586	−35,991,767.286789	5e−12	2e−13	0	8e−11
scrs8	21	904.29695380418	904.29695377855	2e−11	1e−15	3e−17	2e−13

TABLE 8.2 (*continued*).

Problem	itn	c^Tx^k	$b^T\pi^k$	$\frac{\lvert c^Tx^k-b^T\pi^k\rvert}{1+\lvert b^T\pi^k\rvert}$	$\frac{\lVert Ax^k-b\rVert}{1+\lVert x^k\rVert}$	$\frac{\lVert A^T\pi^k+s^k-c\rVert}{1+\lVert s^k\rvert}$	$\frac{\mu^k}{1+\lvert b^T\pi^k\rvert}$
beaconfd	7	33,592.485814826	33,592.485771448	1e−9	2e−10	2e−11	2e−11
scsd6	10	50.500000078275	50.500000064030	2e−10	6e−15	0	2e−19
ship04s	13	1,798,714.7004927	1,798,714.7004444	2e−11	6e−13	1e−16	1e−12
agg2	24	−20,239,252.354277	−20,239,252.356863	1e−10	5e−14	0	1e−11
agg3	21	10,312,115.9350899	10,312,115.9350813	8e−13	2e−14	0	1e−10
scfxm2	19	36,660.261567320	36,660.261564919	6e−11	1e−10	3e−16	1e−11
ship04l	12	1,793,324.5379701	1,793,324.5379704	1e−13	1e−13	2e−16	3e−14
fffff800	38	555,679.56576888	555,679.56340175	4e−9	2e−10	0	2e−11
ship08s	13	1,920,098.2105549	1,920,098.2105340	1e−11	1e−13	0	3e−13
sctap2	12	1,724.8071428587	1,724.8071428563	1e−12	1e−14	0	7e−16
scfxm3	20	54,901.254586753	54,901.254517497	1e−9	2e−10	6e−15	4e−13
ship12s	16	1,489,236.1347549	1,489,236.1338498	1e−11	3e−12	1e−13	1e−15
scsd8	9	905.00000451087	904.99999866657	6e−9	3e−15	0	2e−15
czprob	35	2,185,196.7039679	2,185,196.6964335	3e−9	2e−12	0	1e−14
ship08l	14	1,909,055.2113734	1,909,055.2113891	8e−12	1e−11	9e−17	6e−20
ship12l	16	1,470,187.9194253	1,470,187.9193126	7e−11	2e−13	6e−15	3e−18
25fv47	26	5,501.8458883209	5,501.8458882282	7e−11	1e−14	2e−17	1e−17

TABLE 8.3
Comparison of number of iterations with other implementations.

Problem	AIPM	AKRV2	AKRV1	LMS	GMS	DBDW
afiro	7	15	20	13	20	10
adlittle	10	18	24	17	18	15
scagr7	13	19	24	22	24	18
stochforl	16			19		
sc205	11	20	28	16	32	17
share2b	12	21	29	17	46	14
sharelb	22	33	38	40	35	28
scorpion	12	19	24	18	33	17
scagr25	16	21	29	24	28	21
sctapl	15	23	33	22	53	21
brandy	20	24	38	27	41	21
scsdl	8	16	19	12	13	8
israel	24	29	37	47	36	24
bandm	17	24	30	28	31	21
scfxml	18	30	33	31	37	23
e226	20	30	34	31	38	24
agg	25			32		
scrs8	21	29	39	50	59	25
beaconfd	7	17	23	21	34	17
scsd6	10	18	22	15	15	13
ship04s	13	22	30	21	40	15
agg2	24			32		
agg3	21			32		
scfxm2	19	29	39	37	42	29
ship04l	12	21	28	22	36	17
fffff800	38			59	55	
ship08s	13	21	32	23	34	16
sctap2	12	25	34	23	41	16
scfxm3	20	30	40	39	42	31
ship12s	16	23	35	27	46	16
scsd8	9	18	23	15	15	13
czprob	35	35	52	57	56	41
ship08l	14	23	31	24	22	17
ship12l	16	23	32	27	24	17
25fv47	24		52	48	44	28

of iterations reported in Domich et al. [5] for a variant of the method of centers. Our method and the second-order dual affine scaling method in Adler et al. [1] computes two directions at each iteration. The method implemented in Domich et al. [5] computes three directions at each iteration and solves a linear programming problem defined by using these directions. All other implementations compute only one direction at each iteration.

On the average the number of iterations required by our implementation to solve the mutually tested problems is 40 percent less than that reported in Lustig, Marsten, and Shanno [18]; it is 50 percent less than that reported in Adler et al. [1]; and it is 55 percent less than that reported by Gill, Murray, and Saunders [8]. Compared to the results in Adler et al. [1] for their second-order method, the results show that our method takes about 35 percent fewer iterations. Finally, our implementation required 20 percent fewer iterations than those required in Domich et al. [5].

It is useful to point out that it is possible to further reduce the total number of iterations needed to solve the problems by using higher-order derivatives. This is discussed in a subsequent paper.

The number of iterations was always fewer than the number of iterations required by the second-order dual affine scaling method implemented by Adler et al. [1]. Many of the problems were solved in practically half the number of iterations when compared with [1].

9. Comparison with OB1 and MINOS 5.3. This section compares the cpu times required by our implementation to those required by the implementation of the primal-dual method in OB1 [18] (02/90 version) and the simplex method in MINOS 5.3 [29]. The source codes (also written in FORTRAN) of OB1 (02/90 version) and MINOS 5.3 were compiled using compiler option "−O3." Hence, everything was identical while making these comparisons. All the default options of OB1 (02/90 version) and MINOS 5.3 were used. Printing was turned to minimum level in both cases. In the case of OB1 (02/90 version), crush = 2 was used for all problems.

The times for MINOS 5.3 are those for subroutine M5SOLV only. The TIMER subroutine in OB1 was used to compute its cpu times. The times for OB1 were calculated as follows:

$$\text{OB1 Time} = \text{end of hprep} - \text{after mpsink} + \text{end of obdriv} - \text{after getcmo}.$$

Times required by MINOS 5.3, OB1 (02/90 version), and our implementation do not include times spent in converting the MPS input file into a problem in the standard form. The times required by our implementation include all the time spent after the input files were converted into a problem in the standard form.

The times required by our implementation is given in the second column of Table 9.1. The times required by OB1 (02/90 version) are given in column 3 and the times required by MINOS are given in column 4. Column 5 gives the ratio of times required by OB1 (02/90 version) to our code. Column 6 gives the ratio of times required by MINOS 5.3 to our code.

From these results we find that, on the average, our implementation in the current state performs two times better than the implementation in OB1 (02/90 version). The ratio of cpu times with OB1 (02/90 version) is more or less uniform.

Comparing the results with MINOS 5.3 we find that the proposed implementation is on the average better by a factor of 2.5. In this case, however, the ratio of cpu times varies significantly. MINOS 5.3 was generally superior on problems with few relatively dense columns, whereas our implementation of the primal-dual method was superior on problems with sparse Cholesky factor.

10. Conclusions. Details of a particular implementation of the primal-dual method are given. This implementation requires a considerably smaller number of iterations and saves considerable computational effort. We have given expressions to compute

TABLE 9.1
Comparison of cpu *time with* OB1 *and* MINOS 5.3 *on* SUN 4/110.

Problem	AIPM	OB1	MINOS5.3	$\frac{\text{OB1}}{\text{AIPM}}$	$\frac{\text{MINOS5.3}}{\text{AIPM}}$
afiro	.12	.60	.09	5.0	.7
adlittle	.64	1.81	.70	2.8	1.1
scagr7	1.11	2.75	1.66	2.5	1.5
stochfor1	1.48	2.91	1.50	2.0	1.0
sc205	1.49	3.33	2.16	2.2	1.4
share2b	1.50	3.00	1.46	2.0	1.0
share1b	3.27	9.21	3.98	2.8	1.2
scorpion	2.87	7.06	5.92	2.4	2.0
scagr25	5.23	10.43	15.32	2.0	2.9
sctap1	4.92	9.18	7.71	1.8	1.6
brandy	7.18	15.30	11.55	2.1	1.6
scsd1	2.46	5.46	6.15	2.2	2.5
israel	58.37	127.01	6.11	2.2	.1
bandm	8.01	17.61	22.09	2.2	2.7
scfxm1	10.55	20.82	12.76	2.0	1.2
e226	9.38	15.83	15.30	1.7	1.6
agg	32.88	47.46	7.32	1.4	.2
scrs8	13.31	43.95	40.86	3.3	3.1
beaconfd	2.56	9.28	1.97	3.6	.8
scsd6	5.66	10.56	31.19	1.9	5.5
ship04s	6.92	17.58	6.63	2.5	.95
agg2	65.86	100.92	10.05	1.5	.15
agg3	66.66	94.74	10.95	1.4	0.16
scfxm2	21.69	46.74	51.42	2.1	2.37
ship04l	8.90	24.35	13.20	2.7	.54
fffff800	80.90	140.01	14.33	1.7	.17
ship08s	9.50	23.05	20.43	2.4	2.1
sctap2	30.04	47.75	56.02	1.6	1.9
scfxm3	33.31	72.84	107.40	2.1	3.2
ship12s	14.06	31.78	47.85	2.2	3.4
scsd8	10.38	21.98	230.23	2.1	22.2
czprob	33.78	81.93	166.03	2.4	4.9
ship08l	18.12	42.81	19.34	2.4	1.0
ship12l	28.56	63.11	120.31	2.2	4.2
25fv47	164.53	334.14	941.41	2.0	5.7
Total	766.20	1,507.29	2,011.4		

all the derivatives at a given point of a primal-dual affine scaling trajectory. The implementation described here effectively combines the second derivative with the centering vector. Heuristics for computing centering parameter and step length were given and their effectiveness was demonstrated. A new approach to generating a starting point was used. In addition, the results demonstrate that it is possible to develop fast (robust) implementations of interior point methods, which ensure sufficient reduction in a potential function at each iteration.

Comparison with OB1 (02/90 version) and the simplex method show that our implementation was faster by a factor of 2 and 2.5, respectively.

Acknowledgments. I thank Professor Michael Saunders for making the source code of MINOS 5.3 available. I thank Professor Roy E. Marsten for releasing a copy of OB1 for comparison in this paper. Also, I thank Mr. I. C. Choi for helping out with OB1 and Professor Donald Goldfarb for his constant encouragement, which made this work possible. I also thank the two anonymous referees for their careful reading of this paper and for their suggestions for improvement.

Appendix. Here we prove that the potential function (3.3) can be reduced by a constant amount at each iteration. The development of our proof is based on the analysis in Freund [6]. Let us define $l \equiv 2n+1$,

$$\hat{A} \equiv \begin{bmatrix} A & 0 & -\xi_x \\ 0 & Z^T & -\xi_s \\ c^T & (AA^T)^{-1}A & \xi_a \end{bmatrix},$$

$\hat{b}^T \equiv (b^T, c^TZ, b^T(AA^T)^{-1}Ac)$, and $y = (x^T, s^T, \lambda)^T$. Hence, without loss of generality, consider the problem

$$(PD) \qquad \begin{aligned} &\text{minimize} \quad y_l \\ &\text{s.t.} \quad \hat{A}y = \hat{b}, \\ &\qquad\quad y \geqq 0, \end{aligned}$$

where $y \in \mathfrak{R}^l$. Let us consider the potential function

$$\text{(A.1)} \qquad F(y) \equiv \hat{\rho} \ln y_l - \sum_{i=1}^{l} \ln y_i,$$

where $\hat{\rho} = l + \sqrt{l}$. The function $F(y)$ in (A.1) is the same as the function (3.3). Let $y^k > 0$ be any feasible point of (PD) and let Y^k be the diagonal matrix whose diagonal elements are $(y_1, \dots, y_l)$. Let

$$d \equiv [I - \bar{A}^T(\bar{A}\bar{A}^T)^{-1}\bar{A}](\hat{\rho}e_l - e),$$

where $\bar{A} = \hat{A}Y^k$. Let $y^{k+1} = y^k - (\varepsilon/\|d\|)Y^k d$, $\varepsilon < 1$. Then

$$\begin{aligned} F(y^{k+1}) - F(y^k) &= \hat{\rho} \ln \frac{y_l^{k+1}}{y_l^k} - \sum_{i=1}^{l} \ln \frac{y_i^{k+1}}{y_i^k} \\ &= \hat{\rho} \ln\left(1 - \frac{\varepsilon}{\|d\|} d_l\right) - \sum_{i=1}^{l} \ln\left(1 - \frac{\varepsilon}{\|d\|} d_i\right) \\ &\leqq -\frac{\hat{\rho}\varepsilon}{\|d\|} d_l + \frac{\varepsilon}{\|d\|} \sum_{i=1}^{n} d_i + \frac{\varepsilon^2}{2(1-\varepsilon)} \\ &= -\varepsilon\|d\| + \frac{\varepsilon^2}{2(1-\varepsilon)}. \end{aligned}$$

The inequality above follows by using the fact that $\ln(1+\delta) \leqq \delta$ for $\delta > -1$, and $\ln(1+\delta) \geqq (\delta^2/2(1-\varepsilon))$ if $|\delta| \leqq \varepsilon < 1$.

If $\|d\| \geqq 1$, then for $\varepsilon = .5$, $F(y)$ is reduced by .25.

Otherwise, if $\|d\| < 1$, then from the definition of d, we have

$$\bar{A}^T(\bar{A}\bar{A}^T)^{-1}\bar{A}\left(y_l^k e_l - \frac{y_l^k}{\hat{\rho}} e\right) + \frac{y_l^k}{\hat{\rho}}(d+e) = y_l^k e_l,$$

which gives a feasible solution to the dual of (PD). Furthermore, the duality gap is given by

$$y_l^k \frac{e^T(d+e)}{\hat{\rho}} \leqq y_l^k \frac{l+\sqrt{l}\|d\|}{l+\sqrt{l}} < y_l^k.$$

But y_l^k is also the current objective value. Therefore, the optimal objective value of (PD) must be positive. If this is the case, then either (P) or (D) has no feasible solution, and we would stop.

Hence, if (P) and (D) have a feasible solution then $F(y)$ can be reduced by .25 at each iteration. A failure to reduce $F(y)$ by this amount would imply that either (P) or (D) has no feasible solution.

REFERENCES

[1] I. ADLER, N. KARMARKAR, M. G. C. RESENDE, G. VEIGA (1989), *An implementation of Karmarkar's algorithm for linear programming*, Math. Programming, 44, pp. 297-336.

[2] D. A. BAYER AND J. C. LAGARIAS (1989), *The nonlinear geometry of linear programming:* I. *Affine and projective scaling trajectories*, II. *Legendre transform coordinates and central trajectories*, Trans. Amer. Math. Soc., pp. 499-581.

[3] Y. C. CHENG, D. J. HOUCK, J. M. LIU, M. S. MEKETON, L. SLUTSMAN, R. J. VANDERBEI, AND P. WANG (1989), *The* AT&T KORBX *System*, AT&T Tech. J., 68, pp. 7-19.

[4] I. C. CHOI, C. MONMA, AND D. F. SHANNO (1990), *Further development of a primal-dual interior point method*, ORSA J. Comput., 2, pp. 304-311.

[5] P. D. DOMICH, P. T. BOGGS, J. R. DONALDSON, AND C. WITZGALL (1989), *Optimal* 3-*dimensional methods for linear programming*, NISTIR 89-4225, Center for Computing and Applied Mathematics, U.S. Dept. of Commerce, National Institute of Standards and Technology, Gaithersburg, MD.

[6] R. FREUND (1988), *Polynomial-time algorithms for linear programming based only on primal scaling and projected gradient of a potential function*, Working Paper OR 182-88, Operations Research Center, Massachusetts Institute of Technology, Cambridge, MA.

[7] D. M. GAY (1985), *Electronic mail distribution of linear programming test problems*, Mathematical Programming Society Committee on Algorithms News Letter, 13, pp. 10-12.

[8] P. E. GILL, W. MURRAY, AND M. A. SAUNDERS (1988), *A single-phase dual barrier method for linear programming*, Report SOL 88-10, Systems Optimization Laboratory, Stanford Univ., Stanford, CA.

[9] D. GOLDFARB AND S. MEHROTRA (1988), *A relaxed of Karmarkar's algorithm*, Math. Programming, 40, pp. 285-315.

[10] ——— (1989), *A self-correcting version of Karmarkar's algorithm*, SIAM J. Numer. Anal., 26, pp. 1006-1015.

[11] C. GONZAGA AND M. J. TODD (1989), *An $O(\sqrt{n}L)$ iterations large-step primal-dual affine algorithm for linear programming*, Tech. Report, School of Operations Research and Industrial Engineering, Cornell University, Ithaca, NY.

[12] D. D. HERTOG AND C. ROSS (1989), *A survey of search directions in interior point methods for linear programming*, Report 89-65, Faculty of Mathematics and Informatics/Computer Science, Delft Univ. of Technology, Delft, Holland.

[13] N. KARMARKAR (1984), *A new polynomial time algorithm for linear programming*, Combinatorica, 4, pp. 373-395.

[14] N. KARMARKAR, J. C. LAGARIAS, L. SLUTSMAN, AND P. WANG (1989), *Power series variants of Karmarkar-type algorithms*, AT&T Tech. J., May/June, pp. 20-36.

[15] M. Kojima, S. Mizuno, and A. Yoshise (1989), *A primal-dual interior point algorithm for linear programming*, in Progress in Mathematical Programming, Interior Point and Related Methods, N. Megiddo, ed., Springer-Verlag, New York, pp. 29-47.

[16] ——— (1991), *An $O(\sqrt{n}L)$-iteration potential reduction algorithm for linear complementarity problems*, Math. Programming, 50, pp. 331-342.

[17] I. J. Lustig (1991), *Feasibility issues in a primal-dual interior-method for linear programming*, Math. Programming, 49, pp. 145-162.

[18] I. J. Lustig, R. E. Marsten, and D. F. Shanno (1989), *Computational experience with a primal-dual interior point method for linear programming*, TR J-89-11, Industrial and Systems Engineering Report Series, Georgia Inst. of Technology, Atlanta, GA.

[19] R. E. Marsten, M. J. Saltzman, D. F. Shanno, G. S. Pierce, and J. F. Ballintijn (1989), *Implementation of a dual affine interior point algorithm for linear programming*, ORSA J. Comput. 1, pp. 287-297.

[20] K. A. McShane, C. L. Monma, and D. Shanno (1989), *An implementation of a primal-dual interior point method for linear programming*, ORSA J. Comput., 1, pp. 70-83.

[21] N. Megiddo (1986), *Pathways to the optimal set in linear programming*, in Progress in Mathematical Programming, N. Megiddo, ed., Springer-Verlag, New York, pp. 131-158.

[22] S. Mehrotra (1992), *Implementations of affine scaling methods: Approximate solutions of systems of linear equations using preconditioned conjugate gradient methods*, ORSA J. Comput., 4, pp. 103-118.

[23] ——— (1989), *Implementations of affine scaling methods: Towards faster implementations with complete Cholesky factor in use*, TR-89-15R, Dept. of Industrial Engineering/Management Science, Northwestern Univ., Evanston, IL.

[24] ——— (1991), *On finding a vertex solution using interior point methods*, Linear Algebra Appl., 152, pp. 233-253.

[25] ——— (1990), *On an implementation of primal-dual predictor-corrector algorithms*, presented at the Second Asilomar Workshop on Progress in Mathematical Programming, Asilomar, CA.

[26] C. L. Monma and A. J. Morton (1987), *Computational experience with a dual affine variant of Karmarkar's method for linear programming*, Oper. Res. Lett., 6, pp. 261-267.

[27] R. C. Monteiro and I. Adler (1989), *An $O(n^3L)$ primal-dual interior point algorithm for linear programming*, Math. Programming, 44, pp. 43-66.

[28] R. C. Monteiro, I. Adler, and M. G. C. Resende (1990), *A polynomial-time primal-dual affine scaling algorithm for linear and convex quadratic programming and its power series extension*, Math. Oper. Res., 15, pp. 191-214.

[29] B. A. Murtagh and M. A. Saunders (1983), MINOS 5.0 *user's guide*, Tech. Rep. SOL 83-20, Dept. of Operations Research, Stanford Univ., Stanford, CA.

[30] M. J. Todd (1989), *The affine-scaling direction for linear programming is a limit of projective-scaling direction*, Tech. Rep., School of Operations Research and Industrial Engineering, Cornell Univ., Ithaca, NY.

[31] M. J. Todd and Y. Ye (1990), *A centered projective algorithm for linear programming*, Math. Oper. Res., 15, pp. 508-529.

[32] Y. Ye (1991), *An $O(n^3L)$ potential reduction algorithm for linear programming*, Math. Programming, 50, pp. 239-258.

SIAM J. OPTIMIZATION
Vol. 2, No. 4, pp. 602-618, November 1992

004

ON REGULARIZED LEAST NORM PROBLEMS*

ACHIYA DAX†

Abstract. This paper investigates the regularized least norm problem

$$\text{minimize } F(\mathbf{x}) = (\varepsilon/s)\|\mathbf{x}\|_s^s + \|A\mathbf{x}-\mathbf{b}\|_p,$$

where ε is a positive constant, $1<s<\infty$, and $1<p<\infty$. Let $\mathbf{x}_\varepsilon$ denote the solution that corresponds to a given value of ε, and let $\mathbf{x}^*$ denote the minimum l_s norm solution of the unregularized least l_p norm problem. It is shown that $\mathbf{x}_\varepsilon$ is a continuous function of ε, $\|\mathbf{x}_\varepsilon\|_s \leqq \|\mathbf{x}^*\|_s$, $\lim_{\varepsilon\to\infty}\mathbf{x}_\varepsilon = \mathbf{0}$, and $\lim_{\varepsilon\to 0}\mathbf{x}_\varepsilon = \mathbf{x}^*$. Furthermore, if the system $A\mathbf{x}=\mathbf{b}$ is solvable then there exists a positive constant δ such that $\mathbf{x}_\varepsilon = \mathbf{x}^*$ for all $\varepsilon \in (0, \delta]$. The question of whether $\mathbf{x}_\varepsilon = \mathbf{x}^*$ is related to a new theorem of the alternative. The main result is the observation that the dual of the regularized least norm problem has the form

$$\begin{aligned} &\text{maximize} \quad D(\mathbf{y}) = \mathbf{b}^T\mathbf{y} - (\varepsilon/t)\|A^T\mathbf{y}/\varepsilon\|_t^t \\ &\text{subject to} \quad \|\mathbf{y}\|_q \leqq 1, \end{aligned}$$

where $t = s/(s-1)$ and $q = p/(p-1)$. Moreover, the primal solution $\hat{\mathbf{x}}$ is easily recovered from a dual solution $\hat{\mathbf{y}}$, and vice versa. This pair of points satisfies $D(\hat{\mathbf{y}}) = F(\hat{\mathbf{x}})$ and the classical primal-dual inequality $D(\mathbf{y}) \leqq F(\mathbf{x})$ holds for all $\mathbf{x}\in\mathbb{R}^n$ and $\mathbf{y}\in\mathbb{R}^m$ such that $\|\mathbf{y}\|_q \leqq 1$. The paper presents an iterative improvement process which, under certain conditions, converges toward a solution of the unregularized least norm problem. The inequality $\|\mathbf{y}\|_q \leqq 1$ introduces an obstacle into the solution of the dual problem, but this obstacle may be removed by applying penalty function methods.

Key words. l_p least norm problems, regularization, optimality conditions, behavior of regularized solutions, duality relations

AMS(MOS) subject classifications. 65K99, 65O99

1. Introduction. This paper investigates the regularized least norm problem

$$\text{minimize} \quad F(\mathbf{x}) = (\varepsilon/s)\|\mathbf{x}\|_s^s + \|A\mathbf{x}-\mathbf{b}\|_p \tag{1.1}$$

where ε is a positive constant, $1<s<\infty$, $1<p<\infty$, A is a real $m\times n$ matrix, $\mathbf{b} = (b_1, \ldots, b_m)^T \in \mathbb{R}^m$, and $\mathbf{x} = (x_1, \ldots, x_n)^T \in \mathbb{R}^n$ denotes the vector of unknowns. For the sake of clarity we mention that

$$\|\mathbf{x}\|_s = \left(\sum_{j=1}^n |x_j|^s\right)^{1/s}$$

and

$$\|A\mathbf{x}-\mathbf{b}\|_p = \left(\sum_{i=1}^m |\mathbf{a}_i^T\mathbf{x} - b_i|^p\right)^{1/p},$$

where $\mathbf{a}_i^T = (a_{i1}, \ldots, a_{in})$ denotes the ith row of A. The dual indices of s and p are denoted by t and q, respectively. That is,

$$\frac{1}{p}+\frac{1}{q}=1 \quad\text{and}\quad \frac{1}{s}+\frac{1}{t}=1.$$

One motivation for our research is that solving (1.1) may serve as a possible way to extend the concept of regularization to non-Euclidean norms. The method of

* Received by the editors December 3, 1990; accepted for publication (in revised form) September 30, 1991.

† Hydrological Service, P.O.B. 6381, Jerusalem 91060, Israel.

regularization has been introduced by Tikhonov and others as a means of improving the stability of ill-posed problems (see, e.g., Tikhonov and Arsenin (1977)). In particular it turns out to be a useful tool for handling data fitting problems which result in a system of linear equations

$$A\mathbf{x}=\mathbf{b}$$

such that small changes in the data (i.e., the elements of A or $\mathbf{b}$) cause large changes in the minimum norm solution of the least squares problem

$$\text{minimize} \quad \tfrac{1}{2}\|A\mathbf{x}-\mathbf{b}\|_2^2. \tag{1.2}$$

Suppose, for example, that we have an a priori estimate, $\mathbf{z}_0$, of the solution of (1.2). Then one way to "stabilize" the solution is by adding the term $\frac{1}{2}\varepsilon\|\mathbf{x}-\mathbf{z}_0\|_2^2$ to our objective function and shifting the origin to $\mathbf{z}_0$. This results in a problem of the form

$$\text{minimize} \quad \tfrac{1}{2}\varepsilon\|\mathbf{x}\|_2^2+\tfrac{1}{2}\|\mathbf{Ax}-\mathbf{b}\|_2^2, \tag{1.3}$$

which is sometimes referred to as "regularization in standard form" or "Tikhonov's regularization." The singular value decomposition of A provides an insight into the nature of this problem as well as efficient computational procedures (see, e.g., Elden (1977)). The use of truncated singular value decomposition (SVD) solutions results in a closely related regularization technique (see, e.g., Varah (1979), Hansen (1987), or Chan and Hansen (1990)).

Although the traditional method of data fitting is by the least squares technique, in some applications it is desired to replace (1.2) with the general least norm problem

$$\text{minimize} \quad \|A\mathbf{x}-\mathbf{b}\|_p. \tag{1.4}$$

This raises the questions of how the regularization approach is extended to non-Euclidean norms, and what properties characterize regularized least norm problems. The current research is aimed at answering these questions. When $p\neq 2$ the SVD of A is not quite as helpful and the answers require different tools. It is also worth mentioning that (1.1) is not the only way to extend the regularization approach. The "natural" way is, perhaps, to consider the problem

$$\text{minimize} \quad \varepsilon\|x\|_s^s/s+\|A\mathbf{x}-\mathbf{b}\|_p^p/p. \tag{1.5}$$

This formulation coincides with Tikhonov's regularization when $s=p=2$, and has some computational advantage (see § 6). Nevertheless, as we now show, there are good reasons to start our investigation with (1.1).

The interest that we have in (1.1) is stimulated by the following observations. Recently, Dax (1991) proved that the dual of the regularized l_1 problem

$$\text{minimize} \quad (\varepsilon/2)\|\mathbf{x}\|_2^2+\|A\mathbf{x}-\mathbf{b}\|_1 \tag{1.6}$$

has the form

$$\begin{aligned} &\text{maximize} && \mathbf{b}^T\mathbf{y}-(\varepsilon/2)\|A^T\mathbf{y}/\varepsilon\|_2^2 \\ &\text{subject to} && \|\mathbf{y}\|_\infty\leqq 1, \end{aligned} \tag{1.7}$$

and if $\mathbf{y}$ solves the dual then the vector $A^T\mathbf{y}/\varepsilon$ solves the primal. Similarly, it is shown in Dax (1992a) that the dual of the regularized l_∞ problem

$$\text{minimize} \quad (\varepsilon/2)\|\mathbf{x}\|_2^2+\|A\mathbf{x}-\mathbf{b}\|_\infty \tag{1.8}$$

has the form

$$\begin{aligned} &\text{maximize} && \mathbf{b}^T\mathbf{y}-(\varepsilon/2)\|A^T\mathbf{y}/\varepsilon\|_2^2 \\ &\text{subject to} && \|\mathbf{y}\|_1\leqq 1, \end{aligned} \tag{1.9}$$

and if $\mathbf{y}$ solves the dual then the vector $A^T\mathbf{y}/\varepsilon$ solves the primal. Recall that for any given vector $\mathbf{y}=(y_1,\ldots,y_m)^T\in\mathbb{R}^m$ the norm $\|\mathbf{y}\|_p$ is a continuous function of p in the interval $[0,\infty)$, and

$$\lim_{p\to\infty}\|\mathbf{y}\|_p=\max_i|y_i|=\|\mathbf{y}\|_\infty.$$

Consequently, the dual index of ∞ is defined as 1, and vice versa. The two examples mentioned above suggest that the dual of the regularized l_p problem

$$\text{(1.10)}\qquad \text{minimize}\quad (\varepsilon/2)\|\mathbf{x}\|_2^2+\|A\mathbf{x}-\mathbf{b}\|_p$$

has the form

$$\text{(1.11)}\qquad \begin{aligned}&\text{maximize}\quad \mathbf{b}^T\mathbf{y}-(\varepsilon/2)\|A^T\mathbf{y}/\varepsilon\|_2^2\\ &\text{subject to}\quad \|\mathbf{y}\|_q\leqq 1,\end{aligned}$$

and if $\mathbf{y}$ solves the dual then $A^T\mathbf{y}/\varepsilon$ solves the primal. Assuming that this observation is true, it is tempting to make a further guess and to claim that the dual of (1.1) has the form

$$\text{(1.12)}\qquad \begin{aligned}&\text{maximize}\quad D(\mathbf{y})=\mathbf{b}^T\mathbf{y}-(\varepsilon/t)\|A^T\mathbf{y}/\varepsilon\|_t^t\\ &\text{subject to}\quad \|\mathbf{y}\|_q\leqq 1.\end{aligned}$$

Establishing this conjecture is a major objective of this paper. Another question that must be answered is how a primal solution is obtained from a dual solution when $s\neq 2$.

Both (1.6) and (1.8) can be viewed as special cases of the partially regularized linear programming problem

$$\text{(1.13)}\qquad \begin{aligned}&\text{minimize}\quad (\varepsilon/2)\sum_{j=1}^{k}x_j^2+\mathbf{c}^T\mathbf{x}\\ &\text{subject to}\quad A\mathbf{x}\geqq\mathbf{b},\end{aligned}$$

where $\mathbf{c}$ is a given vector in $\mathbb{R}^n$ and k is a positive integer such that $1\leqq k<n$. It is shown in Dax (1991) that the dual of this problem has the form

$$\text{(1.14)}\qquad \begin{aligned}&\text{maximize}\quad \mathbf{b}^T\mathbf{y}-(\varepsilon/2)\|(\hat{A}^T\mathbf{y}-\hat{\mathbf{c}})/\varepsilon\|_2^2\\ &\text{subject to}\quad \mathbf{y}\geqq\mathbf{0}\quad\text{and}\quad \tilde{A}^T\mathbf{y}=\tilde{\mathbf{c}},\end{aligned}$$

where A and $\mathbf{c}$ are split such that

$$A=[\hat{A},\tilde{A}],\quad \hat{A}\in\mathbb{R}^{m\times k},\quad \tilde{A}\in\mathbb{R}^{m\times(n-k)},\quad \mathbf{c}=\begin{pmatrix}\hat{\mathbf{c}}\\ \tilde{\mathbf{c}}\end{pmatrix},\quad \hat{\mathbf{c}}\in\mathbb{R}^k,\quad \tilde{\mathbf{c}}\in\mathbb{R}^{n-k}.$$

Moreover, let $\mathbf{y}$ solve (1.14) and define

$$\hat{\mathbf{x}}=(\hat{A}^T\mathbf{y}-\hat{\mathbf{c}})/\varepsilon;$$

then $\hat{\mathbf{x}}$ provides the first k components of the primal solution. This observation has been used to establish the duality properties of (1.6) and (1.8). If $k=n$ then (1.12) is transformed to the regularized linear programming problem

$$\text{(1.15)}\qquad \begin{aligned}&\text{minimize}\quad (\varepsilon/2)\|\mathbf{x}\|_2^2+\mathbf{c}^T\mathbf{x}\\ &\text{subject to}\quad A\mathbf{x}\geqq\mathbf{b}.\end{aligned}$$

In this case the duality theory of quadratic programming (see, e.g., Mangasarian (1969)) implies that the dual of (1.15) has the form

$$\text{(1.16)} \qquad \begin{aligned} &\text{maximize} \quad \mathbf{b}^T\mathbf{y}-(\varepsilon/2)\|(A^T\mathbf{y}-\mathbf{c})/\varepsilon\|_2^2 \\ &\text{subject to} \quad \mathbf{y} \geqq \mathbf{0}, \end{aligned}$$

and if $\mathbf{y}$ solves the dual then the vector $(A^T\mathbf{y}-\mathbf{c})/\varepsilon$ solves the primal. This relation forms the basis of a useful successive overrelaxation (SOR) method for large scale linear programming (see Mangasarian (1981)). However, neither (1.13) nor (1.15) are equivalent to (1.1), which forces us to look for a different approach.

Tikhonov (1965) showed that as ε tends to zero the solution of (1.3) tends to the minimum norm solution of (1.2). Yet the two solutions never coincide (unless both points lie at the origin). On the other hand, Mangasarian and Meyer (1979) proved that the regularized LP problem (1.15) has the following property. There exists a positive constant δ such that for any ε from the interval $(0, \delta]$ the solution of (1.15) coincides with the minimum (Euclidean) norm solution of the unregularized LP problem. These observations raise the question of whether (1.1) shares similar properties. The answer is given in the next section.

2. Optimality conditions. This section investigates the optimality condition of (1.1) and the behavior of the solution as ε moves from zero to infinity. We start by proving existence and uniqueness of the solution.

LEMMA 1. *Let S denote the set of all points that solve* (1.4) *and let* $\mathbf{x}^*$ *denote the unique solution of the problem*

$$\text{(2.1)} \qquad \begin{aligned} &\textit{minimize} \quad \|\mathbf{x}\|_s \\ &\textit{subject to} \quad \mathbf{x} \in S. \end{aligned}$$

The problem (1.1) *has a unique solution* $\hat{\mathbf{x}}$, *which satisfies*

$$\text{(2.2)} \qquad \|\hat{\mathbf{x}}\|_s \leqq \|\mathbf{x}^*\|_s.$$

Proof. Recall that S is a closed convex subset of $\mathbb{R}^n$. Hence the existence and uniqueness of $\mathbf{x}^*$ are ensured by the well-known Projection Theorem (e.g., Luenberger (1969)). Since $\mathbf{x}^*$ solves (1.4) the inequality

$$\text{(2.3)} \qquad \|A\mathbf{x}^*-\mathbf{b}\|_p \leqq \|A\mathbf{x}-\mathbf{b}\|_p$$

holds for all $\mathbf{x} \in \mathbb{R}^n$, and the level set

$$\{\mathbf{x} \mid F(\mathbf{x}) \leqq F(\mathbf{x}^*)\}$$

is contained in the compact ball

$$\{\mathbf{x} \mid \|\mathbf{x}\|_s \leqq \|\mathbf{x}^*\|_s\}.$$

Therefore, the existence and uniqueness of $\hat{x}$ are direct consequences of the fact that $F(\mathbf{x})$ is continuous and strictly convex in $\mathbb{R}^n$. □

Let $\nabla F(\mathbf{x})$ denote the gradient vector of $F(\mathbf{x})$ at a point $\mathbf{x} \in \mathbb{R}^n$. Then the jth component of $\nabla F(\mathbf{x})$ has the form

$$\text{(2.4)} \qquad \begin{aligned} \partial F(\mathbf{x})/\partial x_j &= \varepsilon|x_j|^{s-1}\operatorname{sign}(x_j) \\ &\quad + \left(\sum_{i=1}^m a_{ij}|\mathbf{a}_i^T\mathbf{x}-b_i|^{p-1}\operatorname{sign}(\mathbf{a}_i^T\mathbf{x}-b_i)\right) \Big/ \|A\mathbf{x}-\mathbf{b}\|_p^{p-1}, \end{aligned}$$

which shows that $F(\mathbf{x})$ is continuously differentiable at any point $\mathbf{x} \in \mathbb{R}^n$ in which $A\mathbf{x} \neq \mathbf{b}$. Moreover, if the system $A\mathbf{x}=\mathbf{b}$ is inconsistent then $F(\mathbf{x})$ is both strictly convex and continuously differentiable in $\mathbb{R}^n$. In this case $\hat{\mathbf{x}}$ is the only point in $\mathbb{R}^n$ which satisfies $\nabla F(\mathbf{x}) = \mathbf{0}$, while the fact that $\mathbf{x}^*$ solves (1.4) implies that the gradient vector of $\|A\mathbf{x}-\mathbf{b}\|_p$ vanishes at this point, which means that the jth component of $\nabla F(\mathbf{x}^*)$ equals $\varepsilon|x_j^*|^{s-1} \operatorname{sign}(x_j^*)$. If $\mathbf{x}^* = \mathbf{0}$, then clearly $\hat{\mathbf{x}} = 0$. Otherwise $\nabla F(\mathbf{x}^*) \neq \mathbf{0}$ and $\hat{\mathbf{x}} \neq \mathbf{x}^*$. Now the uniqueness of $\mathbf{x}^*$ and the inequality (2.2) imply

$$\|A\mathbf{x}^* - \mathbf{b}\|_p < \|A\hat{\mathbf{x}} - \mathbf{b}\|_p, \tag{2.5}$$

while the uniqueness of $\hat{\mathbf{x}}$ gives

$$\|\hat{\mathbf{x}}\|_s < \|\mathbf{x}^*\|_s. \tag{2.6}$$

It is also easy to verify that $\hat{\mathbf{x}} = 0$ implies $\mathbf{x}^* = \mathbf{0}$. Summarizing the above discussion we obtain the following results.

LEMMA 2. *Assume that the system $A\mathbf{x} = \mathbf{b}$ is inconsistent. In this case $\hat{\mathbf{x}}$ is the only point in $\mathbb{R}^n$ that satisfies $\nabla F(\mathbf{x}) = \mathbf{0}$. If $\mathbf{x}^* = \mathbf{0}$ then $\hat{\mathbf{x}} = \mathbf{0}$. Otherwise $\hat{\mathbf{x}} \neq \mathbf{0}$ and $\hat{\mathbf{x}}$ satisfies* (2.5) *and* (2.6).

Let us now consider the case when $A\mathbf{x}^* = \mathbf{b}$. Recall that $\|\mathbf{x}\|_s^s/s$ is continuously differentiable in $\mathbb{R}^n$ and let $\mathbf{g}$ denote the gradient vector of this function at $\mathbf{x}^*$. Here again $\mathbf{x}^* = \mathbf{0}$ implies $\hat{\mathbf{x}} = \mathbf{0}$. Hence in the forthcoming discussion we assume that $\mathbf{x}^* \neq \mathbf{0}$, which means that $\mathbf{g} \neq \mathbf{0}$. The convexity of $\|\mathbf{x}\|_s^s/s$ implies that the inequality

$$\|\mathbf{x}^* + \mathbf{u}\|_s^s/s \geqq \|\mathbf{x}^*\|_s^s/s + \mathbf{g}^T\mathbf{u} \tag{2.7}$$

holds for all $\mathbf{u} \in \mathbb{R}^n$ (see, e.g., Mangasarian (1969) or McCormick (1983)). Hence, by using the equality $A\mathbf{x}^* = \mathbf{b}$, we obtain that

$$F(\mathbf{x}^* + \mathbf{u}) \geqq F(\mathbf{x}^*) + \varepsilon\mathbf{g}^T\mathbf{u} + \|A\mathbf{u}\|_p \tag{2.8}$$

for all $\mathbf{u} \in \mathbb{R}^n$. If the inequality

$$\varepsilon\mathbf{g}^T\mathbf{u} + \|A\mathbf{u}\|_p < 0 \tag{2.9}$$

has no solution, then clearly $\mathbf{x}^*$ solves (1.1). On the other hand, if (2.9) has a solution $\mathbf{u}^*$, then the differentiability of $\|\mathbf{x}\|_s^s/s$ implies the existence of a positive constant $\eta > 0$ such that

$$F(\mathbf{x}^* + \theta\mathbf{u}^*) < F(\mathbf{x}^*)$$

for all $\theta \in (0, \eta)$. The question of whether (2.9) has a solution is characterized by the following theorem of the alternative, whose proof is given in Dax (1990).

THEOREM 3. *Either the inequality* (2.9) *has a solution $\mathbf{u} \in \mathbb{R}^n$, or the system*

$$A^T\mathbf{y} = \varepsilon\mathbf{g} \quad \textit{and} \quad \|\mathbf{y}\|_q \leqq 1 \tag{2.10}$$

has a solution $\mathbf{y} \in \mathbb{R}^m$, but never both.

A further result of Dax (1990) enables us to decide which of the two systems is solvable.

THEOREM 4. *Let $\mathbf{y}^*$ solve the problem*

$$\begin{aligned} &\textit{minimize} \quad \|A^T\mathbf{y} - \varepsilon\mathbf{g}\|_2^2 \\ &\textit{subject to} \quad \|\mathbf{y}\|_q \leqq 1, \end{aligned} \tag{2.11}$$

and let

$$\mathbf{u}^* = A^T\mathbf{y}^* - \varepsilon\mathbf{g} \tag{2.12}$$

denote the corresponding residual vector. If $\mathbf{u}^*=\mathbf{0}$ *then* $\mathbf{y}^*$ *solves* (2.11). *Otherwise* $\mathbf{u}^*$ *solves* (2.9). *Furthermore, in the second case* $\mathbf{u}^*$ *is the steepest descent vector of* $\varepsilon\mathbf{g}^T\mathbf{u}+\|A\mathbf{u}\|_p$ *at the origin point. That is, the vector* $\mathbf{u}^*/\|\mathbf{u}^*\|_2$ *solves the problem*

$$\text{(2.13)}\qquad \begin{aligned} &\textit{minimize} \quad \varepsilon\mathbf{g}^T\mathbf{u}+\|A\mathbf{u}\|_p\\ &\textit{subject to} \quad \|\mathbf{u}\|_2\leqq 1.\end{aligned}$$

COROLLARY 5. *Assume that* $A\mathbf{x}^*=\mathbf{b}$. *If* $\mathbf{x}^*=\mathbf{0}$ *or* $\mathbf{u}^*=\mathbf{0}$, *then* $\hat{\mathbf{x}}=\mathbf{x}^*$.

Another consequence of the equality $A\mathbf{x}^*=\mathbf{b}$ is that $\mathbf{x}^*$ solves the problem

$$\text{(2.14)}\qquad \begin{aligned} &\text{minimize} \quad \|\mathbf{x}\|_s^s/s\\ &\text{subject to} \quad A\mathbf{x}=\mathbf{b},\end{aligned}$$

while the optimality condition of (2.14) implies the existence of a vector $\mathbf{y}\in\mathbb{R}^*$ such that

$$\text{(2.15)}\qquad A^T\mathbf{y}=\mathbf{g}$$

(see, e.g., Mangasarian (1969) or McCormick (1983)). Hence if we define $\tilde{\mathbf{y}}$ to be the unique solution of the problem

$$\text{(2.16)}\qquad \begin{aligned} &\text{minimize} \quad \|\mathbf{y}\|_q\\ &\text{subject to} \quad A^T\mathbf{y}=\mathbf{g},\end{aligned}$$

then (2.11) can be written in the form

$$\text{(2.17)}\qquad \begin{aligned} &\text{minimize} \quad \|A^T(\mathbf{y}-\varepsilon\tilde{\mathbf{y}})\|_2^2\\ &\text{subject to} \quad \|\mathbf{y}\|_q\leqq 1,\end{aligned}$$

which shows that $\mathbf{u}^*=\mathbf{0}$ if and only if $\varepsilon\|\tilde{\mathbf{y}}\|_q\leqq 1$. Now Corollary 5 can be sharpened as follows.

THEOREM 6. *Assume that the system* $A\mathbf{x}=\mathbf{b}$ *is solvable. If* $\mathbf{x}^*=\mathbf{0}$ *or* $\varepsilon\leqq 1/\|\tilde{\mathbf{y}}\|_q$ *then* $\hat{\mathbf{x}}=\mathbf{x}^*$. *Otherwise* $A\hat{\mathbf{x}}\neq\mathbf{b}$, $0<\|\hat{\mathbf{x}}\|_s<\|\mathbf{x}^*\|_s$, *and* $\hat{\mathbf{x}}$ *is the only point in* $\mathbb{R}^n$ *that satisfies* $\nabla F(\mathbf{x})=\mathbf{0}$.

The last theorem provides a partial answer to the question of what happens when ε tends to zero. The next result completes the answer to this question.

THEOREM 7. *Let* $\{\varepsilon_k\}$ *be a sequence of positive numbers such that*

$$\text{(2.18)}\qquad \lim_{k\to\infty}\varepsilon_k=0,$$

and let $\mathbf{x}_k$ *denote the solution of the problem*

$$\text{(2.19)}\qquad \textit{minimize} \quad F_k(\mathbf{x})=\varepsilon_k\|\mathbf{x}\|_s^s/s+\|A\mathbf{x}-\mathbf{b}\|_p.$$

Then

$$\text{(2.20)}\qquad \lim_{k\to\infty}\mathbf{x}_k=\mathbf{x}^*.$$

Proof. If $\mathbf{x}^*=\mathbf{0}$ or $A\mathbf{x}^*=b$ then (2.20) is a direct consequence of Lemma 2 or Theorem 6. Hence it is sufficient to consider the case when $\mathbf{x}^*\neq 0$ and $A\mathbf{x}^*\neq\mathbf{b}$. In this case Lemma 2 implies that $\mathbf{x}_k$ satisfies the relations

$$\text{(2.21)}\qquad \|A\mathbf{x}_k-\mathbf{b}\|_p>\|A\mathbf{x}^*-\mathbf{b}\|_p>0,$$

$$\text{(2.22)}\qquad \|\mathbf{x}_k\|_s<\|\mathbf{x}^*\|_s,$$

and

$$\text{(2.23)}\qquad \varepsilon_k\mathbf{g}_k+\mathbf{h}_k=\mathbf{0},$$

where $\mathbf{g}_k$ and $\mathbf{h}_k$ denote the gradient vectors of $\|\mathbf{x}\|_s^s/s$ and $\|A\mathbf{x}-\mathbf{b}\|_p$, respectively, at the point $\mathbf{x}_k$. The inequality (2.22) indicates that the sequences $\{\mathbf{x}_k\}$ and $\{\mathbf{g}_k\}$ are bounded. Hence from (2.18) and (2.23) we obtain that

$$\lim_{k\to\infty} \mathbf{h}_k = 0, \tag{2.24}$$

which shows that any cluster point $\tilde{\mathbf{x}}$ of the sequence $\{\mathbf{x}_k\}$ solves (1.4). The existence of such a cluster point is ensured by the fact that the sequence $\{\mathbf{x}_k\}$ is bounded, while (2.22) implies $\|\tilde{\mathbf{x}}\|_s \leqq \|\mathbf{x}^*\|_s$. We have proved, therefore, that any cluster point of the sequence $\{\mathbf{x}_k\}$ converges to $\mathbf{x}^*$. Thus, since this sequence is bounded, the whole sequence converges to $\mathbf{x}^*$. □

The next observation clarifies the situation when ε tends to infinity.

THEOREM 8. *Let $\{\varepsilon_k\}$ be a sequence of positive numbers such that*

$$\lim_{k\to\infty} \varepsilon_k = \infty, \tag{2.25}$$

and let $\mathbf{x}_k$ denote the solution of (2.19). *Then*

$$\lim_{k\to\infty} \mathbf{x}_k = \mathbf{0}. \tag{2.26}$$

Proof. If $\mathbf{x}^* = \mathbf{0}$, then $\mathbf{x}_k = \mathbf{0}$ for all k and the claim is straightforward. Otherwise, when $\mathbf{x}^* \neq \mathbf{0}$, Theorem 6 and Lemma 2 allow us to assume that relations (2.21), (2.22), and (2.23) hold. Thus, by combining (2.23) and the fact that the sequence $\{\mathbf{h}_k\}$ is bounded, we deduce that

$$\lim_{k\to\infty} \mathbf{g}_k = \mathbf{0}, \tag{2.27}$$

which implies (2.26). □

We shall finish this section by showing that the solution of (1.1) is a continuous function of ε.

THEOREM 9. *Let $\{\varepsilon_k\}$ be a sequence of positive numbers such that*

$$\lim_{k\to\infty} \varepsilon_k = \varepsilon, \tag{2.28}$$

and let $\mathbf{x}_k$ denote the solution of (2.19). *Then*

$$\lim_{k\to\infty} \mathbf{x}_k = \hat{\mathbf{x}}, \tag{2.29}$$

where, as before, $\hat{\mathbf{x}}$ denotes the solution of (1.1).

Proof. Here again, if $\mathbf{x}^* = \mathbf{0}$, then $\mathbf{x}_k = \mathbf{0}$ for all k and $\hat{\mathbf{x}} = \mathbf{0}$, so the claim is straightforward. Otherwise Lemma 2 and Theorem 6 allow us to assume that relations (2.21), (2.22), and (2.23) hold for all k. However, this time we must show that any cluster point of the sequence $\{\mathbf{x}_k\}$ solves (1.1). In fact, since $F(\mathbf{x})$ is strictly convex and continuous, it is sufficient to prove that

$$\lim_{k\to\infty} F(\mathbf{x}_k) = F(\hat{\mathbf{x}}). \tag{2.30}$$

Recall that the gradient vector of $F(\mathbf{x})$ at the point $\mathbf{x}_k$ has the form

$$\nabla F(\mathbf{x}_k) = \varepsilon \mathbf{g}_k + \mathbf{h}_k. \tag{2.31}$$

Hence by subtracting (2.23) from (2.31) we obtain

$$\nabla F(\mathbf{x}_k) = (\varepsilon - \varepsilon_k)\mathbf{g}_k. \tag{2.32}$$

On the other hand, the convexity of $F(\mathbf{x})$ yields

$$F(\hat{\mathbf{x}}) \geqq F(\mathbf{x}_k) + (\hat{\mathbf{x}} - \mathbf{x}_k)^T \nabla F(\mathbf{x}_k) \tag{2.33}$$

and

$$0 < F(\mathbf{x}_k) - F(\hat{\mathbf{x}}) \leqq (\mathbf{x}_k - \hat{\mathbf{x}})^T \nabla F(\mathbf{x}_k) = (\varepsilon - \varepsilon_k)\mathbf{g}_k^T(\mathbf{x}_k - \hat{\mathbf{x}}). \tag{2.34}$$

Therefore, since the sequences $\{\mathbf{x}_k\}$ and $\{\mathbf{g}_k\}$ are bounded, (2.34) implies (2.30). □

3. Duality. In this section we show that the dual of (1.1) has the form

$$\begin{aligned} &\text{maximize} \quad D(\mathbf{y}) = \mathbf{b}^T\mathbf{y} - (\varepsilon/t)\|A^T\mathbf{y}/\varepsilon\|_t^t \\ &\text{subject to} \quad \|\mathbf{y}\|_q \leqq 1. \end{aligned} \tag{3.1}$$

To prove this assertion we need some further notation. Let $\mathbf{c}_j = (a_{1j}, \ldots, a_{mj})^T$ denote the jth column of A, and let

$$\mathbf{z}(\mathbf{y}) = (z_1(\mathbf{y}), \ldots, z_n(\mathbf{y}))^T \tag{3.2}$$

be a vector function whose jth component is

$$\mathbf{z}_j(\mathbf{y}) = |\mathbf{c}_j^T\mathbf{y}|^{t-1} \operatorname{sign}(\mathbf{c}_j^T\mathbf{y}). \tag{3.3}$$

Then $A\mathbf{z}(\mathbf{y})/\varepsilon^{t-1}$ is the gradient vector of $(\varepsilon/t)\|A^T\mathbf{y}/\varepsilon\|_t^t$,

$$\|\mathbf{z}(\mathbf{y})/\varepsilon^{t-1}\|_s^s = \|A^T\mathbf{y}/\varepsilon\|_t^t, \tag{3.4}$$

and

$$\mathbf{y}^T A\mathbf{z}(\mathbf{y})/\varepsilon^{t-1} = \varepsilon\|A^T\mathbf{y}/\varepsilon\|_t^t. \tag{3.5}$$

Similarly, we use

$$\mathbf{g}(\mathbf{x}) = (g_1(\mathbf{x}), \ldots, g_n(\mathbf{x}))^T \tag{3.6}$$

to denote the gradient vector of $\|\mathbf{x}\|_s^s/s$. That is,

$$g_j(\mathbf{x}) = |x_j|^{s-1} \operatorname{sign}(x_j). \tag{3.7}$$

Now we can verify that

$$\mathbf{g}(\mathbf{z}(\mathbf{y})/\varepsilon^{t-1}) = A^T\mathbf{y}/\varepsilon. \tag{3.8}$$

In other words, $A^T\mathbf{y}/\varepsilon$ is the gradient vector of $\|\mathbf{x}\|_s^s/s$ at the point $\mathbf{x} = \mathbf{z}(\mathbf{y})/\varepsilon^{t-1}$. Another useful vector function is

$$\mathbf{r}(\mathbf{x}) = (r_1(\mathbf{x}), \ldots, r_m(\mathbf{x}))^T, \tag{3.9}$$

where

$$r_i(\mathbf{x}) = |\mathbf{a}_i^T\mathbf{x} - b_i|^{p-1} \operatorname{sign}(\mathbf{a}_i^T\mathbf{x} - b_i)/(\|A\mathbf{x} - \mathbf{b}\|_p)^{p-1}. \tag{3.10}$$

With these notations at hand

$$\nabla F(\mathbf{x}) = \varepsilon\mathbf{g}(\mathbf{x}) + A^T\mathbf{r}(\mathbf{x}), \tag{3.11}$$

$$\|\mathbf{r}(\mathbf{x})\|_q = 1, \tag{3.12}$$

and

$$(A\mathbf{x} - \mathbf{b})^T\mathbf{r}(\mathbf{x}) = \|A\mathbf{x} - \mathbf{b}\|_p. \tag{3.13}$$

Finally we introduce the vector function

$$\mathbf{w}(\mathbf{y}) = (w_1(\mathbf{y}), \ldots, w_m(\mathbf{y}))^T, \tag{3.14}$$

whose ith component is

$$w_i(\mathbf{y}) = |y_i|^{q-1} \operatorname{sign}(y_i)/(\|\mathbf{y}\|_q)^{q-1}. \tag{3.15}$$

That is, $\mathbf{w}(\mathbf{y})$ is the gradient vector of the function $\|\mathbf{y}\|_q$. Here one can verify that

$$\|\mathbf{w}(\mathbf{y})\|_p = 1 \tag{3.16}$$

and

$$\mathbf{y}^T\mathbf{w}(\mathbf{y}) = \|\mathbf{y}\|_q. \tag{3.17}$$

THEOREM 10. *Let* $\hat{\mathbf{y}}$ *solve* (3.1). *Then the vector*

$$\hat{\mathbf{x}} = \mathbf{z}(\hat{\mathbf{y}})/\varepsilon^{t-1} = \mathbf{z}(\hat{\mathbf{y}}/\varepsilon) \tag{3.18}$$

solves (1.1). *Conversely, let* $\hat{\mathbf{x}}$ *solve* (1.1) *and assume that* $A\hat{\mathbf{x}} \neq b$. *Then the vector*

$$\hat{\mathbf{y}} = -\mathbf{r}(\hat{\mathbf{x}}) \tag{3.19}$$

solves (3.1). *In both cases we have*

$$D(\hat{\mathbf{y}}) = F(\hat{\mathbf{x}}), \tag{3.20}$$

and the primal-dual inequality

$$D(\mathbf{y}) \leqq F(\mathbf{x}) \tag{3.21}$$

holds for all $\mathbf{x} \in \mathbb{R}^n$ *and* $\mathbf{y} \in \mathbb{R}^m$ *such that* $\|\mathbf{y}\|_q \leqq 1$.

Proof. The first part of the proof considers the case in which the dual objective function, $D(\mathbf{y})$, has a maximizer $\tilde{\mathbf{y}}$ such that $\|\tilde{\mathbf{y}}\|_q \leqq 1$. In this case any point $\hat{\mathbf{y}} \in \mathbb{R}^m$ that solves (3.1) is also a maximizer of $D(\mathbf{y})$ and the gradient vector of $D(\mathbf{y})$ vanishes at this point. That is,

$$\mathbf{b} - A\mathbf{z}(\hat{\mathbf{y}})/\varepsilon^{t-1} = \mathbf{0}, \tag{3.22}$$

which means that the vector $\hat{\mathbf{x}} = \mathbf{z}(\hat{\mathbf{y}})/\varepsilon^{t-1}$ satisfies both $A\hat{\mathbf{x}} = \mathbf{b}$ and

$$F(\hat{\mathbf{x}}) = (\varepsilon/s)\|\hat{\mathbf{x}}\|_s^s = (\varepsilon/s)\|\mathbf{z}(\hat{\mathbf{y}})/\varepsilon^{t-1}\|_s^s = (\varepsilon/s)\|A^T\hat{\mathbf{y}}/\varepsilon\|_t^t. \tag{3.23}$$

On the other hand, multiplying (3.22) by $\hat{\mathbf{y}}^T$ gives

$$\begin{aligned} 0 &= \mathbf{b}^T\hat{\mathbf{y}} - \hat{\mathbf{y}}^TA\mathbf{z}(\hat{\mathbf{y}})/\varepsilon^{t-1} = \mathbf{b}^T\hat{\mathbf{y}} - \varepsilon\|A^T\hat{\mathbf{y}}/\varepsilon\|_t^t \\ &= \mathbf{b}^T\hat{\mathbf{y}} - (\varepsilon/t)\|A^T\hat{\mathbf{y}}/\varepsilon\|_t^t - (\varepsilon/s)\|A^T\hat{\mathbf{y}}/\varepsilon\|_t^t = D(\hat{\mathbf{y}}) - F(\hat{\mathbf{x}}). \end{aligned} \tag{3.24}$$

Moreover, combining (3.8) with $\|\hat{\mathbf{y}}\|_q \leqq 1$ shows (as in Corollary 5) that $\hat{\mathbf{x}}$ solves (1.1).

Before starting the second part of the proof we make a further comment on the first case. Assume that $\hat{\mathbf{x}}$ solves (1.1) and that $A\hat{\mathbf{x}} = \mathbf{b}$. Then a further use of the proof of Corollary 5 shows the existence of a vector $\hat{\mathbf{y}}$ such that $\|\hat{\mathbf{y}}\|_q \leqq 1$ and

$$A^T\hat{\mathbf{y}} = \varepsilon\mathbf{g}(\hat{\mathbf{x}}). \tag{3.25}$$

On the other hand (3.8) yields

$$A^T\hat{\mathbf{y}} = \varepsilon\mathbf{g}(\mathbf{z}(\hat{\mathbf{y}})/\varepsilon^{t-1}). \tag{3.26}$$

Hence by comparing these equalities we deduce that

$$\hat{\mathbf{x}} = \mathbf{z}(\hat{\mathbf{y}})/\varepsilon^{t-1}, \tag{3.27}$$

and that $\hat{\mathbf{y}}$ is a maximizer of $D(\mathbf{y})$. In other words, the dual objective function has a maximizer $\tilde{\mathbf{y}}$ such that $\|\tilde{\mathbf{y}}\|_q \leqq 1$ if and only if the solution of (1.1) satisfies $A\hat{\mathbf{x}} = \mathbf{b}$.

The second part of the proof considers the case in which the dual objective function has no maximizer $\tilde{\mathbf{y}}$ such that $\|\tilde{\mathbf{y}}\|_q \leqq 1$. In this case a point $\hat{\mathbf{y}} \in \mathbb{R}^m$ solves (3.1) if and only if

$$\|\hat{\mathbf{y}}\|_q = 1 \tag{3.28}$$

and there exists a positive constant λ such that

$$\mathbf{b} - A\mathbf{z}(\hat{\mathbf{y}})/\varepsilon^{t-1} = \lambda \mathbf{w}(\hat{\mathbf{y}}). \tag{3.29}$$

(These are the well-known Kuhn-Tucker optimality conditions.) Hence by applying (3.16) and (3.18) we deduce that

$$\lambda = \lambda \|\mathbf{w}(\hat{\mathbf{y}})\|_p = \|\lambda \mathbf{w}(\hat{\mathbf{y}})\|_p = \|\mathbf{b} - A\mathbf{z}(\hat{\mathbf{y}})/\varepsilon^{t-1}\|_p = \|\mathbf{b} - A\hat{\mathbf{x}}\|_p. \tag{3.30}$$

It follows, therefore, that if $\|\hat{\mathbf{y}}\|_q = 1$ and $\hat{\mathbf{x}} = \mathbf{z}(\hat{\mathbf{y}})/\varepsilon^{t-1}$, then $\hat{\mathbf{y}}$ solves (3.1) if and only if

$$\mathbf{w}(\hat{\mathbf{y}}) = -(A\hat{\mathbf{x}} - \mathbf{b})/\|A\hat{\mathbf{x}} - \mathbf{b}\|_p. \tag{3.31}$$

Moreover, multiplying (3.29) by $\mathbf{y}^T$ gives

$$\mathbf{b}^T\hat{\mathbf{y}} - \varepsilon \|A^T\hat{\mathbf{y}}/\varepsilon\|_t^t = \lambda = \|A\hat{\mathbf{x}} - \mathbf{b}\|_p \tag{3.32}$$

and

$$\begin{aligned} \mathbf{b}^T\hat{\mathbf{y}} - (\varepsilon/t)\|A^T\hat{\mathbf{y}}/\varepsilon\|_t^t &= (\varepsilon/s)\|A^T\hat{\mathbf{y}}/\varepsilon\|_t^t + \|A\hat{\mathbf{x}} - \mathbf{b}\|_p \\ &= (\varepsilon/s)\|\mathbf{z}(\hat{\mathbf{y}})/\varepsilon^{t-1}\|_s^s + \|A\hat{\mathbf{x}} - \mathbf{b}\|_p \\ &= (\varepsilon/s)\|\hat{\mathbf{x}}\|_s^s + \|A\hat{\mathbf{x}} - \mathbf{b}\|_p. \end{aligned} \tag{3.33}$$

It remains to show that if $\hat{\mathbf{x}}$ solves (1.1) and $A\hat{\mathbf{x}} \neq \mathbf{b}$ then the vector $\hat{\mathbf{y}} = -\mathbf{r}(\hat{\mathbf{x}})$ solves (3.1). On one hand we have

$$\varepsilon \mathbf{g}(\hat{\mathbf{x}}) + A^T\mathbf{r}(\hat{\mathbf{x}}) = \mathbf{0} \tag{3.34}$$

and

$$\mathbf{g}(\hat{\mathbf{x}}) = A^T(-\mathbf{r}(\hat{\mathbf{x}}))/\varepsilon = A^T\hat{\mathbf{y}}/\varepsilon, \tag{3.35}$$

while on the other hand (3.8) yields

$$\mathbf{g}(\mathbf{z}(\hat{\mathbf{y}})/\varepsilon^{t-1}) = A^T\hat{\mathbf{y}}/\varepsilon. \tag{3.36}$$

Thus, by comparing (3.35) with (3.36), we deduce that

$$\hat{\mathbf{x}} = \mathbf{z}(\hat{\mathbf{y}})/\varepsilon^{t-1}. \tag{3.37}$$

In addition, (3.12) gives

$$\|\hat{\mathbf{y}}\|_q = \|\mathbf{r}(\hat{\mathbf{x}})\|_q = 1, \tag{3.38}$$

and the definitions of $\mathbf{w}(\mathbf{y})$ and $\mathbf{r}(\mathbf{x})$ lead to

$$\mathbf{w}(\hat{\mathbf{y}}) = \mathbf{w}(-\mathbf{r}(\hat{\mathbf{x}})) = -(A\hat{\mathbf{x}} - \mathbf{b})/\|A\hat{\mathbf{x}} - \mathbf{b}\|_p. \tag{3.39}$$

Hence, as we noted above, relations (3.37), (3.38), and (3.39) imply that $\hat{\mathbf{y}}$ solves (3.1). □

Another consequence of the above relations is that the vectors $\hat{\mathbf{y}}$ and $\mathbf{b} - A\hat{\mathbf{x}}$ are "aligned." That is,

$$\hat{\mathbf{y}}^T(\mathbf{b} - A\hat{\mathbf{x}}) = \|\hat{\mathbf{y}}\|_q \|\mathbf{b} - A\hat{\mathbf{x}}\|_p. \tag{3.40}$$

It is also worthwhile to mention that the dual of the problem

$$\begin{aligned} &\text{minimize} && (\varepsilon/s)\|\mathbf{x}\|_s^s + \|A\mathbf{x}-\mathbf{b}\|_p \\ &\text{subject to} && \mathbf{a}_i^T\mathbf{x} \geqq b_i \quad \text{for } i = m+1, \dots, \hat{m} \\ &\quad\text{and} && \mathbf{a}_i^T\mathbf{x} = b_i \quad \text{for } i = \hat{m}+1, \dots, \tilde{m} \end{aligned} \tag{3.41}$$

has the form

$$\begin{aligned} &\text{maximize} && \tilde{\mathbf{b}}^T\mathbf{y} - (\varepsilon/t)\|\tilde{A}^T\mathbf{y}/\varepsilon\|_t^t \\ &\text{subject to} && \|\mathbf{y}\|_q \leqq 1 \quad \text{and} \quad y_i \geqq 0 \quad \text{for } i = m+1, \dots, \hat{m}, \end{aligned} \tag{3.42}$$

where $\mathbf{y} = (y_1, \dots, y_{\tilde{m}})^T \in \mathbb{R}^{\tilde{m}}$, $\tilde{\mathbf{b}} = (b_1, \dots, b_{\tilde{m}})^T \in \mathbb{R}^{\tilde{m}}$, and $\tilde{A}$ is an $\tilde{m} \times n$ matrix whose rows are $\mathbf{a}_i^T$, $i = 1, \dots, \tilde{m}$. The proof of this observation is similar to that of Theorem 10.

4. Iterative improvement of regularized solutions. This section presents and analyzes a simple iterative method that can be used to "improve" the solution of the regularized least norm problem. We have seen that one motivation for solving the regularized problem is the existence of an initial estimate $\mathbf{z}_0$. However, when the regularized problem is solved we obtain a new estimate $\mathbf{z}_1$. Thus, by shifting the origin to $\mathbf{z}_1$, it is possible to construct an "improved" regularized problem and "improved" solution $\mathbf{z}_2$. Similarly, $\mathbf{z}_2$ can be used to generate $\mathbf{z}_3$ and so forth. The exact formulation of this process is as follows. Given $\mathbf{z}_k$, we define

$$\mathbf{b}_k = \mathbf{b} - A\mathbf{z}_k \tag{4.1}$$

and calculate $\mathbf{x}_k$, the unique solution of the problem

$$\text{minimize} \quad (\varepsilon/s)\|\mathbf{x}\|_s^s + \|A\mathbf{x} - \mathbf{b}_k\|_p. \tag{4.2}$$

Then the next point is defined as

$$\mathbf{z}_{k+1} = \mathbf{z}_k + \mathbf{x}_k. \tag{4.3}$$

The aim of this section is to answer the question of whether the sequence $\{\mathbf{z}_k\}$ converges and what properties the limit point has, if it exists.

The definition of $\mathbf{z}_{k+1}$ implies the inequality

$$\|A\mathbf{z}_k - \mathbf{b}\|_p \geqq (\varepsilon/s)\|\mathbf{x}_k\|_s^s + \|A\mathbf{z}_{k+1} - \mathbf{b}\|_p, \tag{4.4}$$

which shows that the sequence $\{\|A\mathbf{z}_k - \mathbf{b}\|_p\}$ is monotonic decreasing and bounded from below. Consequently, this sequence converges,

$$\lim_{k\to\infty} \mathbf{x}_k = \mathbf{0}, \tag{4.5}$$

and

$$\lim_{k\to\infty} \mathbf{g}(\mathbf{x}_k) = \mathbf{0}. \tag{4.6}$$

If there exists an index k_0 for which $\mathbf{z}_{k_0}$ solves (1.4), then clearly $\mathbf{z}_k = \mathbf{z}_{k_0}$ for all $k \geqq k_0$. Hence there is no loss of generality in assuming that such an index does not exist, and that $A\mathbf{z}_k \neq \mathbf{b}$ for all k. Now the fact that $\mathbf{x}_k$ solves (4.2) indicates that

$$\varepsilon\mathbf{g}(\mathbf{x}_k) + A^T\mathbf{r}(\mathbf{z}_{k+1}) = \mathbf{0} \tag{4.7}$$

and

$$\lim_{k\to\infty} A^T\mathbf{r}(\mathbf{z}_k) = \mathbf{0}. \tag{4.8}$$

A further justification for applying this process lies in the following observation.

THEOREM 11. *Let* $\tilde{\mathbf{x}}$ *solve* (1.4). *Then*

$$\lim_{k\to\infty} \|A\mathbf{z}_k - \mathbf{b}\|_p = \|A\tilde{\mathbf{x}} - \mathbf{b}\|_p. \tag{4.9}$$

Proof. Let α denote the limit of the sequence $\{\|A\mathbf{z}_k - \mathbf{b}\|_p\}$. If $\alpha = 0$ then the claim is straightforward. Therefore, it is possible to assume that $\alpha > 0$. Since $\|\mathbf{r}(\mathbf{z}_k)\|_q = 1$ the sequence $\{\mathbf{r}(\mathbf{z}_k)\}$ has at least one cluster point, say $\mathbf{r}^* = (r_1^*, \dots, r_m^*)^T \in \mathbb{R}^m$. Let $\{\mathbf{z}_{k_j}\}$ be a subsequence of $\{\mathbf{z}_k\}$ such that

$$\lim_{j\to\infty} \mathbf{r}(\mathbf{z}_{k_j}) = \mathbf{r}^*. \tag{4.10}$$

Then the limit

$$\lim_{j\to\infty} |\mathbf{a}_i^T\mathbf{z}_{k_j} - b_i|^{p-1} \operatorname{sign}(\mathbf{a}_i^T\mathbf{z}_{k_j} - b_i)/\alpha^{p-1} = r_i^* \tag{4.11}$$

implies that the sequence $\{\mathbf{a}_i^T\mathbf{z}_{k_j} - b_i\}$ converges. Consequently, the vector

$$\tilde{\mathbf{r}} = \lim_{j\to\infty} A\mathbf{z}_{k_j} - \mathbf{b} \tag{4.12}$$

is well defined and the least squares problem

$$\text{minimize} \quad \|A\mathbf{z} - \mathbf{b} - \tilde{\mathbf{r}}\|_2^2 \tag{4.13}$$

has a solution $\tilde{\mathbf{z}}$ such that

$$A\tilde{\mathbf{z}} - \mathbf{b} = \tilde{\mathbf{r}}. \tag{4.14}$$

Further use of (4.12) gives

$$\lim_{k\to\infty} \|A\mathbf{z}_k - \mathbf{b}\|_p = \|A\tilde{\mathbf{z}} - \mathbf{b}\|_p, \tag{4.15}$$

while from (4.8) we deduce that

$$A^T\mathbf{r}(\tilde{\mathbf{z}}) = \lim_{j\to\infty} A^T\mathbf{r}(\mathbf{z}_{k_j}) = \mathbf{0}. \tag{4.16}$$

That is, $\tilde{\mathbf{z}}$ solves (1.4). □

It is tempting to conjecture that the sequence $\{\mathbf{z}_k\}$ converges to $\tilde{\mathbf{z}}$. Indeed, if the columns of A are linearly independent then this assertion is a direct consequence of the fact that $\|A\mathbf{x} - \mathbf{b}\|_p$ is strictly convex. However, when the columns of A are linearly dependent this argument does not work. In this case (1.4) has infinitely many solutions but the corresponding residual vector is unique. That is, $A\tilde{\mathbf{z}} - \mathbf{b} = A\hat{\mathbf{z}} - b$ for any pair of solutions. (This observation is a direct corollary of the well-known Minkowski's inequality.) An equivalent way to express the uniqueness of the residual vector is the following. Let $\tilde{\mathbf{z}}$ solve (1.4). Then $\tilde{\mathbf{z}}$ can be written in the form

$$\tilde{\mathbf{z}} = \tilde{\mathbf{u}} + \tilde{\mathbf{v}}, \tag{4.17}$$

where $\tilde{\mathbf{u}} \in \operatorname{Null}(A)$ and $\tilde{\mathbf{v}} \in \operatorname{Range}(A^T)$, and any other solution $\hat{\mathbf{z}}$ of (1.4) has the form $\hat{\mathbf{z}} = \mathbf{u} + \tilde{\mathbf{v}}$ for some $\mathbf{u} \in \operatorname{Null}(A)$. Note that $\tilde{\mathbf{v}}$ is the unique solution of the problem

$$\begin{aligned} &\text{minimize} && \|\mathbf{x}\|_2 \\ &\text{subject to} && \mathbf{x} \in S. \end{aligned} \tag{4.18}$$

We can also verify that $\|A\mathbf{x}-\mathbf{b}\|_p$ is strictly convex on Range (A^T). Hence, if we have a sequence $\{\mathbf{v}_k\}$ of vectors in Range (A^T), then the limit

$$\lim_{k\to\infty} \|A\mathbf{v}_k-\mathbf{b}\|_p = \|A\tilde{\mathbf{v}}-\mathbf{b}\|_p \tag{4.19}$$

implies that

$$\lim_{k\to\infty} \mathbf{v}_k = \tilde{\mathbf{v}}. \tag{4.20}$$

In particular, let us present $\mathbf{z}_k$ in the form

$$\mathbf{z}_k = \mathbf{u}_k + \mathbf{v}_k, \tag{4.21}$$

where $\mathbf{u}_k \in \text{Null}\,(A)$ and $\mathbf{v}_k \in \text{Range}\,(A^T)$. Then the limit (4.15) leads to (4.19) and (4.20). The question of whether the sequence $\{\mathbf{u}_k\}$ converges is not easy to answer. However, in the important case where $s=2$, equality (4.7) is reduced to

$$\varepsilon \mathbf{x}_k + A^T\mathbf{r}(\mathbf{z}_k) = \mathbf{0}, \tag{4.22}$$

which means that

$$\mathbf{u}_k = \mathbf{u}_0 \tag{4.23}$$

and

$$\mathbf{v}_k = \mathbf{v}_0 - \sum_{l=1}^{k} A^T\mathbf{r}(\mathbf{z}_l)/\varepsilon \tag{4.24}$$

for $k=1,2,\dots$.

COROLLARY 12. *Assume that $s=2$. In this case the sequence $\{\mathbf{z}_k\}$ converges to the point $\mathbf{u}_0+\tilde{\mathbf{v}}$. Moreover, if $\mathbf{z}_0 \in$ Range (A^T), it converges to $\tilde{\mathbf{v}}$, the unique solution of* (4.18).

5. Dual penalty function methods. The practical value of Theorem 10 is that it gives an alternative way to solve (1.1). Let us consider, for example, the regularized l_1 problem (1.6). In this case the dual variables have simple bounds, and the dual can be solved by applying a wide range of methods. In particular, when A is large, sparse, and unstructured, we can use a row relaxation scheme that resembles Kaczmarz's method (see Dax (1991)). However, when $p>1$ the inequality $\|\mathbf{y}\|_q \leqq 1$ introduces a certain difficulty into the solution of the dual problem. One way to remove this obstacle is by using a penalty function.

THEOREM 13. *Let $\tilde{\mathbf{y}}$ solve the problem*

$$\textit{maximize} \quad \mathbf{b}^T\mathbf{y} - (\varepsilon/t)\|A^T\mathbf{y}/\varepsilon\|_t^t - \tfrac{1}{2}(\max\{0, \|\mathbf{y}\|_q - 1\})^2/\lambda, \tag{5.1}$$

where λ is a given positive constant. Then the vector

$$\hat{\mathbf{y}} = \tilde{\mathbf{y}}/\max\{1, \|\tilde{\mathbf{y}}\|_q\} \tag{5.2}$$

solves the problem

$$\begin{aligned} &\textit{maximize} \quad \mathbf{b}^T\mathbf{y} - (\nu/t)\|A^T\mathbf{y}/\nu\|_t^t \\ &\textit{subject to} \quad \|\mathbf{y}\|_q \leqq 1, \end{aligned} \tag{5.3}$$

where

$$\nu = \varepsilon/\max\{1, \|\tilde{\mathbf{y}}\|_q\}. \tag{5.4}$$

Furthermore, define

$$\hat{\mathbf{x}} = \mathbf{z}(\hat{\mathbf{y}})/\nu^{t-1} = \mathbf{z}(\tilde{\mathbf{y}})/\varepsilon^{t-1}. \tag{5.5}$$

Then

$$\nu = \varepsilon/(1+\lambda\|A\hat{\mathbf{x}}-\mathbf{b}\|_p) \tag{5.6}$$

and $\hat{\mathbf{x}}$ *solves the problem*

$$\textit{minimize} \quad (\nu/s)\|\mathbf{x}\|_s^s + \|A\mathbf{x}-\mathbf{b}\|_p. \tag{5.7}$$

Proof. If $\|\tilde{\mathbf{y}}\|_q \leqq 1$ then Theorem 13 is a special case of Theorem 10. Hence it is sufficient to consider the case when $\|\tilde{\mathbf{y}}\|_q > 1$. In this case the optimality conditions of (5.1) indicate that $\tilde{\mathbf{y}}$ satisfies

$$\mathbf{b} - A\mathbf{z}(\tilde{\mathbf{y}})/\varepsilon^{t-1} = (\|\tilde{\mathbf{y}}\|_q - 1)\mathbf{w}(\tilde{\mathbf{y}})/\lambda, \tag{5.8}$$

while (3.16) gives

$$(\|\tilde{\mathbf{y}}\|_q - 1)/\lambda = \|\mathbf{b} - A\mathbf{z}(\tilde{\mathbf{y}})/\varepsilon^{t-1}\|_p = \|A\hat{\mathbf{x}}-\mathbf{b}\|_p \tag{5.9}$$

and

$$\|\tilde{\mathbf{y}}\|_q = 1 + \lambda\|A\hat{\mathbf{x}}-\mathbf{b}\|_p. \tag{5.10}$$

Now the equality

$$\mathbf{w}(\hat{\mathbf{y}}) = \mathbf{w}(\tilde{\mathbf{y}}) \tag{5.11}$$

enables us to rewrite (5.8) in the form

$$\mathbf{b} - A\hat{\mathbf{x}} = \|A\hat{\mathbf{x}}-\mathbf{b}\|_p\mathbf{w}(\hat{\mathbf{y}}). \tag{5.12}$$

Thus, by combining (5.5), (5.12), and the equality

$$\|\hat{\mathbf{y}}\|_q = 1, \tag{5.13}$$

we conclude that $\hat{\mathbf{y}}$ solves (5.3). □

The usefulness of this approach is demonstrated when solving the regularized l_∞ problem (1.8). In this example the analog of (5.1) takes the form

$$\text{maximize} \quad \mathbf{b}^T\mathbf{y} - (\varepsilon/2)\|A^T\mathbf{y}/\varepsilon\|_2^2 - \tfrac{1}{2}(\max\{0, \|\mathbf{y}\|_1 - 1\})^2/\lambda, \tag{5.14}$$

and if $\tilde{\mathbf{y}}$ solves this problem then $\hat{\mathbf{x}} = A^T\tilde{\mathbf{y}}/\varepsilon$ solves the problem

$$\text{minimize} \quad (\nu/2)\|\mathbf{x}\|_2^2 + \|A\mathbf{x}-\mathbf{b}\|_\infty, \tag{5.15}$$

where

$$\nu = \varepsilon/(1+\lambda\|A\hat{\mathbf{x}}-\mathbf{b}\|_\infty). \tag{5.16}$$

Now the structure of (5.14) enables us to solve this problem by applying a row relaxation method (see Dax (1992a)).

Another penalty function method is implied by the following observation.

THEOREM 14. *Let* $\tilde{\mathbf{y}}$ *solve the problem*

$$\textit{maximize} \quad \mathbf{b}^T\mathbf{y} - (\varepsilon/t)\|A^T\mathbf{y}/\varepsilon\|_t^t - (\mu/q)\|\mathbf{y}/\mu\|_q^q, \tag{5.17}$$

where μ *is a given positive constant, and define*

$$\nu = \varepsilon/\|\tilde{\mathbf{y}}\|_q. \tag{5.18}$$

Then

$$\hat{\mathbf{y}} = \tilde{\mathbf{y}}/\|\tilde{\mathbf{y}}\|_q \tag{5.19}$$

solves the problem

$$(5.20)\qquad \begin{aligned} &\textit{maximize} \quad \mathbf{b}^T\mathbf{y}-(\nu/t)\|A^T\mathbf{y}/\nu\|_t^t \\ &\textit{subject to} \quad \|\mathbf{y}\|_q \leqq 1, \end{aligned}$$

while

$$(5.21)\qquad \hat{\mathbf{x}}=\mathbf{z}(\hat{\mathbf{y}})/\nu^{t-1}=\mathbf{z}(\tilde{\mathbf{y}})/\varepsilon^{t-1}$$

solves the problem

$$(5.22)\qquad \textit{minimize} \quad (\nu/s)\|\mathbf{x}\|_s^s+\|A\mathbf{x}-\mathbf{b}\|_p.$$

Proof. Note that $\tilde{\mathbf{y}}=\mathbf{0}$ if and only if $\mathbf{b}=\mathbf{0}$. Hence there is no loss of generality in excluding this possibility. The optimality conditions of (5.17) imply that

$$(5.23)\qquad \mathbf{b}-A\mathbf{z}(\tilde{\mathbf{y}})/\varepsilon^{t-1}=\mathbf{w}(\tilde{\mathbf{y}})\|\tilde{\mathbf{y}}/\mu\|_q^{q-1},$$

while (3.16) gives

$$(5.24)\qquad \|\tilde{\mathbf{y}}/\mu\|_q^{q-1}=\|A\hat{\mathbf{x}}-\mathbf{b}\|_p.$$

Hence the relations $\|\hat{\mathbf{y}}\|_q=1$, $\hat{\mathbf{x}}=\mathbf{z}(\hat{\mathbf{y}})/\nu^{t-1}$, and

$$(5.25)\qquad \mathbf{w}(\hat{\mathbf{y}})=\mathbf{w}(\tilde{\mathbf{y}})=-(A\hat{\mathbf{x}}-b)/\|A\hat{\mathbf{x}}-\mathbf{b}\|_p$$

indicate that $\hat{\mathbf{y}}$ solves (5.20). □

Observe that here

$$(5.26)\qquad \nu=\varepsilon/(\mu\|A\hat{\mathbf{x}}-\mathbf{b}\|_p^{p-1}),$$

which means that ν can be greater than ε. The appeal of (5.17) lies in the case where both s and p lie in the interval $(1,2]$. In this case the dual penalty function is twice continuously differentiable, which enables us to apply Newton's method. In the special case where $s=p=2$, the problem (5.17) takes the form

$$(5.27)\qquad \text{minimize} \quad \tfrac{1}{2}\|A^T\mathbf{y}\|_2^2/\varepsilon+\tfrac{1}{2}\|\mathbf{y}\|_2^2/\mu-\mathbf{b}^T\mathbf{y},$$

and $\tilde{\mathbf{y}}$ can be calculated via the SVD of A. Moreover, using the SVD of A we can verify that if $\varepsilon=1$ and $\tilde{\mathbf{y}}$ solves (5.27) then the primal solution $\hat{\mathbf{x}}=A^T\tilde{\mathbf{y}}$ solves the problem

$$(5.28)\qquad \text{minimize} \quad \tfrac{1}{2}\mu^{-1}\|\mathbf{x}\|_2^2+\tfrac{1}{2}\|A\mathbf{x}-\mathbf{b}\|_2^2,$$

which brings us back to "Tikhonov's regularization." It is also worthwhile to note that both (5.1) and (5.17) have the property that the vectors $\tilde{\mathbf{y}}$ and $\mathbf{b}-A\hat{\mathbf{x}}$ are "aligned." That is,

$$(5.29)\qquad \tilde{\mathbf{y}}^T(\mathbf{b}-A\hat{\mathbf{x}})=\|\tilde{\mathbf{y}}\|_q\|\mathbf{b}-A\hat{\mathbf{x}}\|_p.$$

The validity of this assertion is verified by multiplying (5.12) and (5.23) by $\tilde{\mathbf{y}}^T$.

6. Concluding remarks. The analysis of (1.1) reveals several interesting properties of this problem. We have seen that as ε moves from ∞ to 0, the solution point $\mathbf{x}_\varepsilon$ changes continuously from $\mathbf{0}$ to $\mathbf{x}^*$, where $\mathbf{x}^*$ denotes the minimum norm solution of the unregularized least norm problem. Moreover, if the system $A\mathbf{x}=\mathbf{b}$ is solvable then there exists a positive constant δ, such that $\mathbf{x}_\varepsilon=\mathbf{x}^*$ whenever $\varepsilon<\delta$. Other features that characterize (1.1) are the existence of a dual problem and the fact that a primal solution is easily retrieved from a dual solution, and vice versa.

The interest that we have in (1.1) lies in the fact that this problem can be viewed as an extension of (1.6), (1.8), and regularized linear programming problems. However, as was noted in the introduction, the "natural" way to extend Tikhonov's regularization is (1.5). Other forms of regularization that deserve our attention are

$$\text{minimize} \quad \varepsilon\|\mathbf{x}\|_s + \|A\mathbf{x}-\mathbf{b}\|_p^p/p \tag{6.1}$$

and

$$\text{minimize} \quad \varepsilon\|\mathbf{x}\|_s + \|A\mathbf{x}-\mathbf{b}\|_p. \tag{6.2}$$

The results of the current research suggest that the other forms share similar properties and provide tools for investigating this conjecture. Of special interest is the question of what structure the duals of (1.5), (6.1), and (6.2) have. The answer is displayed in Table 1, but proofs and relations between primal and dual solutions are given elsewhere (see Dax (1992b)).

TABLE 1

Duality in regularized least norm problems.

Primal problem	Dual problem
minimize $\varepsilon\|\mathbf{x}\|_s^s/s + \|A\mathbf{x}-\mathbf{b}\|_p^p/p$	maximize $\mathbf{b}^T\mathbf{y} - \varepsilon\|A^T\mathbf{y}/\varepsilon\|_t^t/t - \|\mathbf{y}\|_q^q/q$
minimize $\varepsilon\|\mathbf{x}\|_s^s/s + \|A\mathbf{x}-\mathbf{b}\|_p$	maximize $\mathbf{b}^T\mathbf{y} - \varepsilon\|A^T\mathbf{y}/\varepsilon\|_t^t/t$ subject to $\|\mathbf{y}\|_q \leqq 1$
minimize $\varepsilon\|\mathbf{x}\|_s + \|A\mathbf{x}-\mathbf{b}\|_p^p/p$	maximize $\mathbf{b}^T\mathbf{y} - \|\mathbf{y}\|_q^q/q$ subject to $\|A^T\mathbf{y}/\varepsilon\|_t \leqq 1$
minimize $\varepsilon\|\mathbf{x}\|_s + \|A\mathbf{x}-\mathbf{b}\|_p$	maximize $\mathbf{b}^T\mathbf{y}$ subject to $\|A^T\mathbf{y}/\varepsilon\|_t \leqq 1$ and $\|\mathbf{y}\|_q \leqq 1$

Note that the dual of (1.5) is a special case of (5.17) in which $\mu = 1$. Solving this problem is advantageous when $m \ll n$, and when both s and p lie in the interval $(1, 2]$. Another advantage of the dual approach is that linear constraints are transformed into simple bounds.

A further consequence of the current research is related to minimum norm problems of the form

$$\begin{aligned} &\text{minimize} && \|\mathbf{x}\|_s \\ &\text{subject to} && A\mathbf{x} = \mathbf{b}. \end{aligned} \tag{6.3}$$

The need for solving such problems arises in control theory applications as well as in other areas (e.g., Cadzow (1973)). Luenberger (1969) shows that the dual of (6.3) has the form

$$\begin{aligned} &\text{maximize} && \mathbf{b}^T\mathbf{y} \\ &\text{subject to} && \|A^T\mathbf{y}\|_t \leqq 1, \end{aligned} \tag{6.4}$$

but does not supply an explicit rule for retrieving a primal solution from a dual solution. Yet the results of Theorems 6 and 10 suggest that such a rule exists! Moreover, the difficulty in handling the inequality $\|A^T\mathbf{y}\|_t \leqq 1$ can be resolved by following the approach proposed in this paper. In particular, we can show that the dual of the problem

$$\begin{aligned} &\text{minimize} && \|\mathbf{x}\|_s^s/s \\ &\text{subject to} && A\mathbf{x} = \mathbf{b}, \end{aligned} \tag{6.5}$$

is simply

$$\text{(6.6)} \qquad \text{maximize} \quad \mathbf{b}^T\mathbf{y} - \|A^T\mathbf{y}\|_t^t / t.$$

For a detailed discussion of these ideas, see Dax (1993).

REFERENCES

J. A. CADZOW, *A finite algorithm for the minimum l_∞ solution to a system of consistent linear equations*, SIAM J. Numer. Anal., 10 (1973), pp. 607-617.

T. F. CHAN AND P. C. HANSEN, *Computing truncated singular value decomposition least squares solutions by rank revealing* QR-*factorizations*, SIAM J. Sci. Statist. Comput., 11 (1990), pp. 519-530.

A. DAX, *A new theorem of the alternative*, Math. Programming, 47 (1990), pp. 297-299.

———, *A row relaxation method for large l_1 problems*, Linear Algebra Appl., 156 (1991), pp. 793-818.

———, *A Row Relaxation Method for Large Minimax Problems*, Tech. Rep., Hydrological Service of Israel, Jerusalem, 1992a.

———, *Duality in Regularized Least Norm Problems*, Tech. Rep., Hydrological Service of Israel, Jerusalem, 1992b.

———, *A note on minimum norm solutions*, J. Optim. Theory Appl., 76 (1993), to appear.

L. ELDEN, *Algorithms for the regularization of ill-conditioned least squares problems*, BIT, 17 (1977), pp. 134-145.

P. C. HANSEN, *The truncated* SVD *as a method for regularization*, BIT, 27 (1987), pp. 534-553.

D. G. LUENBERGER, *Optimization by Vector Space Methods*, John Wiley, New York, 1969.

O. L. MANGASARIAN, *Nonlinear Programming*, McGraw-Hill, New York, 1969.

———, *Iterative solution of linear programs*, SIAM J. Numer. Anal., 18 (1981), pp. 606-614.

O. L. MANGASARIAN AND R. R. MEYER, *Nonlinear perturbation of linear programs*, SIAM J. Control Optim., 17 (1979), pp. 745-752.

G. P. MCCORMICK, *Nonlinear Programming*, John Wiley, New York, 1983.

A. N. TIKHONOV, *The stability of algorithms for the solution of degenerate systems of linear algebraic equations*, J. Comput. Math. Phys., 5 (1965), pp. 181-188.

A. N. TIKHONOV AND V. Y. ARSENIN, *Solutions of ill-posed problems*, John Wiley, New York, 1977.

J. M. VARAH, *A practical examination of some numerical methods for linear discrete ill-posed problems*, SIAM Rev., 21 (1979), pp. 100-111.

SIAM J. OPTIMIZATION
Vol. 2, No. 4, pp. 619-634, November 1992

005

ON THE CONTINUITY OF THE SOLUTION MAP IN LINEAR COMPLEMENTARITY PROBLEMS*

M. SEETHARAMA GOWDA†

Abstract. The continuity properties of the solution map $\mathcal{S}:(M, q)\mapsto \mathcal{S}(M, q)$ are investigated, where $\mathcal{S}(M, q)$ denotes the solution set corresponding to the linear complementarity problem LCP (M, q). A Robinson-type upper semicontinuity result is established for $\mathcal{S}$, and a generalization of the Mangasarian-Shiau result concerning the Lipschitzian property of $\mathcal{S}$ in the q-variable is proved. It is also shown that when the matrix is positive semidefinite (or more generally a **G**-matrix), the solution map is Lipschitz continuous with respect to the q-vector if and only if the matrix is a **P**-matrix.

Key words. complementarity problem, copositive matrix, upper and lower semicontinuity, Lipschitz continuity

AMS(MOS) subject classification. 90C33

1. Introduction.

Given a matrix $M\in R^{n\times n}$ and a vector $q\in R^n$, the Linear Complementarity Problem LCP (M, q) is to find a vector x in R^n such that

$$x\geqq 0,\quad Mx+q\geqq 0,\quad \text{and}\quad x^T(Mx+q)=0. \tag{1}$$

The advantage of studying such a problem is well documented in the literature. See, e.g., [20] and [4].

In this article, we study the behavior of the solution set as the data (M, q) changes. This amounts to the study of the continuity properties of the multivalued mapping $\mathcal{S}: R^{n\times n}\times R^n\rightarrow R^n$ defined by

$$\mathcal{S}(M, q)=\text{SOL}\,(M, q),$$

where SOL (M, q) denotes the set of all solutions of LCP (M, q), i.e., the set of all x's satisfying (1). Since $\mathcal{S}$ is a multivalued mapping, its continuity can be studied in any number of ways. In this article, we deal with the upper, lower, and Lipschitz (semi) continuity properties of $\mathcal{S}$.

While there is a large body of literature concerning the stability and continuity of solution in linear and quadratic programming problems, very few articles have been written on the continuity of the solution map in LCPs. Motivated by linear and quadratic programming problems, Robinson [26] studied the stability and continuity properties of solution maps of generalized equations. In that paper Robinson proves the upper semicontinuity property of $\mathcal{S}$ at a pair (M, q) where M is monotone and $\mathcal{S}$ is restricted to monotone (i.e., positive semidefinite) matrices. In § 2 we will extend this result to copositive matrices.

In § 3 we study the lower semicontinuity property. This concept is related to the concepts of robustness [13] and stability [11], [8]. We show in this section that in some important cases, lower semicontinuity implies uniqueness of solution. We observe that for a fully semimonotone matrix M, the problem LCP (M, q) can have at most one robust solution.

Section 4 deals with the (locally) upper Lipschitzian property of $\mathcal{S}$. This is intimately related to a result of Robinson [27], which states that for any matrix M,

* Received by the editors February 11, 1991; accepted for publication (in revised form) October 8, 1991.

† Department of Mathematics and Statistics, University of Maryland Baltimore County, Baltimore, Maryland 21228 (gowda@umbc.bitnet).

the mapping $\Phi(q)=\mathscr{S}(M,q)$ is locally upper Lipschitzian. We give an alternate proof and at the same time extend a result of Mangasarian and Shiau [18], that for a **P**-matrix M, the mapping $\Phi(q)$ is Lipschitzian.

Our final section deals with the Lipschitzian characterization of **P**-matrices. In [18], Mangasarian and Shiau give an example of a positive semidefinite matrix for which the mapping Φ is not Lipschitzian. We show in this article that the mapping Φ corresponding to a positive semidefinite matrix cannot be Lipschitzian unless the matrix is a **P**-matrix. In fact, in Theorem 13 we prove a converse of the Mangasarian-Shiau result: If M is a **G**-matrix for which Φ is Lipschitzian, then M is a **P**-matrix. The set $\mathscr{G}$ of all **G**-matrices includes semimonotone matrices. In particular, the set $\mathscr{G}$ contains copositive matrices, $\mathbf{P}_0$-matrices, and positive semidefinite matrices. The results proved in § 5 partially answer a question posed to the author by Jong-Shi Pang: If the mapping Φ corresponding to a **Q**-matrix is Lipschitzian, should the matrix be a **P**-matrix? We conclude this paper by proving the Lipschitzian property for negative **N**-matrices.

Let us say a few words about the notation. Throughout this paper, $\|M\|$ and $\|q\|$ denote (arbitrary but fixed) norms of the matrix M and the vector q, respectively. B denotes the open unit ball in $(R^n, \|\cdot\|)$. Corresponding to any set E we write $|E|$ for the number of elements in that set. The closure of a set $\mathscr{C}$ in $R^{n\times n}$ is denoted by $\bar{\mathscr{C}}$. For any set $E\subseteq R^n$, E^* denotes the dual of E defined by

$$E^*=\{y\in R^n\colon y^Tx\geqq 0 \quad \forall x\in E\}.$$

The domain of the mapping $\mathscr{S}$ is defined by

$$\operatorname{dom}\mathscr{S}=\{(M,q)\in R^{n\times n}\times R^n\colon \mathscr{S}(M,q)\neq\emptyset\}.$$

Corresponding to any nonzero solution x of LCP (M,q), the set $\alpha=\{i: x_i>0\}$ will be called the support set of x and the submatrix $M_{\alpha\alpha}$ of M will be called the *supporting submatrix* corresponding to x.

2. Upper semicontinuity.

Definition 1. Let $\mathscr{C}$ be a nonempty subset of $R^{n\times n}$, $(M,q)\in\bar{\mathscr{C}}\times R^n$. The mapping $\mathscr{S}$ is said to be $\mathscr{C}$-upper semicontinuous ($\mathscr{C}$-usc for short) at (M,q) if for every $\varepsilon>0$ there exists a $\delta>0$ such that

$$\mathscr{S}(M',q')\subseteq\mathscr{S}(M,q)+\varepsilon B \tag{2}$$

for all $(M',q')\in\mathscr{C}\times R^n$ satisfying $\|M'-M\|+\|q'-q\|<\delta$. If $\mathscr{C}$ is $R^{n\times n}$, we omit the prefix $\mathscr{C}$.

Remarks. Likely candidates for the set $\mathscr{C}$ are singleton sets, the set of all skew-symmetric matrices, the set of all positive semidefinite matrices, the set of all copositive matrices, the set of all semimonotone matrices, $R^{n\times n}$, etc. If we replace (in the above definition) the open set $\mathscr{S}(M,q)+\varepsilon B$ (which contains $\mathscr{S}(M,q)$) by an arbitrary open set U containing $\mathscr{S}(M,q)$, we obtain a stronger definition. That this definition is indeed stronger is illustrated in the example below. We note, however, that the two definitions coincide when $\mathscr{S}(M,q)$ is compact.

Example 1. Let

$$M=\begin{bmatrix}0 & 0\\ 0 & 1\end{bmatrix},\quad q=\begin{bmatrix}0\\ 0\end{bmatrix},\quad q'=\begin{bmatrix}0\\ \varepsilon\ln\varepsilon\end{bmatrix},$$

where $0<\varepsilon<1$. Put $\mathscr{C}=\{M\}$. It can be easily verified that $\mathscr{S}(M,q)$ and $\mathscr{S}(M,q')$ are parallel (half-) lines in R^2 and so $\mathscr{S}(M,q')$ cannot be contained in arbitrary open sets (which contain $\mathcal{S}(M, q)$) even if ε is small. Since $\mathcal{S}$ is easily seen to be $\mathscr{C}$-usc at (M, q), the statement made in the above remark is justified.

Our aim in this section is to describe certain triples $(\mathscr{C}, M, q)$ for which the mapping $\mathcal{S}$ is $\mathscr{C}$-usc at (M, q). We start with a fundamental result due to Robinson [27], which shows that for any matrix M, $\mathcal{S}$ is $\{M\}$-usc at (M, q) for every q.

THEOREM 1. *Let M be any $n \times n$ matrix. Then there is a positive number λ with the following property: Given any $q \in R^n$, there exists an $\delta > 0$ such that*

$$(3) \qquad \mathcal{S}(M, q') \subseteq \mathcal{S}(M, q) + \lambda \|q' - q\| B$$

for all q' with $\|q' - q\| < \delta$.

The next result can be found in [13, Thm. 5.6]. (Actually, in [13] the result is proved for (M, q) with $\mathcal{S}(M, q) \neq \emptyset$. It is easy to show, using normalized vectors, that the result is true even when $\mathcal{S}(M, q)$ is empty.) Before stating the result, we recall that a matrix M is an $\mathbf{R}_0$-matrix if $\mathcal{S}(M, 0) = \{0\}$. It is well known that M is an $\mathbf{R}_0$-matrix if and only if $\mathcal{S}(M, q)$ is bounded for every q.

THEOREM 2. *Let M be an $\mathbf{R}_0$-matrix. Then $\mathcal{S}$ is usc at (M, q) for every q.*

We now concentrate on positive semidefinite matrices and copositive matrices. The following simple example due to Robinson [26] shows that even when M is positive semidefinite, the mapping $\mathcal{S}$ need not be upper semicontinuous.

Example 2. Let $M = [0]$, $q = [1]$. Then $\mathcal{S}(M, q) = \{0\}$ while for any $M' = [-\varepsilon]$ with $\varepsilon > 0$, we have $\mathcal{S}(M', q) = \{0, 1/\varepsilon\}$. Clearly, $\mathcal{S}$ is not usc at (M, q).

In [26], Robinson shows that when M is positive semidefinite and $\mathcal{S}(M, q)$ is nonempty and compact, $\mathcal{S}$ is $\mathscr{C}$-usc at (M, q) where $\mathscr{C}$ is the set of all positive semidefinite matrices. Robinson proves this as a by-product of his theory of generalized equations. Here, we extend this result to copositive matrices. We recall that an $n \times n$ matrix M is said to be *copositive* if $x^T M x \geqq 0$ for all $x \geqq 0$ in R^n. In the following result, "int" refers to the interior.

THEOREM 3. *Let M be an $n \times n$ copositive matrix and let $q \in \operatorname{int} \mathcal{K}^*$ where $\mathcal{K} = \mathcal{S}(M, 0)$. Then $\mathcal{S}$ is $\mathscr{C}$-usc at (M, q) where $\mathscr{C}$ denotes the set of all $n \times n$ copositive matrices.*

Proof. Assume the contrary. Then there is an open set U containing $\mathcal{S}(M, q)$, a sequence $\{(M^k, q^k)\}$ converging to (M, q) in $R^{n \times n} \times R^n$ with each M^k belonging to $\mathscr{C}$, and a sequence $\{x^k\}$ such that for all k, $x^k \in \mathcal{S}(M^k, q^k)$ and $x^k \notin U$. The sequence $\{\|x^k\|\}$ diverges to ∞; otherwise, a subsequential limit of $\{x^k\}$ belongs to $\mathcal{S}(M, q)$ and to the complement of U, a contradiction. Let s be any subsequential limit of $\{x^k / \|x^k\|\}$. Then it is easily seen that $0 \neq s \in \mathcal{S}(M, 0)$. From the copositivity of M^k and the equation $(x^k)(M^k x^k + q^k) = 0$, we get $(x^k)^T q^k \leqq 0$. This leads to the inequality $s^T q \leqq 0$. Since q belongs to the interior of $\mathcal{K}^*$, we must have $s^T q > 0$ and thus we reach a contradiction. □

We make several remarks regarding Theorem 3. First, Theorem 3 does not assert the nonemptiness of $\mathcal{S}(M', q')$ when (M', q') is near (M, q). The nonemptiness of $\mathcal{S}(M, q)$ is (essentially) guaranteed by Lemke's theorem [15]. (See [9] for an explicit reference.) The following result, first observed by Stone [4], answers the nonemptiness question. Stone's proof is based on the degree theory. For a proof based on the basic theorem of complementarity of Eaves, see [10].

THEOREM 4. *Let M be a copositive matrix and $q \in \operatorname{int} \mathcal{K}^*$ where $\mathcal{K} = \mathcal{S}(M, 0)$. Then for all copositive matrices M' and all vectors q' with (M', q') near (M, q), $\mathcal{S}(M', q') \neq \emptyset$.*

We give two illustrations of Theorem 3 by specializing M. First let M be a copositive-star matrix, i.e., let M be copositive and satisfy the condition

$$x \in \mathcal{S}(M, 0) \Rightarrow M^T x \leqq 0.$$

For such a matrix, $\mathcal{K}^* = \{x\colon x \geqq 0, M^Tx \leqq 0\}^* = R^n_+ - M(R^n_+)$. Also, $\operatorname{int}(R^n_+ - M(R^n_+)) = \operatorname{int} R^n_+ - M(R^n_+)$. We have the following corollary.

COROLLARY 1. *Let M be a copositive-star matrix and suppose that* LCP (M, q) *has a strictly feasible solution, i.e., there exists a nonnegative vector* u^* *such that* $Mu^* + q > 0$. *Then the conclusion of Theorem* 3 *holds.*

For the second illustration, let M be copositive-plus, i.e., let M be copositive and satisfy the condition

$$x \geqq 0, \qquad x^TMx = 0 \Rightarrow (M + M^T)x = 0.$$

Then M is copositive-star and (see [16]) $q \in \operatorname{int} \mathcal{K}^*$ if and only if $\mathcal{S}(M, q)$ is nonempty and compact. We thus obtain the conclusion of Theorem 3 when M is copositive-plus and the solution set of LCP (M, q) is nonempty and compact. This of course yields the result of Robinson mentioned before.

The following examples show that the conclusion of Theorem 3 may nor may not hold when $q \notin \operatorname{int} \mathcal{K}^*$.

Example 3. Let $M = [0]$, $q = [0]$. Then $\mathcal{S}(M, q) = R^1_+$. Note that the interiority assumption in Theorem 3 does not hold, but still $\mathcal{S}$ is upper semicontinuous at (M, q).

Example 4. Let

$$M = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}, \quad M' = \begin{bmatrix} 0 & \varepsilon \\ -\varepsilon & 1 \end{bmatrix}, \quad q = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \quad q' - \begin{bmatrix} \varepsilon \ln \varepsilon \\ \varepsilon \ln \varepsilon \end{bmatrix},$$

where $0 < \varepsilon < 1$. It can be easily verified that M and M' are positive semidefinite, and that (M', q') is near (M, q) when ε is small. Furthermore, $\mathcal{K} = \mathcal{S}(M, 0)$ is the nonnegative x-axis in R^2 and so q lies in the boundary of $\mathcal{K}^*$. The vector with components $x_\varepsilon = (1/\varepsilon)\{\varepsilon \ln \varepsilon - \ln \varepsilon\}$ and $y_\varepsilon = -\ln \varepsilon$ is an element of $\mathcal{S}(M', q')$ but is far away from $\mathcal{S}(M, q)$ when ε is small. Thus $\mathcal{S}$ is not upper semicontinuous even when restricted to positive semidefinite matrices.

3. Lower semicontinuity.

DEFINITION 2. Let $\mathcal{C}$ be a nonempty subset of $R^{n \times n}$, $(M, q) \in \bar{\mathcal{C}} \times R^n$. The mapping $\mathcal{S}$ is said to be $\mathcal{C}$-lower semicontinuous ($\mathcal{C}$-lsc for short) at (M, q) if for every $\varepsilon > 0$ there exists a $\delta > 0$ such that

$$(4) \qquad \mathcal{S}(M, q) \subseteq \mathcal{S}(M', q') + \varepsilon B$$

for all $(M', q') \in (\mathcal{C} \times R^n) \cap \operatorname{dom} \mathcal{S}$ satisfying $\|M' - M\| + \|q' - q\| < \delta$. If $\mathcal{C}$ is $R^{n \times n}$, we omit the prefix $\mathcal{C}$.

It is clear that the above definition implies that $\mathcal{S}$ is $\mathcal{C}$-lsc at every x^* in $\mathcal{S}(M, q)$ where $\mathcal{C}$-lsc at x^* means: For every $\varepsilon > 0$ there exists a $\delta > 0$ such that

$$(5) \qquad \mathcal{S}(M', q') \cap (x^* + \varepsilon B) \neq \emptyset$$

for all $(M', q') \in (\mathcal{C} \times R^n) \cap \operatorname{dom} \mathcal{S}$ satisfying $\|M' - M\| + \|q' - q\| < \delta$. (It is not difficult to show that the two formulations are equivalent if $\mathcal{S}(M, q)$ is compact.) In the above two formulations, we demand that (M', q') be in $\operatorname{dom} \mathcal{S}$. This is done in order to accommodate points on the boundary of $\operatorname{dom} \mathcal{S}$. Suppose we remove the restriction that (M', q') be in $(\mathcal{C} \times R^n) \cap \operatorname{dom} \mathcal{S}$ from the definition of $\mathcal{C}$-lower semicontinuity at x^*. The resulting definition is precisely that of *robustness* [13] of x^* for LCP (M, q). Thus we conclude that if a solution x^* is robust for LCP (M, q), then $\mathcal{S}$ is $\mathcal{C}$-lsc at x^* for every nonempty set $\mathcal{C} \subseteq R^{n \times n}$. We remark that the notion of robustness is weaker than the *stability* issue treated in [11] and [8]. (In fact, a solution x^* is stable for LCP (M, q) if and only if it is isolated and robust [13].)

To illustrate the above definition, consider a copositive-star matrix M and let $\mathscr{C}$ denote the set of all copositive matrices. Let $q \in R^n$ be such that $|\mathscr{S}(M, q)| = 1$ and $q \in \operatorname{int} \mathscr{K}^*$. (The second condition is superfluous if M is copositive-plus.) By combining Theorems 3 and 4 we see that $\mathscr{S}$ is $\mathscr{C}$-lsc at (M, q).

We show in this section that in some important cases lower semicontinuity implies uniqueness of solution. The underlying idea in the proofs of all the results in this section is the following. If $\mathscr{S}$ is $\mathscr{C}$-lsc at (M, q), then (4) holds for every ε. By choosing (M', q') so that $|\mathscr{S}(M', q')| = 1$, we can cause $\mathscr{S}(M, q)$ to lie in a ball of arbitrarily small radius. This implies that $\mathscr{S}(M, q)$ is a singleton set. The conditions we need in order to apply the above idea depend on whether q is zero or nonnegative or not nonnegative.

Before stating our first lower semicontinuity result, we recall that a solution x^* of LCP(M, q) is *nondegenerate* with respect to M if $x^* + Mx^* + q > 0$ and that LCP(M, q) is nondegenerate (or q is nondegenerate with respect to M) if every solution of LCP(M, q) is nondegenerate. For a given $n \times n$ matrix M, the set of all degenerate q's is contained in a finite union of proper subspaces of R^n (see [6, Lemma 4]). It follows that every open set, consisting only of q's for which $\mathscr{S}(M, q)$ is nonempty, must contain a nondegenerate q. In particular, the set of all nondegenerate positive vectors is dense in R^n_+, and when the matrix M is a **Q**-matrix (i.e., when $\mathscr{S}(M, q) \neq \emptyset$ for all $q \in R^n$) the set of all nondegenerate vectors is dense in R^n.

THEOREM 5. *Let $M \in \bar{\mathscr{C}} \subset R^{n \times n}$. If $\mathscr{S}$ is $\mathscr{C}$-lsc at $(M, 0)$, then M is an $\mathbf{R}_0$-matrix.*

Proof. Take any $\varepsilon > 0$. Then

$$\mathscr{S}(M, 0) \subseteq \mathscr{S}(M', q') + \varepsilon B$$

for all $(M', q') \in (\mathscr{C} \times R^n) \cap \operatorname{dom} \mathscr{S}$ sufficiently close to $(M, 0)$. In particular, this inclusion holds for a pair (M', q') where q' is positive and nondegenerate with respect to M'. But for such an (M', q'), the set $\mathscr{S}(M', q')$ is finite [19]. The above inclusion shows that $\mathscr{S}(M, 0)$ is bounded. Since $\mathscr{S}(M, 0)$ is a cone, we must have $\mathscr{S}(M, 0) = \{0\}$. This completes the proof. □

We recall some more definitions. We say that a matrix M is *semimonotone* if for every nonzero nonnegative x there exists an index i such that $x_i > 0$ and $x_i(Mx)_i \geqq 0$. An equivalent definition [6] is: For all $d > 0$, $\mathscr{S}(M, d) = \{0\}$. We note that copositive matrices are semimonotone. A matrix M is said to be *strictly semimonotone* if for every nonzero nonnegative x there exists an index i such that $x_i > 0$ and $x_i(Mx)_i > 0$; equivalently [6] if for all $d \geqq 0$, $\mathscr{S}(M, d) = \{0\}$. Cottle [2] characterizes a strictly semimonotone matrix as a **Q**-matrix for which every principal submatrix is also a **Q**-matrix. Our next definition involves a *fully semimonotone* matrix. To decribe this, we consider a matrix M and an index set $\alpha \subseteq \{1, 2, \ldots, n\}$. Let β denote the complement of α in $\{1, 2, \ldots, n\}$. If the principal submatrix $M_{\alpha\alpha}$ of M is nonsingular, we call the matrix

$$\hat{M} = \begin{bmatrix} M_{\alpha\alpha}^{-1} & -M_{\alpha\alpha}^{-1} M_{\alpha\beta} \\ M_{\beta\alpha} M_{\alpha\alpha}^{-1} & M_{\beta\beta} - M_{\beta\alpha} M_{\alpha\alpha}^{-1} M_{\alpha\beta} \end{bmatrix} \tag{6}$$

a *principal pivot transform* of M obtained by pivoting on the submatrix $M_{\alpha\alpha}$. (The transform corresponding to $\alpha = \emptyset$ is M itself.) We shall say that a matrix M is *fully semimonotone* if every principal pivot transform is semimonotone. We note that positive semidefinite matrices (more generally, $\mathbf{P}_0$-matrices, i.e., matrices with nonnegative principal minors) are fully semimonotone.

The following observations regarding a matrix M and its principal pivot transforms are useful. For a matrix M, suppose $M_{\alpha\alpha}$ is nonsingular and let z and w be two vectors satisfying $w=Mz+q$. The triple $(\hat{M}, \hat{q}, \hat{z})$ is called the principal pivot transform of (M, q, z) (obtained by pivoting on $M_{\alpha\alpha}$) where $\hat{M}$ is defined above and

$$z=\begin{bmatrix} z_\alpha \\ z_\beta \end{bmatrix}, \quad q=\begin{bmatrix} q_\alpha \\ q_\beta \end{bmatrix}, \quad w=\begin{bmatrix} w_\alpha \\ w_\beta \end{bmatrix},$$

$$\hat{q}=\begin{bmatrix} -M_{\alpha\alpha}^{-1}q_\alpha \\ q_\beta - M_{\beta\alpha}M_{\alpha\alpha}^{-1}q_\alpha \end{bmatrix}, \quad \hat{z}=\begin{bmatrix} w_\alpha \\ z_\beta \end{bmatrix}. \tag{7}$$

It is easily seen that $\hat{z}$ is a solution of LCP $(\hat{M}, \hat{q})$ if and only if z is a solution of LCP (M, q) and $|\mathscr{S}(\hat{M}, \hat{q})| = |\mathscr{S}(M, q)|$. Furthermore, there exist constants C_1 and C_2 depending only on the matrix M such that for $w=Mz+q$ and $y=Mx+p$,

$$\|\hat{z}-\hat{x}\| \leqq C_1\{\|z-x\|+\|q-p\|\} \tag{8}$$

and

$$\|q-p\| \leqq C_2\|\hat{q}-\hat{p}\|. \tag{9}$$

The transformation $(M, q, z) \to (\hat{M}, \hat{q}, \hat{z})$ has the following *local homeomorphism property.* When $M_{\alpha\alpha}$ is nonsingular, a triple $(N, r, u) \in R^{n\times n}\times R^n \times R^n$ is close to (M, q, z) if and only if $(\hat{N}, \hat{r}, \hat{u})$ is close to $(\hat{M}, \hat{q}, \hat{z})$ where $(\hat{N}, \hat{r}, \hat{u})$ is obtained by pivoting on the submatrix $N_{\alpha\alpha}$. For more properties of the principal pivot transforms, see [3], [24], and [30].

THEOREM 6. *Let M be a semimonotone matrix and $q \geqq 0$. Consider the following statements*:

(a) *$\mathscr{S}$ is lower semicontinuous at (M, q).*

(b) *$\mathscr{S}$ is $\{M\}$-lower semicontinuous at (M, q).*

(c) LCP (M, q) *has a unique solution.*

Then (a)$\Rightarrow$(b)$\Rightarrow$(c). *The implication* (c)$\Rightarrow$(a) *holds if M is an $\mathbf{R}_0$-matrix.*

Proof. The implication (a)$\Rightarrow$(b) is obvious. Suppose (b) holds. Then for any $\varepsilon>0$, the inclusion (4) holds for all (M', q') close to (M, q) where $M'=M$ and $(M', q') \in$ dom $\mathscr{S}$. Pick a positive q' that is close to q. Since $M'=M$ is semimonotone, $\mathscr{S}(M', q') = \{0\}$. Because of (4), $\mathscr{S}(M, q)$ is contained in a ball of arbitrarily small radius. Hence LCP (M, q) has a unique solution. As regards the implication (c)$\Rightarrow$(a): If M is an $\mathbf{R}_0$-matrix, we apply Theorems 2 and 3 in [8]. □

THEOREM 7. *Let M be a fully semimonotone matrix, $0 \not\leqq q \in R^n$ with $\mathscr{S}(M, q) \neq \emptyset$. Consider the following statements*:

(a) *$\mathscr{S}$ is lower semicontinuous at (M, q) and $\mathscr{S}(M, q') \neq \emptyset$ for all q' near q.*

(b) *$\mathscr{S}$ is $\{M\}$-lower semicontinuous at (M, q) and $\mathscr{S}(M, q') \neq \emptyset$ for all q' near q.*

(c) *$\mathscr{S}$ is lower semicontinuous at (M, q) and there exists a solution $x^* \in \mathscr{S}(M, q)$ for which the supporting submatrix $M_{\alpha\alpha}$ is nonsingular.*

(d) LCP (M, q) *has a unique solution.*

Then (a)$\Rightarrow$(b)$\Rightarrow$(d) *and* (c)$\Rightarrow$(d). *The implication* (d)$\Rightarrow$(a) *holds if M is an $\mathbf{R}_0$-matrix and the implication* (d)$\Rightarrow$(c) *holds if the supporting submatrix $M_{\alpha\alpha}$ of the* (*unique*) *solution x^* of* LCP (M, q) *is nonsingular.*

Proof. (a)$\Rightarrow$(b) is obvious. Suppose (b) holds. Then for any $\varepsilon>0$, the inclusion (4) holds for all (M', q') close to (M, q) where $M'=M$ and $(M', q') \in$ dom $\mathscr{S}$. Since $\mathscr{S}(M, q') \neq \emptyset$ for all q' close to q, we can pick a q' close to q such that q' is nondegenerate with respect to M. Since M is fully semimonotone, an observation due to Stone [30, p. 119] shows that $\mathscr{S}(M', q') = \mathscr{S}(M, q')$ is a singleton set. Because of (4), $\mathscr{S}(M, q)$ is contained in a ball of arbitrarily small radius. Hence LCP (M, q) has a unique solution. Thus we have (d).

Now suppose that (c) holds. Then by pivoting on $M_{\alpha\alpha}$, we can transform LCP (M, q) into an equivalent problem LCP (M_0, q_0) where $q_0 \geqq 0$. It can be easily verified (using the local homeomorphism property described earlier) that $\mathcal{S}$ is $\{M_0\}$-lower semicontinuous at (M_0, q_0). Since M is fully semimonotone, M_0 is semimonotone and so by Theorem 6, LCP (M_0, q_0) has a unique solution. This implies that LCP (M, q) has a unique solution.

If (d) holds and M is an $\mathbf{R}_0$-matrix, we apply Theorems 2 and 3 in [8] and get (a). If (d) holds and the supporting submatrix $M_{\alpha\alpha}$ of the (unique) solution x^* of LCP (M, q) is nonsingular, we apply Corollary 2 in [8] and get (c). □

Remarks. The condition that $\mathcal{S}(M, q') \neq \emptyset$ for all q' near q in parts (a) and (b) of the above theorem holds [16] when $\mathcal{S}(M, q)$ has a nondegenerate vertex solution. This condition can be slightly weakened by requiring q to lie in the closure of the interior of the set $K(M) = \{r \in R^n : \mathcal{S}(M, r) \neq \emptyset\}$. This weaker condition is satisfied, for example, when $q \in R^n_+ - M(R^n_+) = K(M)$ (i.e., when M is a $\mathbf{Q}_0$-matrix and LCP (M, q) has a feasible solution).

The proofs of Theorems 6 and 7 can be modified to get Corollary 2.

COROLLARY 2. (a) *If M is semimonotone, then for every $q \geqq 0$, the zero vector can be the only possible robust solution for* LCP (M, q).

(b) *If M is fully semimonotone, then for every q, the problem* LCP (M, q) *can have at most one robust solution.*

4. Lipschitz continuity of $\mathcal{S}(M, q)$. This section deals with the Lipschitz continuity of the mapping $\mathcal{S}$. The concepts we will define below are somewhat similar to Robinson's locally upper Lipschitzian property stated in Theorem 1. In what follows we let $\mathscr{E}$ denote a nonempty subset of R^n and let $\mathscr{C}$ denote a nonempty subset of $R^{n \times n}$. We drop the prefix $\mathscr{C}$ if $\mathscr{C}$ is $R^{n \times n}$.

DEFINITION 3. (a) We say that $\mathcal{S}$ is $\mathscr{C}$-locally upper Lipschitzian at $(M, q) \in \bar{\mathscr{C}} \times R^n$ if there is a positive number λ and a $\delta > 0$ such that

$$\mathcal{S}(M', q') \subseteq \mathcal{S}(M, q) + \lambda(\|M' - M\| + \|q' - q\|)B$$

for all $(M', q') \in \mathscr{C} \times R^n$ satisfying $\|M' - M\| + \|q' - q\| < \delta$.

(b) We say that $\mathcal{S}$ is $\mathscr{C}$-locally lower Lipschitzian at $(M, q) \in \bar{\mathscr{C}} \times R^n$ if there is a positive number λ and a $\delta > 0$ such that

$$\mathcal{S}(M, q) \subseteq \mathcal{S}(M', q') + \lambda(\|M' - M\| + \|q' - q\|)B$$

for all $(M', q') \in (\mathscr{C} \times R^n) \cap \operatorname{dom} \mathcal{S}$ satisfying $\|M' - M\| + \|q' - q\| < \delta$.

(c) $\mathcal{S}$ is said to be Lipschitzian on $\mathscr{C} \times \mathscr{E}$ if there is a positive number L such that

$$\mathcal{S}(M, q) \subseteq \mathcal{S}(N, p) + L(\|M - N\| + \|q - p\|)B$$

for all (M, q) and (N, p) in $(\mathscr{C} \times \mathscr{E}) \cap \operatorname{dom} \mathcal{S}$.

It is clear from the above definitions that if $\mathcal{S}$ is $\mathscr{C}$-locally upper Lipschitzian ($\mathscr{C}$-locally lower Lipschitzian) at (M, q), then it is $\mathscr{C}$-upper semicontinuous ($\mathscr{C}$-lower semicontinuous) at (M, q). The following result gives a converse.

THEOREM 8. *Suppose that $\mathcal{S}(M, q)$ is nonempty, compact, and $\mathcal{S}$ is $\mathscr{C}$-usc at (M, q). Then $\mathcal{S}$ is $\mathscr{C}$-locally upper Lipschitzian at (M, q).*

Proof. It follows from the hypothesis that there are constants $\eta > 0$ and $1 \leqq L < \infty$ such that

$$\|x\| \leqq L \quad \forall x \in \mathcal{S}(M', q'), \tag{10}$$

for all $(M', q') \in \mathscr{C} \times R^n$ with $\|M' - M\| + \|q' - q\| < \eta$. By Theorem 1, there are numbers λ and ε such that for all q'' with $\|q'' - q\| < \varepsilon$, we have

$$\mathscr{S}(M, q'') \subseteq \mathscr{S}(M, q) + \lambda \|q'' - q\| B. \tag{11}$$

Now consider (M', q') satisfying the conditions

$$M' \in \mathscr{C}, \qquad L\|M' - M\| + \|q' - q\| < \min\{\varepsilon, \eta\}. \tag{12}$$

Let $x \in \mathscr{S}(M', q')$ so that $x \in \mathscr{S}(M, q'')$ where $q'' = (M' - M)x + q'$. By (12) and (10),

$$\|q'' - q\| \leqq L\|M' - M\| + \|q' - q\| < \varepsilon$$

so that by (11),

$$\mathscr{S}(M, q'') \subseteq \mathscr{S}(M, q) + \lambda\{L\|M' - M\| + \|q' - 1\|\}B.$$

Therefore, $x \in \mathscr{S}(M, q) + \lambda L\{\|M' - M\| + \|q' - q\|\}B$. This gives the desired result. □

Combining Theorems 2 and 8, we get the following corollary.

COROLLARY 3. *Let M be an $\mathbf{R}_0$-matrix. Then $\mathscr{S}$ is locally upper Lipschitzian at (M, q) for every vector q with $\mathscr{S}(M, q) \neq \emptyset$.*

In Definitions 3(a) and (b), the vector q' is allowed to vary in a neighborhood of q. We show below that when M is an $\mathbf{R}_0$-matrix, this restriction can be removed. First we prove a lemma.

LEMMA 1. *Suppose that M is an $\mathbf{R}_0$-matrix. Then there is an $\varepsilon > 0$ and a positive number L such that*

$$\|x'\| \leqq L(\|q'\| + \|M' - M\|) \tag{13}$$

for all (M', q', x') with $\|M' - M\| < \varepsilon$, $q' \in R^n$ and $x' \in \mathscr{S}(M', q')$.

Proof. Assume the contrary. Then there are sequences $\{M^k\} \subset R^{n \times n}$, $\{q^k\} \subset R^n$, $\{x^k\} \subset R^n$ such that $M^k \to M$, $x^k \in \mathscr{S}(M^k, q^k)$, and

$$\|x^k\| > k(\|q^k\| + \|M^k - M\|) \tag{14}$$

for all k. If $\{x^k\}$ is unbounded, we can assume without loss of generality that $\|x^k\| \to \infty$. Then the above inequality shows that $\|q^k\| / \|x^k\| \to 0$, in which case a subsequential limit of $\{x^k / \|x^k\|\}$ belongs to $\mathscr{S}(M, 0)$. Since this subsequential limit is nonzero and M is an $\mathbf{R}_0$-matrix, we reach a contradiction, and hence $\{x^k\}$ must be bounded. But then, (14) shows that $q^k \to 0$. By Corollary 3, there is a positive λ such that for all large k,

$$\mathscr{S}(M^k, q^k) \subseteq \mathscr{S}(M, 0) + \lambda(\|q^k\| + \|M^k - M\|)B.$$

Since $\mathscr{S}(M, 0) = \{0\}$, we reach a contradiction to (14). This proves the lemma. □

THEOREM 9. *Let M be an $\mathbf{R}_0$-matrix and $q \in R^n$ such that $\mathscr{S}(M, q) \neq \emptyset$. Then:*

(a) *There exist a $K > 0$ and an $\varepsilon > 0$ such that for all $(M', q') \in \mathscr{C} \times R^n$ with $\|M' - M\| < \varepsilon$ the inclusion*

$$\mathscr{S}(M', q') \subseteq \mathscr{S}(M, q) + K(\|M' - M\| + \|q' - q\|)B$$

holds.

(b) *If $\mathscr{S}$ is $\mathscr{C}$-locally lower Lipschitzian at (M, q), then there exist a $K > 0$ and an $\varepsilon > 0$ such that for all $(M', q') \in (\mathscr{C} \times R^n) \cap \operatorname{dom} \mathscr{S}$ with $\|M' - M\| < \varepsilon$ the inclusion*

$$\mathscr{S}(M, q) \subseteq \mathscr{S}(M', q') + K(\|M' - M\| + \|q' - q\|)B$$

holds.

Proof. We prove (a), the proof of (b) being similar. Suppose that (a) is false. Then there are sequences $\{M^k\}$ in $\mathscr{C}$, $\{q^k\}, \{x^k\}, \{u^k\}$ in R^n such that $M^k \to M$, $x^k \in \mathscr{S}(M^k, q^k)$, $u^k \in \mathscr{S}(M, q)$, and

$$(15) \qquad \min_{u \in \mathscr{S}(M,q)} \|x^k - u\| = \|x^k - u^k\| > k(\|M^k - M\| + \|q^k - q\|)$$

for all k. (Note that $\mathscr{S}(M, q)$ is compact so that the above "min" is well defined.) We claim that $\{\|q^k\|\}$ is bounded. Suppose not and assume without loss of generality that $\|q^k\| \to \infty$. Lemma 1 shows that the sequence $\{x^k/\|q^k\|\}$ is bounded. By dividing the above inequality by $k\|q^k\|$ and letting k go to ∞ through a suitable subsequence, we conclude that a subsequential limit of $\{q^k/\|q^k\|\}$ is zero. This, of course, is false, and hence $\{q^k\}$ is bounded. But then, by Lemma 1, $\{x^k\}$ is bounded. The above inequality shows that $q^k \to q$. By Corollary 3, $\mathscr{S}$ is locally upper Lipschitzian at (M, q) and so there is a positive λ such that for all large k,

$$\mathscr{S}(M^k, q^k) \subseteq \mathscr{S}(M, q) + \lambda(\|M^k - M\| + \|q^k - q\|)B.$$

This contradicts (15), and hence we have proved (a). □

We conclude this section by proving a Lipschitzian property of $\mathscr{S}$. The result says that $\mathscr{S}$ is Lipschitzian in (M, q) when (M, q) varies over a compact subset of the set $\mathscr{P} \times R^n$ where $\mathscr{P}$ is the set of all **P**-matrices. This is a generalization of the Mangasarian-Shiau result [18], which says that the mapping $\Phi(q) := \mathscr{S}(M, q)$ is Lipschitzian in q over all of R^n when M is a **P**-matrix. Mangasarian and Shiau prove their result using complementary cones. In the present paper we give an alternate proof of the Mangasarian-Shiau result. The generalization and the alternate proof of the Mangasarian-Shiau result may have been known to several researchers and may perhaps be implicit in the works of [22], [23], and [28]. We recall a few things about **P**-matrices. Let M be a **P**-matrix so that by definition every principal minor of M is positive. It is well known that for any $q \in R^n$, $|\mathscr{S}(M, q)| = 1$. In fact, this is a characterization of **P**-matrices [19]. Another equivalent characterization due to Gale and Nikaido is that for every nonzero vector x,

$$\max_{1 \leqq i \leqq n} x_i(Mx)_i > 0.$$

It can be easily verified that the quantity

$$(16) \qquad \alpha(M) := \min_{\|x\|_\infty = 1} \max_{1 \leqq i \leqq n} x_i(Mx)_i$$

is well defined and positive where $\|x\|_\infty$ denotes the ∞-norm of the vector x. Furthermore, $\alpha(M)$ is continuous on $\mathscr{P}$. In what follows, we write $\|\mathscr{S}(M, q) - \mathscr{S}(M, r)\|_\infty$ for the ∞-norm of the vector $x - y$ where $\mathscr{S}(M, q) = \{x\}$ and $\mathscr{S}(M, r) = \{y\}$. The ∞-norm of a matrix M (i.e., the norm of the operator $M : (R^n, \|\cdot\|_\infty) \to (R^n, \|\cdot\|_\infty)$) is denoted by $\|M\|_\infty$.

THEOREM 10. *Let M be a* **P**-*matrix. Then*

$$(17) \qquad \|\mathscr{S}(M, q) - \mathscr{S}(M, r)\|_\infty \leqq \alpha(M)^{-1}\|q - r\|_\infty$$

for all q and r in R^n.

Proof. Let $\mathscr{S}(M, q) = \{x\}$ and $\mathscr{S}(M, r) = \{y\}$ so that $x \geqq 0$, $u = Mx + q \geqq 0$, $x^T u = 0$, $y \geqq 0$, $v = My + r \geqq 0$, and $y^T v = 0$. Then for every index i,

$$\begin{aligned}(x - y)_i\{M(x - y)\}_i &= (x - y)_i\{Mx + q - (My + r) + (r - q)\}_i \\ &= (x - y)_i(r - q)_i - x_i v_i - y_i u_i,\end{aligned}$$

where we have used the complementary conditions $x_iu_i=0$ and $y_iv_i=0$. Since x_iv_i and y_iu_i are nonnegative, we have

$$\begin{aligned}\alpha(M)\|x-y\|_\infty^2 &\leqq \max_{1\leqq i\leqq n}(x-y)_i\{M(x-y)\}_i\\ &\leqq \max_{1\leqq i\leqq n}(x-y)_i(r-q)_i\\ &\leqq \|x-y\|_\infty\|r-q\|_\infty.\end{aligned}$$

This gives the desired inequality. □

THEOREM 11. *Let $\mathscr{C}$ be a compact subset of $\mathscr{P}$ and $\mathscr{E}$ be a bounded subset of R^n. Then $\mathscr{S}$ is Lipschitzian on $\mathscr{C}\times\mathscr{E}$. In fact,*

$$\|\mathscr{S}(M,q)-\mathscr{S}(N,p)\|_\infty\leqq L(\|M-N\|_\infty+\|q-p\|_\infty)$$

for all (M,q) and (N,p) in $\mathscr{C}\times\mathscr{E}$ where

$$L=\max\{\theta\delta^{-2},\delta^{-1}\},\quad \delta=\min_{M\in\mathscr{C}}\alpha(M),\quad \theta=\max_{q\in\mathscr{E}}\|q\|_\infty.$$

Proof. First of all, we observe that L is finite since α (as a function on $\mathscr{P}$) is continuous and $\mathscr{C}$ is compact. Now let (M,q) and (N,p) be any two elements of $\mathscr{C}\times\mathscr{E}$. Let $\mathscr{S}(M,q)=\{x\}$ and $\mathscr{S}(N,p)=\{z\}$. By putting $r=0$ in (17), we get

$$\|x\|_\infty\leqq\alpha(M)^{-1}\|q\|_\infty.$$

Since x is the only solution of LCP $(N,(M-N)x+q)$, we deduce from Theorem 11 (applied to N) that

$$\begin{aligned}\|x-z\|_\infty &\leqq \alpha(N)^{-1}\|(M-N)x+q-p\|_\infty\\ &\leqq \alpha(N)^{-1}\{\|M-N\|_\infty\|x\|_\infty+\|q-p\|_\infty\}\\ &\leqq \alpha(N)^{-1}\{\alpha(M)^{-1}\|M-N\|_\infty\|q\|_\infty+\|q-p\|_\infty\}\\ &\leqq \delta^{-1}\{\theta\delta^{-1}\|M-N\|_\infty+\|q-p\|_\infty\}\\ &\leqq L(\|M-N\|_\infty+\|q-p\|_\infty).\end{aligned}$$

This completes the proof. □

Remarks. Note that in Theorem 10 the vectors q and r are unrestricted while in Theorem 11 the vectors q and p vary over a bounded set $\mathscr{E}$. It is easy to see that Theorem 11 is valid for $\mathscr{E}=R^n$ if and only if $\mathscr{C}$ is a singleton set. Indeed, suppose that Theorem 11 is valid for $\mathscr{C}\times R^n$ for some $\mathscr{C}$. By putting $q=p=kr$ (where $r\in R^n$ is arbitrary) we deduce that

$$\|\mathscr{S}(M,kr)-\mathscr{S}(N,kr)\|_\infty\leqq L\|M-N\|_\infty$$

for all natural numbers k. Since $\mathscr{S}$ is positively homogeneous in the second variable, we can divide both sides of the above inequality by k and let k go to ∞ to conclude that $\mathscr{S}(M,r)=\mathscr{S}(N,r)$ for all $r\in R^n$. Let x be any nonnegative vector. Then x is an element in $\mathscr{S}(M,-Mx)=\mathscr{S}(N,-Mx)$ from which we get the inequality $Nx-Mx\geqq 0$. Similarly, $Mx-Nx\geqq 0$ so that $Mx=Nx$. Since $x\geqq 0$ is arbitrary, we conclude that $M=N$.

5. Lipschitz continuity of $\mathscr{S}$ when M is fixed. In this section we fix the matrix $M\in R^{n\times n}$ and consider $\mathscr{S}$ as a function of q alone. We write

$$\Phi(q):=\mathscr{S}(M,q)$$

and

$$\operatorname{dom}\Phi=\{q\in R^n: \Phi(q)\neq\emptyset\}.$$

We say that Φ is *upper semicontinuous* (*lower semicontinuous*) at q if $\mathcal{S}$ is $\{M\}$-upper semicontinuous ($\{M\}$-lower semicontinuous). By specializing Definition 3 to $\mathcal{C}=\{M\}$ and $\mathcal{E}=R^n$, we get the following types of Lipschitz continuity for Φ.

(a) We say that Φ is *locally upper Lipschitzian* at q if there is a positive number λ and an $\varepsilon>0$ such that for all $q'\in R^n$ satisfying $\|q'-q\|<\varepsilon$ one has the inclusion

$$\Phi(q')\subseteq\Phi(q)+\lambda\|q'-q\|B. \tag{18}$$

(b) We say that Φ is *locally lower Lipschitzian* at q if there is a positive number λ and an $\varepsilon>0$ such that for all $q'\in\operatorname{dom}\Phi$ satisfying $\|q'-q\|<\varepsilon$ one has the inclusion

$$\Phi(q)\subseteq\Phi(q')+\lambda\|q'-q\|B. \tag{19}$$

(c) Φ is said to be *Lipschitzian* if there is a positive number L such that for all q and p in $\operatorname{dom}\Phi$,

$$\Phi(q)\subseteq\Phi(p)+L\|q-p\|B. \tag{20}$$

In view of Theorem 1, for any matrix M, the corresponding Φ is locally upper Lipschitzian at all q. Note that one single λ works for all q. We shall see later (in the proof of Theorem 14) that the existence of such a universal λ is useful. As noted earlier, Theorem 1 shows that Φ is upper semicontinuous at all q. It is obvious that Φ is lower semicontinuous at q whenever it is locally lower Lipschitzian at q.

In the definitions (a) and (b) above, the vector q' is allowed to vary in a neighborhood of q. By specializing Theorem 9 to $\mathcal{C}=\{M\}$, we deduce that these local upper or lower Lipschitzian properties are equivalent to "global" upper or lower Lipschitzian properties when the matrix is a $\mathbf{R}_0$-matrix.

THEOREM 12. *Suppose that M is fully semimonotone. Then the following are equivalent*:

(a) *Φ is Lipschitzian.*

(b) *Φ is locally lower Lipschitzian at all $q\in\operatorname{dom}\Phi$.*

(c) *Φ is lower semicontinuous at all $q\in\operatorname{dom}\Phi$.*

(d) *M is a $\mathbf{P}$-matrix.*

Proof. The implications (a)$\Rightarrow$(b) and (b)$\Rightarrow$(c) are obvious and the implication (d)$\Rightarrow$(a) follows from Theorem 10. We prove the implication (c)$\Rightarrow$(d). Suppose that Φ is lower semicontinuous at all $q\in\operatorname{dom}\Phi$. By Theorem 5 (applied to $\mathcal{C}=\{M\}$), M is an $\mathbf{R}_0$-matrix. Since M is assumed to be semimonotone, by [21], M is a $\mathbf{Q}$-matrix. By Theorems 6 and 7, for every $q\in R^n$, LCP (M,q) has a unique solution. It follows (see [29] or [19, Cor. 4.3]) that M is a $\mathbf{P}$-matrix. □

The above proof actually reveals that if M is fully semimonotone and Φ is lower semicontinuous at the zero vector and at q, then LCP (M,q) has a unique solution.

The above theorem naturally leads to the question of whether the implication (a)$\Rightarrow$(d) holds under more general conditions. This is answered in the results below. *We say that a matrix M is a* $\mathbf{G}$*-matrix if for some $d>0$, the zero vector is the only solution of* LCP (M,d). $\mathbf{G}$-matrices are generalizations of semimonotone matrices. For properties of such matrices, we refer the reader to [7] and [10].

THEOREM 13. *Let M be a* $\mathbf{G}$*-matrix. Then Φ is Lipschitzian if and only if M is a* $\mathbf{P}$*-matrix.*

The proof of this result is based on the following lemmas. In these lemmas we assume without loss of generality that the norm (of a vector) refers to the Euclidean norm.

LEMMA 2. *Let M be any matrix for which $\Phi(q)=\mathscr{S}(M,q)$ is Lipschitzian. If $\hat{M}$ is any principal pivot transform of M, then the mapping*

$$\Psi(q):=\mathscr{S}(\hat{M},q)$$

is also Lipschitzian.

Proof. Let $\hat{M}$ be obtained by pivoting on the principal submatrix $M_{\alpha\alpha}$ so that $\hat{M}$ is given by (6). Suppose that $\Phi(q)=\mathscr{S}(M,q)$ satisfies (20). Consider two vectors r and s in dom Ψ. Given any $x\in\Psi(r)$, there exist q and $p\in R^n$, $z\in\Phi(q)$ such that $\hat{z}=x$, $\hat{q}=r$, $\hat{p}=s$. From (20), there exists $u\in\Phi(p)$ such that

$$\|z-u\|\leqq L\|q-p\|.$$

In view of (8), the vector $\hat{u}\in\Psi(s)$ satisfies the inequality

$$\|\hat{z}-\hat{u}\|\leqq C_1\{\|z-u\|+\|q-p\|\}.$$

The above inequalities together with (9) give

$$\Psi(r)\subseteq\Psi(s)+C\|r-s\|B$$

for a suitable constant C depending on the matrix M. This completes the proof of the lemma. □

LEMMA 3. *If M is a* G*-matrix and Φ is Lipschitzian, then M is strictly semimonotone.*

Proof. Suppose that Φ is Lipschitzian and let d be a positive vector such that $\Phi(d)=\{0\}$. To get the desired result, it is enough to show that $\Phi(q)=\{0\}$ for all $q\geqq 0$. Now let Ω be the set of all positive vectors e such that $\Phi(e)=\{0\}$. The set Ω is nonempty because $d\in\Omega$. We claim that Ω is an open set. Fix an $e\in\Omega$. For all $u\in B$ we have from the Lipschitzian property,

$$\Phi(ke+u)\subseteq\Phi(ke)+L\|u\|B\quad\forall k=1,2,\dots.$$

Since $\Phi(ke)=k\Phi(e)=\{0\}$, the sets $\Phi(ke+u)$ are uniformly bounded. For any $x\in\Phi(ke+u)$ we have

$$x\geqq 0,\quad y=Mx+u+ke\geqq 0,\quad y^Tx=0.$$

It follows that for all large k, $x=0$ so that for all such k, $\Phi(ke+u)=\{0\}$. Since $ke+u>0$ for large k, the set $\{t(ke+u)\colon t>0,\ u\in B\}$ is an open set which contains e and which is contained in Ω. Hence Ω is an open set. Now we show that Ω is closed in R_{++} ($=$the set of all positive vectors in R^n). To this end, let $p\in R_{++}$ be in the closure of Ω. Since every neighborhood of p contains a point e of Ω and for any such point $\Phi(e)=\{0\}$, it follows from the Lipschitzian property that $\Phi(p)$ must be contained in an arbitrarily small ball around the zero vector. Hence $\Phi(p)=\{0\}$ so that $p\in\Omega$. Thus the set Ω is both open and closed in R_{++}. Since R_{++} is connected, $\Omega=R_{++}$. To complete the proof, let $q\geqq 0$. Then every neighborhood of q contains a point of Ω. By the Lipschitzian property we see that $\Phi(q)$ is contained in a ball of arbitrarily small radius. Since $0\in\Phi(q)$, we see that $\Phi(q)=\{0\}$. This completes the proof of the lemma. □

LEMMA 4. *Suppose that M is stricly semimonotone and Φ is Lipschitzian. Then every proper principal submatrix N of M is strictly semimonotone and the corresponding multifunction $\Lambda(r):=\mathrm{SOL}\,(N,r)$ is Lipschitzian.*

Proof. Since strict semimonotonicity property is inherited by all principal submatrices, we only show the Lipschitzian property of Λ. Let $N=M_{\alpha\alpha}$ where α is a proper subset of $\{1,2,\dots,n\}$. Without loss of generality let α be the first $|\alpha|$ natural numbers. Let β be the complement of α in $\{1, 2, \ldots, n\}$, $r, s \in R^{|\alpha|}$, and e be the vector of ones in $R^{|\beta|}$. For $k = 1, 2, \ldots$, let

$$r^k = \begin{bmatrix} r \\ ke \end{bmatrix}, \quad s^k = \begin{bmatrix} s \\ ke \end{bmatrix}, \quad e^k = \begin{bmatrix} 0 \\ ke \end{bmatrix}.$$

We claim that for all large k,

$$\Phi(r^k) = \left\{ \begin{bmatrix} u \\ 0 \end{bmatrix} : u \in \Lambda(r) \right\}. \tag{21}$$

To see this, we observe, from the Lipschitzian property, the inclusion

$$\Phi(r^k) \subseteq \Phi(e^k) + L\|r\|B.$$

Since strict semimonotonicity of M implies that $\Phi(e^k) = \{0\}$ for all k, we see from the above inclusion that $\Phi(r^k)$ is uniformly bounded for all k. It follows that for large k, the β-part of any solution of LCP (M, r^k) is zero. This gives one inclusion toward the equality of sets in (21). The reverse inclusion follows immediately from the observation that since N is strictly semimonotone, it is an $\mathbf{R}_0$-matrix and hence $\Lambda(r)$ is bounded. We thus have (21) for large k. Now fix r and s in dom Λ. The Lipschitzian property of M gives

$$\Phi(r^k) \subseteq \Phi(s^k) + L\|r - s\|B$$

for all k. By choosing a large k we can apply (21) and conclude that

$$\Lambda(r) \subseteq \Lambda(s) + L\|r - s\|B',$$

where B' denotes the open unit ball in $R^{|\alpha|}$. Thus we have shown that Λ is Lipschitzian. □

Proof of Theorem 13. In view of Theorem 10, we need only show that M is a **P**-matrix whenever Φ is Lipschitzian. Because of Lemma 3, it is enough to show that: If a $k \times k$ matrix M is strictly semimonotone and the corresponding multifunction Φ is Lipschitzian, then M is a **P**-matrix. We induct on k. When $k = 1$, the entry in M is positive and hence the matrix is a **P**-matrix. Now assume the result for all $k = 1, 2, \ldots, (n-1)$. Let M be an $n \times n$ matrix that is strictly semimonotone and whose multifunction Φ is Lipschitzian. From Lemma 4 and the induction hypothesis we conclude that every proper principal submatrix of M is a **P**-matrix. To complete the proof we need only show that M has positive determinant. Since M is strictly semimonotone, every entry in the diagonal of M is positive. Let a denote the (1, 1)-entry in M. We put $\alpha = \{1\}$, $r = [-1]$, and $N = [a]$ and we define r^k as in the proof of Lemma 4. We choose k large so that (21) holds and $\hat{r}^k > 0$ where $(\hat{M}, \hat{r}^k)$ is obtained from (M, r^k) by pivoting on the (1, 1)-entry in M. For brevity, we write q for r^k. Since SOL (N, r) is a singleton set, by (21),

$$|\mathcal{S}(M, q)| = 1.$$

Furthermore, $|\mathcal{S}(\hat{M}, \hat{q})| = |\mathcal{S}(M, q)| = 1$. Since $\hat{q} > 0$, $\mathcal{S}(\hat{M}, \hat{q}) = \{0\}$, i.e., $\hat{M}$ is a **G**-matrix. Since Φ is assumed to be Lipschitzian, by Lemma 2, the multifunction Ψ corresponding to $\hat{M}$ is Lipschitzian. By Lemma 3, $\hat{M}$ is strictly semimonotone. From Lemma 4, we conclude that the principal submatrix $T := \hat{M}_{(n-1)(n-1)}$ is strictly semimonotone and its corresponding multifunction is Lipschitzian. From the induction hypothesis, this submatrix is a **P**-matrix. In particular, the determinant of T is positive. But T is the Schur complement [1] of the matrix $M_{11} = N$, and so the well-known Schur formula [1]

$$\det M = (\det M_{11})(\det T)$$

holds. We conclude that the determinant of M is positive. This completes the proof of the theorem. □

The following are easy consequences of Theorem 13.

COROLLARY 4. *Suppose that M is a* **Q**-*matrix for which Φ is Lipschitzian. If some principal pivot transform of M is a* **G**-*matrix, then M is a* **P**-*matrix.*

Proof. If some principal pivot transform $\hat{M}$ of M is a **G**-matrix, then by Lemma 2 and Theorem 13, $\hat{M}$ is a **P**-matrix. Hence M is a **P**-matrix. □

COROLLARY 5. *Suppose that M is a* **Q**-*matrix for which Φ is Lipschitzian. If there is a nondegenerate q such that* LCP (M, q) *has a unique solution, then M is a* **P**-*matrix.*

Proof. Let x^* be the unique solution of LCP (M, q) with $x^* + Mx^* + q > 0$. Since the supporting submatrix corresponding to x^* is nonsingular [13], by pivoting on this submatrix we can get $(\hat{M}, \hat{q}, \hat{x}^*)$. Since $|\mathcal{S}(\hat{M}, \hat{q})| = |\mathcal{S}(M, q)| = 1$ and $\hat{q} \geqq 0$, we see that $\hat{x}^* = 0$ and hence $\hat{q} = \hat{x}^* + \hat{M}\hat{x}^* + \hat{q} = x^* + Mx^* + q > 0$. This shows that $\hat{M}$ is a **G**-matrix. We now apply the previous corollary to get the result. □

Remarks. It is not known at this stage whether the above corollary is valid without the existence of a nondegenerate vector for which the LCP has a unique solution. It is also not known whether Theorem 13 holds under the weaker assumption of Φ being locally lower Lipschitzian at every $q \in \operatorname{dom} \Phi$.

Our next result deals with **N**-matrices. Recall that a matrix M is an **N**-matrix [14], [25] if every principal minor of M is negative.

THEOREM 14. *Let M be an* **N**-*matrix with all entries negative. Then Φ is Lipschitzian.*

Proof. Since M is a negative matrix, the domain of Φ is R^n_+. Also $\Phi(q) = \{0\}$ for all $q \geqq 0$, $q \not> 0$. By Kojima and Saigal [14] or by Parthasarathy and Ravindran [25], for every $q > 0$ we have $\Phi(q) = \{0, f(q)\}$ where $f(q)$ is the unique nonzero solution of LCP (M, q). To establish the Lipschitzian property of Φ, we first prove the existence of a positive constant C such that

$$\|f(q) - f(p)\| \leqq C\|q - p\|B \quad \forall q > 0, \quad p > 0. \tag{22}$$

We proceed as in Lemma 3.1 of Mangasarian and Shiau [18]. Corresponding to each $J \subseteq \{1, 2, \ldots, n\}$, we let $Q(J)$ denote the set of all vectors q for which the system

$$\begin{aligned} M_j x + q_j &\geqq 0, \quad x_j = 0 \quad \forall j \in J, \\ M_j x + q_j &= 0, \quad x_j \geqq 0 \quad \forall j \notin J \end{aligned} \tag{23}$$

has a solution where M_j denotes the jth row of M, etc. We observe that $Q(\{1, 2, \ldots, n\}) = R^n_+$. If $J \neq \{1, 2, \ldots, n\}$ and $0 < q \in Q(J)$, then $f(q)$ is the unique solution of (23). Now for any $q > 0$ and $p > 0$, the line segment

$$[q, p] = \{(1-t)q + tp: 0 \leqq t \leqq 1\}$$

is contained in the union of $Q(J)$ as J varies over all proper subsets of $\{1, 2, \ldots, n\}$ including the empty set. (Note that each $q > 0$ must be in the above union since (23) corresponding to $\{1, 2, \ldots, n\}$ gives only one solution, namely zero, and we know that $\Phi(q)$ contains a nonzero solution.) The proof of Lemma 3.1 in [18] can be modified to show that Lemma 3.1 is valid with J_i contained in but not equal to $\{1, 2, \ldots, n\}$. Using this and the argument employed in the proof of Theorem 3.2 in [18], we deduce the existence of a C satisfying (22). Now suppose that q and p are arbitrary. Since the Lipschitzian property is obvious if both q and p have (some) zero components, assume without loss of generality that $0 \not< q \geqq 0$ and $p > 0$. By Theorem 1, there is a universal constant λ such that

$$\Phi(q') \subseteq \Phi(q) + \lambda\|q' - q\|B$$

for all q' near q. Now let $q' > 0$ be the line segment $[q, p]$ close to q so that the above inclusion holds. We also have, from (22),

$$\Phi(p) \subseteq \Phi(q') + C\|p - q'\|B.$$

Above inclusions give

$$\Phi(p) \subseteq \Phi(q) + (C + \lambda)\|p - q\|B.$$

Finally, $\Phi(q) = \{0\} \subseteq \{0, f(p)\} \subseteq \Phi(p) + (C + \lambda)\|p - q\|B$. Thus we have shown that Φ has the Lipschitzian property with the Lipschitzian constant $C + \lambda$. This completes the proof. □

From Theorems 13 and 14, we infer that when M is either a **P**-matrix or an **N**-matrix with all negative entries, the corresponding Φ is Lipschitzian. Since every negative **N**-matrix is a **Z**-matrix (i.e., a matrix with nonpositive off-diagonal entries), it follows that Theorem 13 is not valid for **Z**-matrices. The following example shows that the Lipschitzian property of Φ is not limited to **P** and **N** matrices.

Example 5. Let

$$M = \begin{bmatrix} -1 & -2 \\ -2 & -1 \end{bmatrix}, \qquad \hat{M} = \begin{bmatrix} -1 & -2 \\ 2 & 3 \end{bmatrix}.$$

Since M is an **N**-matrix with negative entries, by Theorem 14, the mapping Φ corresponding to M is Lipschitzian. By Lemma 2 the mapping $\hat{\Psi}$ corresponding to the principal pivot transform $\hat{M}$ is Lipschitzian. (Note that $\hat{M}$ is obtained from M by pivoting on the (1, 1) entry in M.) We observe that $\hat{M}$ is neither a **P**-matrix nor an **N**-matrix.

Our next example shows that Theorem 14 is false for an arbitrary **N**-matrix.

Example 6. Let

$$M = \begin{bmatrix} -1 & 1 \\ 2 & -1 \end{bmatrix}, \quad q = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \quad q' = \begin{bmatrix} -\varepsilon \\ 1 \end{bmatrix},$$

where $\varepsilon > 0$ and small. It is clear that M is an **N**-matrix. For this M, $\Phi(q)$ has two elements and $\Phi(q')$ has exactly one element for any ε. It is clear that the inclusion

$$\Phi(q) \subseteq \Phi(q') + L\|q - q'\|B$$

cannot hold for a fixed L and all ε. Thus Φ corresponding to M is not Lipschitzian.

6. Concluding remarks. In this paper we have described continuity properties of the solution map $\mathcal{S}$. The results proved in this paper naturally lead to several interesting questions. For example, is Theorem 3 (or a variation of that) true for other matrices, such as **L**-matrices and **G**-matrices? Are Theorems 6 and 7 true for other types of matrices? Is it possible to characterize matrices M for which Φ is Lipschitzian? The answers to all of the above questions are not presently known. We end this paper by noting that the question of Pang (stated in the introduction) remains unresolved.

REFERENCES

[1] R. W. COTTLE, *Manifestations of the Schur complement*, Linear Algebra Appl., 8 (1974), pp. 189-211.

[2] ———, *Completely Q-matrices*, Math. Programming, 19 (1980), pp. 347-351.

[3] R. W. COTTLE AND G. B. DANTZIG, *Complementary pivot theory of mathematical programming*, Linear Algebra Appl., 1 (1968), pp. 103-125.

[4] R. W. COTTLE, J.-S. PANG, AND R. E. STONE, *The Linear Complementarity Problem*, Academic Press, Boston, 1992.

[5] R. D. DOVERSPIKE, *Some perturbation results for the linear complementarity problem*, Math. Programming, 23 (1982), pp. 181-192.

[6] B. C. EAVES, *The linear complementarity problem*, Management Sci., 17 (1971), pp. 612-634.

[7] C. B. GARCIA, *Some classes of matrices in linear complementarity theory*, Math. Programming, 5 (1973), pp. 299-310.

[8] M. S. GOWDA AND J.-S. PANG, *On solution stability of the linear complementarity problem*, Math. Oper. Res., 17 (1992), pp. 77-83.

[9] ———, *Some existence results for multivalued complementarity problems*, Res. Rep., Dept. of Mathematics and Statistics, Univ. of Maryland Baltimore County, Baltimore, MD, April 1990 (revised November 1990); Math. Oper. Res., to appear.

[10] ———, *The basic theorem of complementarity revisited*, Res. Rep., Dept. of Mathematics and Statistics, Univ. of Maryland Baltimore County, Baltimore, MD, October 1990 (revised April 1991); Math. Programming, to appear.

[11] C. D. HA, *Stability of the linear complementarity problem at a solution point*, Math. Programming, 31 (1985), pp. 327-338.

[12] ———, *Application of degree theory in stability of the complementarity problem*, Math. Oper. Res., 12 (1987), pp. 368-376.

[13] M. J. M. JANSEN AND S. H. TIJS, *Robustness and nondegenerateness for linear complementarity problems*, Math. Programming, 37 (1987), pp. 293-308.

[14] M. KOJIMA AND R. SAIGAL, *On the number of solutions to a class of linear complementarity problems*, Math. Programming, 17 (1979), pp. 136-139.

[15] C. E. LEMKE, *On complementary pivot theory*, in Mathematics of Decision Sciences, Part I, G. B. Dantzig and A. F. Veinott, Jr., eds., American Mathematical Society, Providence, RI, 1968, pp. 95-114.

[16] O. L. MANGASARIAN, *Characterizations of bounded solutions of linear complementarity problems*, Math. Programming Stud., 19 (1982), pp. 153-166.

[17] O. L. MANGASARIAN AND T.-H. SHIAU, *Error bounds for monotone linear complementarity problems*, Math. Programming, 36 (1986), pp. 81-89.

[18] ———, *Lipschitz continuity of solutions of linear inequalities, programs, and complementarity problems*, SIAM J. Control Optim., 25 (1987), pp. 583-595.

[19] K. G. MURTY, *On the number of solutions to the linear complementarity problem and spanning properties of complementary cones*, Linear Algebra Appl., 5 (1972), pp. 65-108.

[20] ———, *Linear complementarity, linear and nonlinear programming*, Heldermann-Verlag, West Berlin, Germany, 1987.

[21] J.-S. PANG, *On Q-matrices*, Math. Programming, 17 (1979), pp. 243-247.

[22] ———, *Two characterization theorems in complementarity theory*, Oper. Res. Lett., 7 (1988), pp. 27-31.

[23] ———, *Newton's method for B-differentiable equations*, Math. Oper. Res., 15 (1990), pp. 311-341.

[24] T. D. PARSONS, *Applications of principal pivoting*, in Proceedings of the Princeton Symposium on Mathematical Programming, H. W. Kuhn, ed., Princeton University Press, Princeton, NJ, 1970, pp. 567-581.

[25] T. PARTHASARATHY AND G. RAVINDRAN, *N-matrices*, Linear Algebra Appl., 139 (1990), pp. 89-102.

[26] S. M. ROBINSON, *Generalized equations and their solutions, Part* I; *Basic theory*, Math. Programming Stud., 10 (1979), pp. 128-141.

[27] ———, *Some continuity properties of polyhedral multifunctions*, Math. Programming Stud., 14 (1981), pp. 206-214.

[28] R. MATHIAS AND J.-S. PANG, *Error bounds for the linear complementarity problem with a P-matrix*, Linear Algebra Appl., 132 (1990), pp. 123-136.

[29] H. SAMELSON, R. M. THRALL, AND O. WESLER, *A partition theorem for Euclidean n-space*, Proc. Amer. Math. Soc., 9 (1958), pp. 805-807.

[30] R. E. STONE, *Geometric aspects of the linear complementarity problem*, Ph.D. thesis, Dept. of Operations Research, Stanford Univ., Stanford, CA, 1981.

SIAM J. OPTIMIZATION
Vol. 2, No. 4, pp. 635-648, November 1992

006

LINEAR INEQUALITY SCALING PROBLEMS*

URIEL G. ROTHBLUM†

Abstract. Let $a \in R^m$, $b \in R^m$, and $C \in R^{m\times n}$ be given, where a is strictly positive. A *C-scaling* of the vector a is defined to be a vector $a' \in R^n$ with $a'_j = a_j[\Pi_{k=1}^m (u_k)^{C_{kj}}]$ for some strictly positive vector $u \in R^m$. The problem of finding a C-scaling of the vector a that satisfies the linear system $Cx = b$ is called the *linear equality scaling problem* (LESP). The current paper considers the *linear inequality scaling problem* (LISP), which concerns the identification of a C-scaling of a which satisfies the linear inequality system $Cx \leqq b$, where it is required that $u \leqq 1$ and that $u_i = 1$ for each i with $(Cx)_i < b_i$. It is shown that LISP generalizes LESP and that it unifies a number of matrix-scaling problems that have been studied recently. Further, it is shown that LISP can be reduced to one of two convex optimization problems and these reductions are used to characterize solutions to LISP and to derive necessary and sufficient conditions for their existence. In addition, uniqueness of solutions is established and perturbed relaxations of LISP are considered.

Key words. scaling, inequality systems, linear

AMS(MOS) subject classifications. 15A39, 15A12, 90C25

1. Introduction. A *scaling* of an array of numbers, e.g., a vector or a matrix, is obtained by multiplying each of the elements of that array by a corresponding positive number where some structure is imposed on the multiplying coefficients. For example, if $A = (A_{ij})$ is a (rectangular) matrix and $B = (B_{ij})$ is a matrix with $B_{ij} = d_i A_{ij} e_j$ for some positive vectors d and e, we call B a (D, E)-scaling of A. Also, if $A = (A_{ij})$ is a square matrix and $B = (B_{ij})$ satisfies $B_{ij} = d_i A_{ij} (d_j)^{-1}$ for some positive vector d, we call B a (D, D^{-1})-scaling of A. *Scaling problems* concern the identification of scalings of given arrays where certain specified properties are to be satisfied. Problems where the target properties are defined via linear equations are of particular interest; see Bacharach (1970), King (1981), Eaves et al. (1985), Bapat (1982), Raghavan (1984, 1985), Schneider and Zenios (1987), Rothblum and Schneider (1989), and references therein. A unifying approach to such problems was recently derived and analyzed by Bapat and Raghavan (1989), Franklin and Lorenz (1989), and Rothblum (1989). The purpose of this paper is to extend the formulation of scaling problems to situations where the desired properties concern linear inequalities. Instances of such matrix scaling problems were recently considered by Balinski and Demange (1987, 1988) and Schneider (1989, 1989a); see § 5 for further details. We also note that our approach generalizes the one for linear equations by the standard transformation of equality constraints to pairs of inequalities; see the end of this section and § 4 for further details.

Before presenting the scaling problem to be considered in this paper we summarize some definitions and notation that will be used. We call a vector $w \in R^p$ *nonnegative*, written $w \geqq 0$, if all the coordinates of w are nonnegative; we call w *strictly positive*, written $w \gg 0$, if all the coordinates of w are positive; and we call w *semipositive*, written $w > 0$, if $w \geqq 0$ and $w \neq 0$. Also, we write $w \leqq v$, $w \ll v$, or $w < v$, for w, $v \in R^p$ if, respectively, $v - w \geqq 0$, $v - w \gg 0$, or $v - w > 0$. Corresponding notation will be used for matrices. The notation $\| \|_\infty$ will be used for the l_∞ norm, i.e., for a vector $x \in R^p$, $\|x\|_\infty = \max\{|x_i|: 1 \leqq i \leqq p\}$. Finally, we denote by 1 the vector whose coordinates are all 1. The dimension of this vector will be clear from the context.

*Received by the editors October 23, 1989; accepted for publication (in revised form) October 25, 1991.

† Faculty of Industrial Engineering and Management, Technion-Israel Institute of Technology, Haifa 32000, Israel and Rutgers Center for Operations Research (RUTCOR), Rutgers University, P.O. Box 5062, New Brunswick, New Jersey 08904.

Throughout the remainder of this paper, let $a \in R^n$, $b \in R^m$, and $C \in R^{m \times n}$ be given where a is a strictly positive vector. The *linear inequality scaling problem* (LISP) *with data a, b, and C* is defined to be the problem of finding vectors $u \in R^m$ and $a' \in R^n$ satisfying

$$0 \ll u \leqq 1, \tag{1}$$

$$a'_j = a_j \left[\prod_{i=1}^{m} (u_i)^{C_{ij}} \right] \quad \text{for } j = 1, \dots, n, \tag{2}$$

$$Ca' \leqq b, \tag{3}$$

and

$$u_i = 1 \quad \text{whenever } (Ca')_i < b_i. \tag{4}$$

Studying solutions to (1)-(4) is the main goal of our paper. In particular, we characterize such solutions and obtain necessary and sufficient conditions for their existence and for the existence of solutions for perturbed relaxations of the problem (to be formally defined in§ 3). Also, we derive uniqueness results.

Of course, if a is nonnegative rather than strictly positive, then for every C-scaling a' of a, $a'_i = 0$ whenever $a_i = 0$. It follows that when considering the LISP with data a, b, and C, we can drop the zero coordinates of a and the corresponding columns of C. Thus, without loss of generality, we may assume that a is strictly positive. Still, in applications where the vector a represents a matrix, the inclusion of the zero coordinates is convenient and natural because it facilitates the use of matrix notation. When we develop the theoretical analysis of the LISP we will assume that the given vector a is strictly positive; but, in the discussion of examples, we allow zero coordinates whenever convenient.

A vector a' is called a *C-scaling* of the vector a if a' has the representation (2) for some strictly positive vector $u \in R^m$, in which case u is called a *weight vector corresponding to a'*. Such a weight vector is called *normalized* if $u \leqq 1$, i.e., if $\|u\|_\infty \leqq 1$. We call a C-scaling a' of *a normalized* if it has a corresponding weight vector that is normalized. Thus LISP is the problem of finding a normalized C-scaling a' of a that satisfies the linear inequalities

$$\sum_{i=1}^{m} C_{ij} a'_j \leqq b_i \quad \text{for } i = 1, \dots, m, \tag{5}$$

such that for some normalized weight vector u corresponding to a' we have that (5) holds with equality for each i for which $u_i < 1$.

The *linear equality scaling problem* (LESP) *with data a, b, and C*, is defined as the problem of finding a C-scaling a' of a that satisfies $Ca' = b$. We note that given a, b, and C, the LISP with data a, $\left(\begin{smallmatrix} b \\ -b \end{smallmatrix}\right)$, and $\left(\begin{smallmatrix} C \\ -C \end{smallmatrix}\right)$ is the problem of finding v, w, $\in R^m$ and $a' \in R^n$ satisfying

$$0 \ll v \leqq 1, \qquad 0 \ll w \leqq 1,$$

$$a'_j = a_j \left[\prod_{i=1}^{m} (v_i w_i^{-1})^{C_{ij}} \right] \quad \text{for } j = 1, \dots, n,$$

and

$$Ca' = b.$$

Taking the change of variable $u_i = v_i / w_i$, the above problem reduces to finding $u \in R^m$ and $a' \in R^n$ that satisfy

$$0 \ll u, \quad a'_j = a_j \left[\prod_{i=1}^{m} (u_i)^{C_{ij}} \right] \quad \text{for } j = 1, \ldots, n,$$

and

$$Ca' = b,$$

which is the linear equality scaling problem with data a, b, and C. So, the LESP is a special instance of the LISP.

Our development relies on the identification of two convex optimization problems whose optimal solutions correspond to the solution of the LISP. Variants of these optimization problems were previously applied by Bapat and Raghavan (1989), Franklin and Lorenz (1989), and Rothblum (1989) to study the LESP. We show in § 4 how the results of the current paper can be used to derive results for LESP obtained in the above references.

2. Characterization. We will find it useful to apply the change of variables $s_i = \ln(u_i)$ to (1)–(4), resulting in the following system:

$$s \leqq 0, \tag{1'}$$

$$a'_j = a_j \exp\left(\sum_{i=1}^{m} s_i C_{ij} \right) \quad \text{for } j = 1, \ldots, n, \tag{2'}$$

$$Ca' \leqq b, \tag{3'}$$

and

$$s_i[(Ca')_i - b_i] = 0 \quad \text{for } i = 1, \ldots, m. \tag{4'}$$

In particular, a pair (a', u) satisfies (1)–(4) if and only if the pair (a', s) satisfies (1′)–(4′). So, the LISP reduces to the problem of identifying a solution to (1′)–(4′).

The following nonlinear optimization problems are useful for studying the LISP:

$$\text{Program I:} \quad \min \sum_{j=1}^{n} x_j [\ln(x_j / a_j) - 1]$$

$$\text{subject to} \quad Cx \leqq b, \quad x \geqq 0,$$

where, as usual, $x_j[\ln(x_j)]$ is defined as zero when $x_j = 0$, and

$$\text{Program II:} \quad \min \sum_{j=1}^{n} a_j \exp\left(\sum_{i=1}^{m} y_i C_{ij} \right) - \sum_{i=1}^{m} y_i b_i$$

$$\text{subject to} \quad y \leqq 0.$$

We note that the objective functions of these programs are convex. This fact is immediate for the objective of Program I (by differentiation twice) and was verified for Program II in Rothblum (1989, Lemma 1). Also, standard arguments show that Program II is equivalent to the dual of Program I; see the Appendix in Rothblum (1989a) for details.

The next two lemmas use convexity arguments to characterize optimal solutions of the above two programs. In particular, the issue in Lemma 1 is a careful examination of the boundary of the feasible region of Program I. Our approach here follows Franklin and Lorenz (1989), who considered a variant of Program I with the equality constraints $Cx = b$ rather than the inequality constraints $Cx \leqq b$. The approach in the analysis of

Program II in Lemma 2 follows Rothblum (1989), who considered the unconstrained variant of Program II. Proofs of the two lemmas are included for the sake of completeness.

LEMMA 1. *Assume that Program* I *is feasible. Then*:

(a) *Program* I *has a unique solution*;

(b) *if* x^* *is the optimal solution of Program* I *and* x *is any feasible solution of that program, then* $\{j=1,\dots,n: x_j^*>0\}\supseteq\{j=1,\dots,n: x_j>0\}$;

(c) *the optimal solution of Program* I *is strictly positive if and only if* $\{x\in R^n: Cx\leqq b,\ x\gg 0\}\neq\varnothing$; *and*

(d) *a strictly positive vector* $x\in R^n$ *is the optimal solution of Program* I *if and only if there exists a vector* λ *in* R^m *such that* (x,λ) *satisfy*

$$\ln(x_j/a_j)+(\lambda^T C)_j=0 \quad \textit{for } j=1,\dots,n, \tag{6}$$

$$Cx\leqq b, \tag{7}$$

$$\lambda\geqq 0 \tag{8}$$

and

$$\lambda_i(Cx-b)_i=0 \quad \textit{for } i=1,\dots,m. \tag{9}$$

Proof. Let $R_+=\{t\in R: t\geqq 0\}$. For $j=1,\dots,n$ consider the function $f_j: R_+\to R$ defined by $f_j(t)=t[\ln(t/a_j)-1]$ for $t>0$ while $f_j(0)=0$. Each f_j is continuous and strictly convex on R_+, is twice differentiable on $R_+\backslash\{0\}$, attains a unique minimum over R_+ at a_j, and $f_j(a_j)=-a_j$.

Let $\underset{\sim}{x}$ be any (fixed) feasible solution of Program I. As $\lim_{t\to\infty} f_j(t)=\infty$ for each j, we have that for some $K_j>0$, $f_j(t)\geqq f(\underset{\sim}{x})+\sum_j a_j$ for all $t>K_j$. Let $K\equiv\max\{K_1,\dots,K_n\}$. Now, if y is feasible for Program I and $\|y\|_\infty>K$, then for some $k=1,\dots,n$, $y_k>K_k$ and

$$f(y)\geqq f_k(y_k)+\sum_{j\neq k} f_j(a_j)\geqq\left[f(\underset{\sim}{x})+\sum_j a_j\right]+\sum_{j\neq k}(-a_j)>f(\underset{\sim}{x}).$$

Hence, $\min\{f(y): Cy\leqq b,\ y\geqq 0\}=\min\{f(y): Cy\leqq b, y\geqq 0,\ \|y\|_\infty\leqq K\}$ and standard compactness arguments show that Program I attains a minimum. The uniqueness of the optimal solution of Program I follows immediately from the convexity of its feasible region and the strict convexity of each of the functions f_j. So, part (a) has been established.

Let x^* be the (unique) optimal solution of Program I and let x be any feasible solution of that program. For each $0\leqq\varepsilon\leqq 1$, let $x(\varepsilon)\equiv(1-\varepsilon)x^*+\varepsilon x$ and let $J\equiv\{j=1,\dots,n: x_j^*>0 \text{ or } x_j>0\}$. Then, for $0\leqq\varepsilon\leqq 1$, $x(\varepsilon)$ is feasible for Program I; furthermore, for $0<\varepsilon<1$, $x(\varepsilon)_j>0$ and

$$(d/d\varepsilon)f[x(\varepsilon)]=\sum_{j\in J}(x_j-x_j^*)\ln[x(\varepsilon)_j/a_j].$$

Let $J_0\equiv\{j\in J: x_j^*=0\}$ and we will show that $J_0=\varnothing$. Now, for $j\in J_0$, $x_j>0$ and

$$(d/d\varepsilon)f[x(\varepsilon)]=\sum_{j\in J_0} x_j\ln(\varepsilon x_j)+0(1)=\left(\sum_{j\in J_0} x_j\right)\ln(\varepsilon)+0(1) \quad \text{as } \varepsilon\to 0.$$

So, if $J_0\neq\varnothing$, then $(\sum_{j\in J_0} x_j)\neq 0$ and the limit of the above derivative as $\varepsilon\to 0$ is $-\infty$, implying that $f[x(\varepsilon)]<f(x^*)$ for all sufficiently small positive ε. This contradicts the optimality of x^* and therefore proves that $J_0=\varnothing$, i.e., $\{j=1,\dots,n: x_j^*>0\}\supseteq\{j=1,\dots,n: x_j>0\}$. So the proof of (b) is complete. Also, (c) is immediate from (b).

We finally establish (d). We first show that if x^*—the optimal solution of Program I—is strictly positive, then there exists a vector λ in R^m such that the pair (x^*, λ) satisfies (6)-(9). Let $\mathrm{I} \equiv \{i = 1, \ldots, m: (Cx^*)_i = b_i\}$ and let z be any vector in R^n satisfying $(Cz)_i \leqq 0$ for all $i \in \mathrm{I}$. Then for sufficiently small positive ε, $x^* + \varepsilon z$ is strictly positive and is feasible for Program I; in particular, for such ε, $f(x^* + \varepsilon z) \geqq f(x^*)$, implying that

$$[(\nabla f)(x)|_{x=x^*}]^T z = (d/d\varepsilon) f(x^* + \varepsilon z)|_{\varepsilon=0} \geqq 0.$$

Thus, letting C_I be the submatrix of C corresponding to the rows of I, we have that

$$C_\mathrm{I} z \leqq 0 \quad \text{implies that } -[(\nabla f)(x)|_{x=x^*}]^T z \leqq 0.$$

By Farkas's Lemma (see, e.g., Schrijver (1986, Lemma 7.1d, p. 89)), we conclude that for some nonnegative vector $\lambda \in R^m$ with $\lambda_i = 0$ for all $i \in \{1, \ldots, m\} \backslash \mathrm{I}$,

$$\lambda^T C = -[(\nabla f)(x)]^T|_{x=x^*} - (\ln(x_1/a_1), \ldots, \ln(x_n/a_n)).$$

So, λ and x^* satisfy (6)-(9), the latter following from the fact that $\lambda_i = 0$ for all $i \in \{1, \ldots, m\} \backslash \mathrm{I}$.

Next assume that $\underset{\sim}{x}$ is a positive vector in R^n such that $\underset{\sim}{x}$ and a vector λ in R^m satisfy (6)-(9). Suppose that $\underset{\sim}{x} \neq x^*$. As $\underset{\sim}{x}$ is feasible for Program I, and x^* is the unique optimal solution of Program I, the convexity of f assures that

$$[(\nabla f)(x)|_{x=\underset{\sim}{x}}]^T (x^* - \underset{\sim}{x}) = (d/d\varepsilon) f[(1-\varepsilon)\underset{\sim}{x} + \varepsilon x^*]|_{\varepsilon=0} \leqq f(x^*) - f(\underset{\sim}{x}) < 0.$$

As (6) asserts that $[(\nabla f)(x)|_{x=\underset{\sim}{x}}]^T = -\lambda^T C$, we get from the above, (9), (8), and the feasibility of x^* for Program I, that

$$\begin{aligned} 0 > [(\nabla f)(x)|_{x=\underset{\sim}{x}}]^T (x^* - \underset{\sim}{x}) &= -\lambda^T C (x^* - \underset{\sim}{x}) = -\lambda^T (Cx^* - b) + \lambda^T (C\underset{\sim}{x} - b) \\ &= -\lambda^T (Cx^* - b) \geqq 0, \end{aligned}$$

a contradiction, which proves that $x = x^*$. □

We observe that part (d) of the above lemma is a statement about the necessity and sufficiency of the Kuhn-Tucker conditions for optimality for Program I. The elaborate arguments were needed because of the nondifferentiability of f on the boundary of its domain. A general result about the necessity and sufficiency of Kuhn-Tucker conditions for convex optimization problems where the objective function is known to be differentiable only on the relative interior of its domain (without any separability assumptions) was recently obtained in Schneider (1989, Thm. 7). Schneider's result could have been used to construct a more direct proof of (d). Still, we used independent arguments because they are more elementary.

LEMMA 2. *A vector* $y \in R^m$ *is an optimal solution of Program* II *if and only if*

$$y \leqq 0, \tag{10}$$

$$\sum_{j=1}^{n} C_{ij} a_j \exp\left(\sum_{k=1}^{m} y_k C_{kj}\right) \leqq b_i \quad \textit{for } i = 1, \ldots, m, \tag{11}$$

and

$$y_i \left[\sum_{j=1}^{n} C_{ij} a_j \exp\left(\sum_{k=1}^{m} y_k C_{kj}\right) - b_i\right] = 0 \quad \textit{for } i = 1, \ldots, m. \tag{12}$$

Proof. As the objective of Program II is convex and its constraints are defined via linear inequalities, we have that the Kuhn-Tucker conditions are necessary and sufficient for optimality; see Avriel (1976, Thms. 4.38 and 4.39, pp. 97-98). For Program II, the Kuhn-Tucker conditions assert that $y \leqq 0$ and for some vector $\mu \in R^m$,

$$\sum_{j=1}^{n} C_{ij} a_j \exp\left(\sum_{k=1}^{m} y_k C_{kj}\right) - b_i + \mu_i = 0 \quad \text{for } i = 1, \dots, m,$$

$$\mu_i \geqq 0 \quad \text{for } i = 1, \dots, n,$$

and

$$y_i \mu_i = 0 \quad \text{for } i = 1, \dots, m.$$

Trivially, these conditions are equivalent to (10)-(12). □

We next use the above two lemmas to show that solving LISP is equivalent to solving either Program I or Program II.

THEOREM 1 (characterization). *Let* $s \in R^m$ *and* $a' \in R^n$. *Then the following are equivalent*:

(a) *the pair* (a', s) *satisfies* (1')-(4');

(b) a' *is an optimal solution of Program* I, $a' \gg 0$, *and the pair* $(a', -s)$ *satisfies* (6)-(9); *and*

(c) s *is an optimal solution of Program* II *and*

$$a_j' = a_j \left[\exp\left(\sum_{i=1}^{m} s_i C_{ij}\right)\right] \quad \textit{for } j = 1, \dots, n. \tag{13}$$

Proof. (a)⟹(b): Assume that the pair (a', s) satisfies (1')-(4'). Then (2') implies that $a' \gg 0$ and that

$$\ln(a_j'/a_j) = (s^T C)_j \quad \text{for } j = 1, \dots, n.$$

Combining this fact with (1'), (3'), and (4') shows that the pair $(a', -s)$ satisfies (6)-(9). Further, Lemma 1 (d) assures that in this case a' is an optimal solution of Program I.

(b)⟹(a): Observing that (6) asserts that

$$x_j = a_j \exp\left(-\sum_{i=1}^{m} \lambda_i C_{ij}\right) \quad \text{for } j = 1, \dots, n,$$

we have that the pair (a', s) satisfies (1')-(4') if and only if $(a', -s)$ satisfies (6)-(9). In particular, we have that (b) implies (a).

(a)⟹(c): Assume that the pair (a', s) satisfies (1')-(4'). Then (2') coincides with (13) and, on substituting (2') into (3') and (4'), we have that

$$\sum_{j=1}^{n} C_{ij} a_j \exp\left(\sum_{k=1}^{m} s_k C_{kj}\right) \leqq b_i \quad \text{for } i = 1, \dots, m$$

and

$$s_i \left[\sum_{j=1}^{n} C_{ij} a_j \exp\left(\sum_{k=1}^{m} s_k C_{kj}\right) - b_i\right] = 0 \quad \text{for } i = 1, \dots, m.$$

As (1') asserts that $s \leqq 0$, we have that s satisfies (10). It now follows from Lemma 2 that s is an optimal solution of Program II. So, (c) has been established.

(c)⟹(a): Assume that s is an optimal solution for Program II and that (13) holds. Then Lemma 2 implies that s satisfies (10)-(12). In particular, on substituting the representation of a' given by (13), we get that

$$s \leqq 0, \qquad Ca' \leqq b,$$

and

$$y_i(Ca'-b)_i=0 \quad \text{for } i=1,\dots,m.$$

These conditions combine with (13) to show that the pair (a', s) satisfies (1′)-(4′). □

In order to characterize solutions to the original system (1)-(4) we apply a change of variable $z_i=\exp(y_i)$, $i=1,\dots,m$, to Program II, yielding the following program:

$$\text{Program II}': \qquad \min \sum_{j=1}^{n} a_j\left(\prod_{i=1}^{m} z_i^{C_{ij}}\right)-\sum_{i=1}^{m}\ln(z_i)b_i$$

$$\text{subject to} \quad 0\ll z\leqq 1.$$

Corollary 1. *Let $u\in R^m$ and $a'\in R^n$. Then the following are equivalent*:

(a) *the pair (a', u) satisfies* (1)-(4);

(b) *a' is an optimal solution of Program* II′, *$u\gg 0$, and the pair $(a', -s)$ satisfies* (6)-(9) *where $s\in R^m$ is defined by $s_i=\ln(u_i)$ for $i=1,\dots,m$; and*

(c) *u is an optimal solution of Program* II′ *and*

$$(14) \qquad a_j'=a_j\left(\prod_{i=1}^{m} u_i^{C_{ij}}\right) \quad \textit{for } j=1,\dots,n.$$

Proof. The equivalence of the above three assertions is immediate from Theorem 1 and the change of variables used to convert (1)-(4) into (1′)-(4′) and Program II into Program II′. □

3. Existence and uniqueness. Our next result gives a number of characterizations for the existence of a solution to LISP with (the given) data a, b, and C.

Theorem 2 (existence). *The following are equivalent*:

(a) LISP *with data a, b, and C has a solution*;

(b) $\{x\in R^n : Cx\leqq b, x\gg 0\}\neq\emptyset$;

(c) *there exists no $\lambda\in R^m$ that satisfies $\lambda\geqq 0$, $\lambda^T b\leqq 0$, and either $\lambda^T C\neq 0$ or $\lambda^T b\neq 0$*;

(d) *Program* I *has an optimal solution u that is strictly positive*;

(e) *Program* II *has an optimal solution*; *and*

(e′) *Program* II′ *has an optimal solution.*

Proof. (a)⇒(b): Assume that a' and u satisfy (1)-(4). Then the strict positivity of a and u implies that a' is strictly positive (see (2)). Combining this fact with (3) shows that $a'\in\{x\in R^n : Cx\leqq b, x\gg 0\}$; hence this latter set is nonempty, establishing (b).

(b)⇔(c): It follows from Motzkin's Theorem, e.g., Schrijver (1986, Cor. 7.1k, p. 94) that $\{x\in R^n : Cx\leqq b, x\gg 0\}\neq\emptyset$ if and only if there exists no $\lambda\in R^m$ and $\mu\in R^n$ such that

$$\lambda^T C-\mu^T=0, \quad \lambda\geqq 0, \quad \mu\geqq 0, \quad \lambda^T b\leqq 0 \quad \text{and either} \quad \mu\neq 0 \quad \text{or} \quad \lambda^T b\neq 0,$$

i.e., there exists no $\lambda\in R^m$ satisfying the conditions spelled out in (c). So, indeed, (b) and (c) are equivalent.

(b)⇔(d): This equivalence is immediate from Lemma 1(c).

(c)⇒(e): Assume that (c) holds, i.e.,

$$(15) \qquad \lambda\in R^m, \quad \lambda\geqq 0, \quad \lambda^T C\geqq 0, \quad \lambda^T b\leqq 0 \quad \Rightarrow \quad \lambda^T C=0, \quad \lambda^T b=0$$

and we will show that Program II has an optimal solution. Denote the objective function of Program II by $h(\cdot)$, i.e., for $y \in R^m$, $h(y) = \sum_{j=1}^n a_j \exp(y^TC)_j - b^Ty$. A *direction of recession* of $h(\cdot)$ is defined to be a vector $d \in R$ for which

$$\sup_{y \in R^m} \{h(y+d) - h(y)\} \leqq 0. \tag{16}$$

As the function $h(\cdot)$ is convex, we have from Rockafellar (1970, Thm. 173, p. 267) that existence of an optimal solution for Program II is implied by the assertion that every direction of recession $d \leqq 0$ of $h(\cdot)$ satisfies

$$h(y+d) - h(y) = 0 \quad \text{for all } y \in R^m. \tag{17}$$

Thus, it suffices to show that if $d \leqq 0$ satisfies (16), then it must satisfy (17).

Assume that $d \leqq 0$ satisfies (16). Substituting the explicit expression of $h(\cdot)$ into (16), we have that for every $y \in R^m$,

$$\sum_{j=1}^n a_j \exp[(y^TC)_j + (d^TC)_j] - \sum_{j=1}^n a_j \exp(y^TC)_j - b^Td \leqq 0, \tag{18}$$

or equivalently,

$$\sum_{j=1}^n a_j\{\exp(y^TC)_j[\exp(d^TC)_j - 1]\} \leqq b^Td. \tag{19}$$

We next argue that $d^TC \leqq 0$. Let $J_+ \equiv \{j = 1, \dots, n : (d^TC)_j > 0\}$ and let $J_- \equiv \{j = 1, \dots, n : (d^TC)_j \leqq 0\}$ and we will show that $J_+ = \emptyset$. Suppose that $J_+ \neq \emptyset$ and $p \in J_+$. Then for each $M > 0$, (19) with $y = Md$ implies that

$$\begin{aligned} a_p \exp[M(d^TC)_p][\exp(d^TC)_p - 1] &\leqq \sum_{j \in J_+} a_j \exp[M(d^TC)_j][\exp(d^TC)_j - 1] \\ &\leqq b^Td - \sum_{j \in J_-} a_j \exp[M(d^TC)_j][\exp(d^TC)_j - 1] \\ &\leqq b^Td + \sum_{j \in J_-} a_j, \end{aligned}$$

implying that $a_p \exp[M(d^TC)_p][\exp[M(d^TC)_p - 1]$ is bounded from above in M. But this conclusion is false as $(d^TC)_p > 0$. This contradiction proves that $J_+ = \emptyset$, i.e., $d^TC \leqq 0$. We also have from (19), again with $y = Md$, that for each $M > 0$

$$b^Td \geqq \sum_{j=1}^n a_j\{\exp[M(d^TC)_j][\exp(d^TC)_j - 1]\}.$$

As $d^TC \leqq 0$, the right-hand side of the above inequality converges to zero as $M \to \infty$ and we conclude that $b^Td \geqq 0$. So, we have seen that (16) implies that $d^TC \leqq 0$ and $d^Tb \geqq 0$. We next combine this fact with (15) to conclude that if $d \leqq 0$ satisfies (16), then necessarily $d^TC = 0$ and $d^Tb = 0$, implying that (19) and (18) must hold as equalities. So, (17) must hold, thereby completing our proof that (c)$\Rightarrow$(e).

(e)$\Leftrightarrow$(e'): This equivalence is immediate as Program II and Program II' are derived from each other via a change of variables, $z_i = \exp(y_i)$ and $y_i = \ln(z_i)$, $i = 1, \dots, m$, respectively.

(e')$\Rightarrow$(a): Assume that (e') holds and that $u \in R^m$ is an optimal solution of Program II. Given u, define $a' \in R^n$ by

$$a'_j = a_j\left(\prod_{i=1}^m u_i^{C_{ij}}\right) \quad \text{for } j = 1, \dots, n.$$

Then Corollary 1 assures that (a', u) satisfies (1)-(4), implying that LISP with data a, b, and C has a solution. □

Remark. We observe that conditions (b) and (c) of Theorem 2 are independent of the vector a. Thus, Theorem 2 implies that if $b \in R^m$ and $C \in R^{m \times n}$ are fixed and either condition (a), (d), (e), or (e′) is satisfied for a, b, and C for one (strictly positive) vector a in R^n, then the same conclusion holds for all (strictly positive) vectors a in R^n.

Theorem 2 demonstrates that LISP need not have a solution. We next examine relaxations of LISP, which can sometimes be solved in cases where LISP itself cannot. In particular, we consider two types of relaxations of LISP. One concerns the relaxation of the constraints $Cx \leqq b$ through perturbations $Cx \leqq b + \varepsilon 1$ for arbitrarily small positive ε. The other concerns the replacement of some of the positive elements of the data vector a by zero elements. Interestingly, solvability of these two different relaxations is equivalent.

We need one additional piece of notation. For a subset J of $\{1, \ldots, n\}$, we let C^J denote the submatrix of C corresponding to the columns indexed by J. Similarly, for a vector $x \in R^n$, we let x_J be the corresponding subvector of x.

THEOREM 3 (existence for perturbed relaxations). *The following are equivalent*:

(a1) *for every* $\varepsilon > 0$, LISP *with data* a, $b + \varepsilon 1$, *and* C *has a solution*;

(a2) *for some subset* J *of* $\{1, \ldots, n\}$, LISP *with data* a_J, b, *and* C^J *has a solution*;

(b1) *for every* $\varepsilon > 0$, $\{x \in R^n : Cx \leqq b + \varepsilon 1,\ x \gg 0\} \neq \emptyset$;

(b2) $\{x \in R^n : Cx \leqq b,\ x \geqq 0\} \neq \emptyset$;

(c) *there is no vector* λ *in* R^m *for which* $\lambda \geqq 0$, $\lambda^T C \geqq 0$, *and* $\lambda^T b < 0$;

(d) *Program* I *is feasible and has an optimal solution*;

(e) *the objective of Program* II *is bounded from below*; *and*

(e′) *the objective of Program* II′ *is bounded from below.*

Proof. The equivalences (a1)⇔(b1) and (a2)⇔(b2) follow from Theorem 2 (the latter relying on the convention that $C^J x_J$ is defined as zero when J is empty); the equivalence (b2)⇔(c) follows from a variant of Farkas's Lemma, e.g., Schrijver (1986, Cor. 7.1f, p. 90); the equivalence (b2)⇔(d) follows from Lemma 1; and the equivalence (e)⇔(e′) follows from the fact that Programs II and II′ are derivable from each other via the change of variable $z_i = \exp(y_i)$ for $i = 1, \ldots, m$. We complete the proof by showing that (b1)⇔(b2) and (b2)⇒(e)⇒(c).

(b1)⇒(b2): Assume that (b1) holds. Then for every $\varepsilon > 0$,

$$\{x \in R^n\colon Cx \ll b + 1\varepsilon,\ x \gg 0\} \supseteq \{x \in R^n\colon Cx \leqq b + 2^{-1}\varepsilon 1,\ x \gg 0\} \neq \emptyset,$$

implying that

$$\begin{aligned} &\{x \in R^n, s \in R^m\colon \|Cx + Is - b\|_\infty \leqq \varepsilon,\ x \gg 0,\ s \gg 0\} \\ &\quad \supseteq \{x \in R^n, s \in R^m\colon Cx + Is = b + 2^{-1}\varepsilon 1,\ x \gg 0,\ s \gg 0\} \neq \emptyset. \end{aligned}$$

It now follows from standard results about linear inequalities, e.g., the Appendix of Rothblum and Schneider (1989), that $\{x \in R^n,\ s \in R^m\colon Cx + Is = b,\ x \geqq 0,\ s \geqq 0\} \neq \emptyset$, i.e., $\{x \in R^n\colon Cx \leqq b,\ x \geqq 0\} \neq \emptyset$.

(b2)⇒(b1): Assume that (b2) holds, i.e., there exists a vector $x^* \in R^m$ satisfying $Cx^* \leqq b$ and $x^* \geqq 0$. Then for every $\varepsilon > 0$, $x_\varepsilon \equiv x^* + \varepsilon(\|C1\|_\infty + 1)^{-1} 1 \in \{x \in R^n\colon Cx \leqq b + \varepsilon 1,\ x \gg 0\}$.

(b2)⇒(e): Assume that (b2) holds, i.e., there exists a vector $x^* \in R^n$ satisfying $Cx^* \leqq b$ and $x^* \geqq 0$. Let $J \equiv \{i = 1, \ldots, n\colon x_i^* > 0\}$. Then $\{x_J \in R^{|J|}\colon C^J x_J \leqq b,\ x_J \gg 0\} \neq \emptyset$ and the equivalence of (b) and (e) in Theorem 2 implies that the minimization problem

$$\begin{aligned} &\min \sum_{j \in J} a_j \exp\left(\sum_{i=1}^m y_i C_{ij}\right) - \sum_{i=1}^m y_i b_i \\ &\text{subject to} \quad y \leqq 0 \end{aligned}$$

has a minimum. As each term $a_j \exp(y^TC)_j$ for $j \in \{1, \ldots, m\} \backslash J$ is positive, the minimal objective value of the above problem is a lower bound on the objective value of Program II.

(e)$\Rightarrow$(c): Suppose (c) is not satisfied and there exists a vector $\lambda \in R^m$ with $\lambda \geqq 0$, $\lambda^T C \geqq 0$, and $\lambda^T b < 0$. Then for every $M > 0$, $-M\lambda$ is feasible for Program II and

$$\sum_{j=1}^{n} a_j \exp[(-M\lambda)^T C]_j - (-M\lambda)^T b \leqq \sum_{j=1}^{n} a_j + M\lambda^T b.$$

As $\lim_{M\to\infty} M\lambda^T b = -\infty$, we conclude that the objective function of Program II is unbounded from below, i.e., (e) is not satisfied. □

Remarks. 1. We note that condition (c) of either Theorem 2 or Theorem 3 can be replaced with a finite set of constraints on C and b; see Eaves and Rothblum (1992) for details. Particularly useful conditions that can be derived explicitly arise in problems where the matrix C is an incidence matrix of a graph. This happens, for example, in matrix scaling problems; see further discussion in §§ 4 and 5.

2. The proof of the equivalence of (a2) and (b2) in Theorem 3 (using Theorem 2) shows that (a2) is satisfied for a specific subset J of $\{1, \ldots, n\}$ if and only if $\{x \in R^n: C^J x_J \leqq b, x_J \gg 0\} \neq \emptyset$. As there exists a maximal subset J of $\{1, \ldots, n\}$ for which $\{x \in R^n: C^J x_J \leqq b, x_J \gg 0\} \neq \emptyset$, we have that this set is the maximal set J for which LISP with data a_J, b, and C^J has a solution.

Our next result concerns uniqueness of solutions to LISP.

THEOREM 4 (uniqueness). *There exists at most one vector a' for which there exists a vector $u \in R^m$ such that (a', u) is a solution to* LISP *with data a, b, and C. Furthermore, if (a', u) is a solution to* LISP *with data a, b, and C, then $(a', \bar{u})$ is another solution if and only if*

$$0 \ll \bar{u} \leqq 1, \tag{20}$$

$$\sum_{i=1}^{n} \ln(u_i) C_{ij} = \sum_{i=1}^{n} \ln(\bar{u}_i) C_{ij} \quad \textit{for } j = 1, \ldots, n, \tag{21}$$

and

$$\bar{u}_i = 1 \quad \textit{whenever } (Ca')_i < b_i \quad \textit{for } i = 1, \ldots, m. \tag{22}$$

Proof. Assume that (a', u) and $(\bar{a}', \bar{u})$ are solutions of (1)-(4). Then Corollary 1 implies that both a' and $\bar{a}'$ are optimal solutions of Program I, and therefore by the uniqueness of a solution to that program (see Lemma 1), we have that $a' = \bar{a}'$. Furthermore, we get from Corollary 1 that if (a', u) and $(a', \bar{u})$ satisfy (1)-(4), then

$$a'_j = a_j \left(\prod_{i=1}^{m} u_i^{C_{ij}} \right) = a_j \left(\prod_{i=1}^{m} \bar{u}_i^{C_{ij}} \right) \quad \text{for } j = 1, \ldots, n,$$

immediately implying (21). Also, (20) and (22) follow directly from the fact that $(a', \bar{u})$ satisfies (1) and (4), respectively.

Next, assume that (a', u) is a solution of (1)-(4) and that $\bar{u}$ satisfies (20)-(22). First, (20) and (22) show that $(a', \bar{u})$ satisfies (1) and (4), and as (a', u) satisfies (3), so does $(a', \bar{u})$. Finally, on exponentiation of (21), we get that

$$\prod_{k=1}^{m} u_k^{C_{kj}} = \prod_{k=1}^{m} \bar{u}_k^{C_{kj}} \quad \text{for } j = 1, \ldots, n,$$

implying that

$$a'_j = a_j\left(\prod_{k=1}^{m} u_k^{C_{kj}}\right) = a_j\left(\prod_{k=1}^{m} \bar{u}_k^{C_{kj}}\right) \quad \text{for } j = 1, \dots, n,$$

thereby showing that $(a', \bar{u})$ satisfies (2). □

4. Linear equality scaling problems. We continue to let $a \in R^n$, $b \in R^n$, and $C \in R^{m \times n}$ be given where a is strictly positive. Recall that the linear equality scaling problem (LESP) with data a, b, and C is defined as the problem of finding a vector $u \in R^n$ and $a' \in R^N$ satisfying

$$0 \ll u, \tag{23}$$

$$a'_j = a_j\left(\prod_{i=1}^{m} u_i^{C_{ij}}\right) \quad \text{for } j = 1, \dots, n, \tag{24}$$

and

$$Ca' = b. \tag{25}$$

Of course, the change of variables $s_i = \ln(u_i)$ makes (23) superfluous and converts (24) to

$$a'_j = a_j \exp\left(\sum_{i=1}^{m} s_i C_{ij}\right) \quad \text{for } j = 1, \dots, n. \tag{24'}$$

We have seen in § 1 that the LESP with data a, b, and C is equivalent to the LISP with data a, $\binom{b}{-b}$, and $\binom{C}{-C}$. Programs I and II can be written for this instance of the LISP as

$$\text{Program I*:} \quad \min \sum_{j=1}^{n} x_j[\ln(x_j/a_j) - 1]$$

$$\text{subject to} \quad Cx = b, \qquad x \geqq 0,$$

and

$$\text{Program II*:} \quad \min \sum_{j=1}^{n} a_j \exp\left(\sum_{i=1}^{m} y_i C_{ij}\right) - \sum_{i=1}^{m} y_i b_i,$$

where a standard change of variable was applied in the derivation of Program II*. Programs I* and II* were used, respectively, in Franklin and Lorenz (1989) and Rothblum (1989) to analyze the LESP. Earlier work of Marshall and Olkin (1968), Bacharach (1970), Bachem and Korte (1979), Eaves et al. (1985) and Rothblum and Schneider (1989) and others considered special cases of Programs I* and II* for instances of LESP that concern matrix scalings. We next specialize the results of §§ 2 and 3 to obtain many of the results of the above references. The arguments require some standard transformations that are left to the reader.

THEOREM 1* (characterization). *Let $s \in R^m$ and $a' \in R^n$. Then the following are equivalent*:

(a) *the pair (a', s) satisfies* (24′) *and* (25);

(b) *a' is an optimal solution of Program* I*, *$a' \gg 0$, and the pair $(a', -s)$ satisfies* (26), (27); *and*

(c) *s is an optimal solution of Program* II* *and* (24′) *holds.*

THEOREM 2* (existence). *The following are equivalent:*

(a) LESP *with data* a, b, *and* C *has a solution*;

(b) $\{x \in R^n\colon Cx = b,\ x \gg 0\} \neq \varnothing$;

(c) *there exists no* $\eta \in R^m$ *which satisfies* $\lambda^T C \geqq 0$, $\lambda^T b \leqq 0$ *and either* $\lambda^T C \neq 0$ *or* $\lambda^T b \neq 0$;

(d) *Program* I* *has an optimal solution* u *which is strictly positive*; *and*

(e) *Program* II* *has an optimal solution.*

The remark following Theorem 2 shows that, given $b \in R^m$ and $C \in R^{m \times n}$, if either condition (a), (d), or (e) of Theorem 2* holds for a, b, and C for some (strictly positive) vector a, then the same conclusion holds for all (strictly positive) vectors a.

THEOREM 3* (existence for perturbed relaxations). *The following are equivalent:*

(a) *for every* $\varepsilon > 0$, *there exist vectors* $s \in R^m$ *and* $a' \in R^n$ *satisfying* (24') *and* $\|Ca' - b\|_\infty \leqq \varepsilon$;

(b1) *for every* $\varepsilon > 0$, $\{x \in R^n\colon \|Cx - b\|_\infty \leqq \varepsilon,\ x \gg 0\} \neq \varnothing$;

(b2) $\{x \in R^n\colon Cx = b,\ x \geqq 0\} \neq \varnothing$;

(c) *there exists no* $\lambda \in R^m$ *for which* $\lambda^T C \geqq 0$ *and* $\lambda^T b < 0$;

(d) *Program* I* *is feasible and has an optimal solution*; *and*

(e) *the objective of Program* II* *is bounded from below.*

We observed in § 3 that condition (c) of either Theorem 2 or 3 can be replaced with a finite set of constraints on C and b. A particularly useful form of such (finite) constraints is available for matrix scalings (see § 5), e.g., the Menon-Schneider conditions for matrix scalings with prespecified row-sums and column-sums. See Rothblum and Schneider (1989) for further details.

THEOREM 4* (uniqueness). *There exists at most one vector* a' *for which there exists a vector* $u \in R^m$ *such that* (a', u) *is a solution to* LESP *with data* a, b, *and* C. *Further, if* (a', u) *is such a solution, then* $(a', \bar{u})$ *is another solution if and only if* $\bar{u} \gg 0$ *and*

$$\sum_{i=1}^m \ln(u_i) C_{ij} = \sum_{i=1}^m \ln(\bar{u}_i) C_{ij} \quad \textit{for } j = 1, \ldots, n.$$

We can combine the LISP and the LESP and consider scaling problems with both equalities and inequalities. The results of §§ 2 and 3 and those of this section can then be combined correspondingly. We omit the details for the general case because they are straightforward (though a little cumbersome). Still, we demonstrate the idea for LESP with upper and lower bounds.

Suppose that in addition to the given a, b, and C we have two nonnegative vectors r and t in R^n. Consider the LISP with data

$$a, \quad \begin{pmatrix} b \\ -b \\ r \\ -t \end{pmatrix}, \quad \text{and} \quad \begin{pmatrix} C \\ -C \\ I \\ -I \end{pmatrix}.$$

Using standard transformations, this problem reduces to identifying $a' \in R^n$, $u \in R^m$, and $\eta \in R^n$ such that

$$0 \ll u, \qquad 0 \ll \eta,$$

$$a'_j = a_j \left(\prod_{k=1}^m u_k^{C_{kj}} \right) \eta_j \quad \text{for } j = 1, \ldots, n,$$

$$Ca' = b \quad \text{and} \quad t \leqq a' \leqq r,$$

$$a'_j = r_j \quad \text{whenever } \eta_j < 1,$$

and

$$a'_j = t_j \quad \text{whenever } \eta_j > 1.$$

Furthermore, Programs I and II for the above LISPs are then given, respectively, by

$$\textit{Program } \mathrm{I}\dagger\text{:} \qquad \min \sum_{j=1}^{n} x_j[\ln(x_j/a_j)-1]$$
$$\text{subject to} \quad Cx = b, \quad t \leqq x \leqq r$$

and

$$\textit{Program } \mathrm{II}\dagger\text{:} \qquad \min \sum_{j=1}^{n} a_j \exp\left(\sum_{i=1}^{m} z_i C_{ij} + y_i - x_i\right) - b^T z - r^T y + s^T x$$
$$\text{subject to} \quad x \leqq 0, \quad y \leqq 0.$$

Lemmas 1 and 2 and Theorems 1-4 can now be modified correspondingly; we omit the details.

5. Matrix scaling. Many important applications concern scaling problems where the vector a that is to be multiplied by scaling coefficients is actually a multidimensional array of numbers. For example, when a represents a matrix, we refer to problems that concern corresponding scalings as *matrix scaling problems.*

One type of commonly used scaling of a matrix $A = (A_{ij})$ has the form $(D_i A_{ij} E_j)$, where D and E are diagonal matrices having positive diagonal elements. We note that such scalings are C-scalings of the array (A_{ij}) where the matrix C is the bipartite node-node incidence matrix of the (un)directed graph that corresponds to the matrix A; see Rothblum (1989) for details. We observe that in this case, with a' corresponding to the scaled matrix, we have that Ca' represents the vector whose elements are the row-sums and column-sums of the scaled matrix. So, the constraints $Ca' = b$ or $Ca' \leqq b$ represent, respectively, prespecification or upper bounds on row- and column-sums; see Rothblum (1989) for a more explicit explanation of the equality case. Schneider (1989) considered the equality case with upper and lower bounds and showed how this problem generalized a number of matrix scaling problems with equality constraints. Balinski and Demange (1987, 1988) considered the problem where upper and lower bounds on the row- and column-sums are given.

Another type of useful scaling of a square matrix $A = (A_{ij})$ has the form $(D_i A_{ij} D_j^{-1})$, where D is a diagonal matrix having positive diagonal elements. Such scalings are C-scalings of the array (A_{ij}) where the matrix C is the node-node incidence matrix of the directed graph that corresponds to the matrix A; see Rothblum (1989) for details. We observe that in this case, with a' corresponding to the scaled matrix, we have that Ca' represents the vector whose elements are the differences between row-sums and corresponding column-sums of the scaled matrix. Further, the constraints $Ca' = b$ or $Ca' < b$ represent, respectively, prespecification or upper bounds on these differences; see Rothblum (1989) for a more explicit explanation of the equality case.

REFERENCES

M. AVRIEL (1976), *Nonlinear Programming—Analysis and Methods*, Prentice-Hall, Englewood Cliffs, NJ.

M. BACHARACH (1970), *Biproportional Matrices and Input-Output Change*, Cambridge University Press, Cambridge, U.K.

A. BACHEM AND B. KORTE (1979), *On the* RAS *algorithm*, Computing, 23, pp. 189-198.

M. L. BALINSKI AND G. DEMANGE (1988), *Algorithms for proportional matrices in reals and integers*, manuscript.

M. L. BALINSKI AND G. DEMANGE (1987), *An axiomatic approach to proportionality between matrices,* manuscript.

R. BAPAT (1982), D_1AD_2 *theorems for multidimensional matrices,* Linear Algebra Appl., 48, pp. 437–442.

R. BAPAT AND T. E. S. RAGHAVAN (1989), *An extension of a theorem of Darroch and Ratcliff in loglinear models and its application to scaling multidimensional matrices,* Linear Algebra Appl., 114/115, pp. 705-715.

B. C. EAVES, A. HOFFMAN, U. G. ROTHBLUM, AND H. SCHNEIDER (1985), *Line-sum-symmetric scalings of square nonnegative matrices,* Math. Programming Stud., 15, pp. 124–141.

B. C. EAVES AND U. G. ROTHBLUM (1992), *Elimination of quantifiers of linear variables and corresponding transfer principles,* Math. Programming, 53, pp. 307–321.

J. FRANKLIN AND J. LORENZ (1989), *On scaling of multidimensional matrices,* Linear Algebra Appl., 114/115, pp. 715–735.

B. KING (1981), *What is* SAM? *A layman's guide to social accounting matrices,* World Bank Staff Working Paper No. 463, The World Bank, Washington, D.C.

A. W. MARSHALL AND I. OLKIN (1968), *Scaling of matrices to achieve specified row- and column-sums,* Numer. Math., 12, pp. 83-90.

T. E. S. RAGHAVAN (1984), *On pairs of multidimensional matrices,* Linear Algebra Appl., 62, pp. 263–268.

——— (1985), *On pairs of multidimensional matrices and their applications,* Contemp. Math., 46, pp. 339–354.

R. T. ROCKAFELLAR (1970), *Convex Analysis,* Princeton University Press, Princeton, NJ.

U. G. ROTHBLUM (1989), *Generalized scalings satisfying linear equations,* Linear Algebra Appl., 114/115, pp. 765–784.

———(1989a), *Linear inequality scaling problems,* RUTCOR Research Report 40-89, RUTCOR, Rutgers Univ., New Brunswick, NJ.

U. G. ROTHBLUM AND H. SCHNEIDER (1989), *Scalings of matrices having pre-specified row-sums and column-sums,* Linear Algebra Appl., 114/115, pp. 737-764.

M. H. SCHNEIDER (1989), *Matrix scaling, entropy minimization and conjugate duality* I: *Existence conditions,* Linear Algebra Appl., 114/115, pp. 785-813.

———(1989a), *Matrix scaling, entropy minimization and conjugate duality* II: *The dual problem,* manuscript.

M. H. SCHNEIDER AND S. ZENIOS (1987), *A comparison study of algorithms for matrix balancing,* OR Group Report Series 88-02, Dept. of Mathematical Sciences, The Johns Hopkins Univ., Baltimore, MD.

A. SCHRIJVER (1986), *Theory of Linear and Integer Programming,* John Wiley, New York.

SIAM J. OPTIMIZATION
Vol. 2, No. 4, pp. 649-664, November 1992

007

NEW PROXIMAL POINT ALGORITHMS FOR CONVEX MINIMIZATION*

OSMAN GÜLER†

Abstract. This paper introduces two new proximal point algorithms for minimizing a proper, lower-semicontinuous convex function $f:\mathbf{R}^n \to R \cup \{\infty\}$. Under this minimal assumption on f, the first algorithm possesses the global convergence rate estimate $f(x_k) - \min_{x\in\mathbf{R}^n} f(x) = O(1/(\sum_{j=0}^{k-1}\sqrt{\lambda_j})^2)$, where $\{\lambda_k\}_{k=0}^{\infty}$ are the proximal parameters. It is shown that this algorithm converges, and global convergence rate estimates for it are provided, even if minimizations are performed inexactly at each iteration. Both algorithms converge even if f has no minimizers or is unbounded from below. These algorithms and results are valid in infinite-dimensional Hilbert spaces.

Key words. proximal point algorithms, global convergence rates, augmented Lagrangian algorithms, convex programming

AMS(MOS) subject classifications. primary 90C25; secondary 49D45, 49D37

1. Introduction.

In this paper we present two new proximal point algorithms for the minimization problem

$$\min_{x\in\mathbf{R}^n} f(x), \tag{1.1}$$

where $f:\mathbf{R}^n \to R\cup\{\infty\}$ is a proper lower-semicontinuous convex function, using the terminology established in Aubin and Ekeland [1] and Rockafellar [17].

The classical proximal point algorithm was introduced into optimization literature by Martinet [11]. It is based on the notion of proximal mapping J_λ,

$$J_\lambda x := \arg\min_{z\in\mathbf{R}^n}\left\{f(z) + \frac{1}{2\lambda}\|z-x\|^2\right\}, \tag{1.2}$$

introduced earlier by Moreau [12]. The proximal point algorithm solves a single optimization problem by solving a sequence of optimization problems (1.2): it starts from a point $x_0\in\mathbf{R}^n$ and generates the sequence $\{x_k\}_{k=0}^{\infty}$, where

$$x_{k+1} = J_{\lambda_k}x_k := \arg\min_{x\in\mathbf{R}^n}\left\{f(x) + \frac{1}{2\lambda_k}\|x-x_k\|^2\right\}, \tag{1.3}$$

and where $\{\lambda_k\}_{k=0}^{\infty}$ is a sequence of positive numbers. The proximal point algorithm was popularized by Rockafellar [18], who showed that the algorithm converges even if the auxiliary minimizations in (1.3) are performed inexactly, which is an important consideration in practice. Güler [7] analyzed the algorithm further and provided global convergence rate estimates for it in terms of the objective residual $f(x_k) - \min_{x\in\mathbf{R}^n} f(x)$.

The minimization problem (1.1) is general enough to include the generic convex programming problem

$$\min_{x\in C} f_0(x) \quad \text{s.t. } f_i(x) \leqq 0, \quad i = 1, \dots, m, \tag{1.4}$$

where C is a closed convex subset of $\mathbf{R}^n$ and $f_i:\mathbf{R}^n\to\mathbf{R}$, $i = 0, 1, \dots, m$ are convex functions; see Rockafellar [19].

* Received by the editors December 26, 1990; accepted for publication (in revised form) August 7, 1991.

† Department of Management Sciences, The University of Iowa, Iowa City, Iowa 52242.

The usual application of the proximal point algorithm to convex programming is not to the primal program (1.4), but to its dual

$$(1.5) \qquad \max_{y\in\mathbf{R}^m}\left\{\inf_{x\in C} f_0(x)+\sum_{i=1}^m y_if_i(x)\right\} \quad \text{s.t. } y\geqq 0.$$

The resulting algorithm is called the augmented Lagrangian method. It was introduced into optimization literature independently by Hestenes [9] and Powell [16]. Augmented Lagrangian methods have many advantages over penalty methods; see Bertsekas [2] and Rockafellar [19].

The algorithms developed here are close in spirit to the classical proximal point algorithm discussed above. The only difference is that our algorithms generate an additional sequence $\{y_k\}_{k=0}^{\infty}$ of points in $\mathbf{R}^n$, and calculate x_{k+1} from

$$(1.6) \qquad x_{k+1}=J_{\lambda_k}y_k:=\arg\min_{x\in\mathbf{R}^n}\left\{f(x)+\frac{1}{2\lambda_k}\|x-y_k\|^2\right\}.$$

The main work in the new algorithm is in the calculation of x_{k+1} in (1.6), the calculation of y_k being trivial. As with the classical algorithm, we show that the minimization in (1.6) can be performed inexactly.

For any feasible $x\in\mathbf{R}^n$, the algorithms here possess the global convergence rate estimate

$$(1.7) \qquad f(x_k)-f(x)=O\left(\frac{1}{(\sum_{j=0}^{k-1}\sqrt{\lambda_j})^2}\right).$$

This is faster than the available rate

$$(1.8) \qquad f(x_k)-f(x)=O\left(\frac{1}{\sum_{j=0}^{k-1}\lambda_j}\right)$$

obtained by Güler [7] for the classical proximal point algorithm.

The paper is organized as follows. In § 2 we present the first proximal point algorithm for (1.1). We state the algorithm and estimate its convergence rate under the assumption that exact minimizations are performed in (1.6). In § 3, we present a version of the algorithm that requires only inexact minimizations in (1.6). In § 4, we show that the algorithms in §§ 2 and 3 have the property that $f(x_k)\to\inf_{x\in\mathbf{R}^n}f(x)$, even in cases where f has no minimizers or is unbounded from below. We also present a monotonic version of the algorithm in which $f(x_{k+1})\leqq f(x_k)$. Some concluding remarks are made in § 5. In the Appendix, we present our second proximal point algorithm.

2. The proximal point algorithm. In this section we develop a new proximal point algorithm for problem (1.1). The inspiration for the algorithm comes from a paper by Nesterov [14] in which an optimal algorithm is developed for smooth convex minimization.

The idea of the algorithm is to generate recursively a sequence $\{\varphi_k\}_{k=0}^{\infty}$ of simple convex quadratic functions (with a diagonal matrix in the quadratic term) that approximate $f(x)$ in such a way that at step $k\geqq 0$, the difference $\varphi_k(x)-f(x)$ is reduced by a fraction $1-\alpha_k$, that is, for all $x\in\mathbf{R}^n$,

$$(2.1) \qquad \varphi_{k+1}(x)-f(x)\leqq(1-\alpha_k)(\varphi_k(x)-f(x)),$$

where α_k is a number in the interval $[0, 1)$.

If (2.1) is satisfied for each $k \geqq 0$, we obtain by induction

$$\varphi_k(x) - f(x) \leqq \left(\prod_{j=0}^{k-1} (1-\alpha_j) \right) (\varphi_0(x) - f(x)).$$

Defining

$$\beta_k = \prod_{j=0}^{k-1} (1-\alpha_j), \tag{2.2}$$

we have

$$\varphi_k(x) - f(x) \leqq \beta_k(\varphi_0(x) - f(x)). \tag{2.3}$$

If, at step k, we have at hand a point x_k such that

$$f(x_k) \leqq \varphi_k^* := \min_{z \in \mathbf{R}^n} \varphi_k(z), \tag{2.4}$$

then we obtain from (2.3) the global convergence estimate

$$f(x_k) - f(x) \leqq \beta_k(\varphi_0(x) - f(x)). \tag{2.5}$$

This is a significant bound only if $f(x) < \infty$, that is, if x is feasible. If f has a minimizer x^*, (2.5) specializes to

$$f(x_k) - f^* \leqq \beta_k(\varphi_0(x^*) - f^*). \tag{2.6}$$

If $\beta_k \to 0$, then $\{x_k\}$ is a minimizing sequence for f. The magnitude of the constant β_k is a measure of the convergence rate of $f(x_k)$ to f^*.

We define the quadratic functions $\varphi_k(x)$, $k \geqq 0$, recursively, as follows:

$$\begin{aligned} \varphi_0(x) &:= f(x_0) + \frac{A}{2} \|x - x_0\|^2, \\ \varphi_{k+1}(x) &:= (1-\alpha_k)\varphi_k(x) \\ &\quad + \alpha_k(f(J_{\lambda_k} y_k) + \langle (y_k - J_{\lambda_k} y_k)/\lambda_k, x - J_{\lambda_k} y_k \rangle). \end{aligned} \tag{2.7}$$

Here A and λ_k are positive numbers and α_k is a number in the interval $[0, 1)$. The point x_0 is feasible, that is, $f(x_0) < \infty$. Here the point $y_k \in \mathbf{R}^n$ can be arbitrary. Later, it will be chosen to satisfy certain desirable properties.

LEMMA 2.1. *For all* $k \geqq 0$, *the quadratic functions* $\varphi_k(x)$ *defined above satisfy inequality* (2.1), *that is,*

$$\varphi_{k+1}(x) - f(x) \leqq (1-\alpha_k)(\varphi_k(x) - f(x)).$$

Proof. Since $J_{\lambda_k} y_k$ is the minimizer in (1.6), we have by the subdifferentiation formula (see Rockafellar [18, pp. 889]), $0 \in \partial f(J_{\lambda_k} y_k) + (J_{\lambda_k} y_k - y_k)/\lambda_k$, that is,

$$(y_k - J_{\lambda_k} y_k)/\lambda_k \in \partial f(J_{\lambda_k} y_k). \tag{2.8}$$

Since f is convex, for any $x \in \mathbf{R}^n$, we have

$$f(x) \geqq f(J_{\lambda_k} y_k) + \langle (y_k - J_{\lambda_k} y_k)/\lambda_k, x - J_{\lambda_k} y_k \rangle. \tag{2.9}$$

Thus

$$\begin{aligned} \varphi_{k+1}(x) - f(x) &= (1-\alpha_k)(\varphi_k(x) - f(x)) \\ &\quad + \alpha_k(f(J_{\lambda_k} y_k) + \langle (y_k - J_{\lambda_k} y_k)/\lambda_k, x - J_{\lambda_k} y_k \rangle - f(x)) \\ &\leqq (1-\alpha_k)(\varphi_k(x) - f(x)). \end{aligned}$$

□

It is not obvious a priori how points $x_k \in \mathbf{R}^n$ can be chosen to satisfy inequality (2.4). Toward this goal, we first note that the quadratic function $\varphi_k(x)$ can be written in the canonical form

$$\varphi_k(x) = \varphi_k^* + \frac{A_k}{2} \|x - \nu_k\|^2, \tag{2.10}$$

where φ_k^* is the minimum value of the function $\varphi_k(x)$ in $\mathbf{R}^n$ and ν_k is its minimizer. Clearly $A_0 = A$ and $\nu_0 = x_0$. Using (2.7) and (2.10) it is easy to show that for $k \geqq 0$,

$$A_{k+1} = (1 - \alpha_k) A_k = \beta_{k+1} A, \tag{2.11}$$

$$\nu_{k+1} = \nu_k - \frac{\alpha_k}{A_{k+1}\lambda_k} (y_k - J_{\lambda_k} y_k). \tag{2.12}$$

We will determine the points $\{x_k\}$ satisfying (2.4) recursively. Suppose we already have a point x_k satisfying inequality (2.4). The following result indicates how y_k and x_{k+1} can be chosen such that x_{k+1} also satisfies (2.4). It is the main result of this section and uses ideas from Nesterov [14, Lemma 1].

THEOREM 2.1. *If, for some $k \geqq 0$, x_k satisfies the inequality* (2.4), *that is*, $f(x_k) \leqq \varphi_k^*$, *then for any $y_k \in \mathbf{R}^n$, $\lambda_k > 0$, and $\alpha_k \in [0, 1)$, the following inequality holds*:

$$\begin{aligned} \varphi_{k+1}^* \geqq f(J_{\lambda_k} y_k) &+ \frac{1}{2\lambda_k}\left(2 - \frac{\alpha_k^2}{A_{k+1}\lambda_k}\right) \|y_k - J_{\lambda_k} y_k\|^2 \\ &+ \frac{1}{\lambda_k} \langle y_k - J_{\lambda_k} y_k, (1 - \alpha_k) x_k + \alpha_k \nu_k - y_k \rangle. \end{aligned} \tag{2.13}$$

Proof. We obtain from (2.7), (2.10), and (2.11)

$$\begin{aligned} \varphi_{k+1}^* &:= \varphi_{k+1}(\nu_{k+1}) \\ &= (1 - \alpha_k)\varphi_k(\nu_{k+1}) + \alpha_k f(J_{\lambda_k} y_k) \\ &\quad + \frac{\alpha_k}{\lambda_k} \langle y_k - J_{\lambda_k} y_k, \nu_{k+1} - J_{\lambda_k} y_k \rangle \\ &= (1 - \alpha_k)\varphi_k^* + \frac{A_{k+1}}{2} \|\nu_{k+1} - \nu_k\|^2 + \alpha_k f(J_{\lambda_k} y_k) \\ &\quad + \frac{\alpha_k}{\lambda_k} \langle y_k - J_{\lambda_k} y_k, \nu_{k+1} - J_{\lambda_k} y_k \rangle. \end{aligned} \tag{2.14}$$

Since by assumption $\varphi_k^* \geqq f(x_k)$, we obtain from (2.9) that

$$\varphi_k^* \geqq f(x_k) \geqq f(J_{\lambda_k} y_k) + \langle (y_k - J_{\lambda_k} y_k)/\lambda_k, x_k - J_{\lambda_k} y_k \rangle.$$

Using this in (2.14), we obtain

$$\begin{aligned} \varphi_{k+1}^* \geqq f(J_{\lambda_k} y_k) &+ \frac{A_{k+1}}{2} \|\nu_{k+1} - \nu_k\|^2 \\ &+ \frac{1}{\lambda_k} \langle y_k - J_{\lambda_k} y_k, (1 - \alpha_k) x_k + \alpha_k \nu_{k+1} - J_{\lambda_k} y_k \rangle. \end{aligned} \tag{2.15}$$

The term $(1 - \alpha_k) x_k + \alpha_k \nu_{k+1} - J_{\lambda_k} y_k$ above can be written as

$$((1 - \alpha_k) x_k + \alpha_k \nu_k - y_k) + \alpha_k (\nu_{k+1} - \nu_k) + (y_k - J_{\lambda_k} y_k).$$

Substituting the value of $\nu_{k+1}-\nu_k$ from the formula (2.12) into the second term above, we see that the scalar product term in (2.15) can be written as

$$\langle y_k - J_{\lambda_k}y_k, (1-\alpha_k)x_k+\alpha_k\nu_k-y_k\rangle+\left(1-\frac{\alpha_k^2}{A_{k+1}\lambda_k}\right)\|y_k-J_{\lambda_k}y_k\|^2. \tag{2.16}$$

Also, substituting the value of $\nu_{k+1}-\nu_k$ in (2.12) into the second term of (2.15), we obtain

$$\frac{A_{k+1}}{2}\,\|\nu_{k+1}-\nu_k\|^2=\frac{\alpha_k^2}{2A_{k+1}\lambda_k^2}\,\|y_k-J_{\lambda_k}y_k\|^2. \tag{2.17}$$

We obtain (2.13) by using lines (2.16) and (2.17) in (2.15). This proves the theorem. □

COROLLARY 2.1. *If, in Theorem* 2.1, *we choose*

$$y_k=(1-\alpha_k)x_k+\alpha_k\nu_k, \tag{2.18}$$

then

$$\varphi_{k+1}^*\geqq f(J_{\lambda_k}y_k)+\frac{1}{2\lambda_k}\left(2-\frac{\alpha_k^2}{A_{k+1}\lambda_k}\right)\|y_k-J_{\lambda_k}y_k\|^2. \tag{2.19}$$

Corollary 2.1 suggests many possibilities for obtaining convergent proximal point algorithms. For example, we can choose

$$x_{k+1}=J_{\lambda_k}y_k:=\arg\min_{z\in\mathbf{R}^n}\left\{f(z)+\frac{1}{2\lambda_k}\,\|z-y_k\|^2\right\}, \tag{2.20}$$

$$\alpha_k^2=A_{k+1}\lambda_k:=(1-\alpha_k)A_k\lambda_k. \tag{2.21}$$

COROLLARY 2.2. *If* y_k *is chosen as in* (2.18), x_{k+1} *is chosen as in* (2.20), *and* α_k *is chosen as in* (2.21), *then*

$$\varphi_{k+1}^*\geqq f(x_{k+1})+\frac{1}{2\lambda_k}\,\|y_k-x_{k+1}\|^2\geqq f(x_{k+1}). \tag{2.22}$$

Our proximal point algorithm chooses y_k, x_{k+1}, and α_k according to Corollary 2.2.

THE PROXIMAL POINT ALGORITHM.

Initialization. Choose a feasible starting point $x_0\in\mathbf{R}^n$ ($f(x_0)<\infty$), and constants $\lambda_0>0$ and $A>0$. Define $\nu_0:=x_0$, $A_0:=A$.

Step k, $k\geqq 0$:

(a) Choose $\lambda_k>0$, and calculate $\alpha_k>0$ from the equation $\alpha_k^2=(1-\alpha_k)A_k\lambda_k$, that is,

$$\alpha_k=\frac{\sqrt{(A_k\lambda_k)^2+4A_k\lambda_k}-A_k\lambda_k}{2}. \tag{2.23}$$

(b) Define

$$\begin{aligned} y_k&=(1-\alpha_k)x_k+\alpha_k\nu_k,\\ x_{k+1}&:=J_{\lambda_k}y_k=\arg\min_{z\in\mathbf{R}^n}\left\{f(z)+\frac{1}{2\lambda_k}\,\|z-y_k\|^2\right\},\\ \nu_{k+1}&=\nu_k+\frac{1}{\alpha_k}\,(x_{k+1}-y_k),\\ A_{k+1}&=(1-\alpha_k)A_k. \end{aligned} \tag{2.24}$$

Remark 2.1. In the algorithm above, the starting point x_0 must be feasible. However, this is not a serious restriction since one preliminary iteration of the classical proximal point algorithm can generate such a feasible point x_0.

In order to estimate the convergence rate of the algorithm, we must estimate the magnitude of β_k, as the inequality (2.6) shows. The following result gives tight bounds for β_k. The upper bound below is an extension of a result in Nesterov [14]. The lower bound is needed in § 3, Lemma 3.3.

LEMMA 2.2.

$$\frac{1}{(1+\sqrt{A}\sum_{j=0}^{k-1}\sqrt{\lambda_j})^2} \leqq \beta_k \leqq \frac{1}{(1+(\sqrt{A}/2)\sum_{j=0}^{k-1}\sqrt{\lambda_j})^2}. \tag{2.25}$$

Proof. We first prove the upper bound on β_k in (2.25). From (2.2), we obtain $\beta_{k+1}=(1-\alpha_k)\beta_k$, which implies $\alpha_k=1-\beta_{k+1}/\beta_k$. Also, from (2.11), $A_{k+1}=\beta_{k+1}A$. Substituting this in (2.21) results in

$$\left(1-\frac{\beta_{k+1}}{\beta_k}\right)^2=\beta_{k+1}A\lambda_k.$$

We make the substitution $\beta_k=\mu_k^{-2}$ in the equality above. Taking the square roots of both sides of this equality and then multiplying both sides of the resulting equality by μ_{k+1}^2, we obtain

$$\mu_{k+1}^2-\mu_k^2=\mu_{k+1}\sqrt{A\lambda_k}. \tag{2.26}$$

It is easy to show that $2\mu_{k+1}(\mu_{k+1}-\mu_k)\geqq\mu_{k+1}^2-\mu_k^2$. Using this in (2.26) results in

$$2\mu_{k+1}(\mu_{k+1}-\mu_k)\geqq\mu_{k+1}\sqrt{A\lambda_k},$$

which implies $\mu_{k+1}-\mu_k\geqq\sqrt{A\lambda_k}/2$. Summing this inequality for $j=0,1,\dots,k-1$ and noting $\mu_0=1$, we obtain

$$\mu_k\geqq 1+\frac{\sqrt{A}}{2}\sum_{j=0}^{k-1}\sqrt{\lambda_j}.$$

Substituting $\mu_k=\beta_k^{-1/2}$ above proves the upper bound on β_k.

It remains to prove the lower bound on β_k. Note that $\beta_{k+1}\leqq\beta_k$ implies $\mu_{k+1}\geqq\mu_k$. Thus $\mu_{k+1}(\mu_{k+1}-\mu_k)\leqq\mu_{k+1}^2-\mu_k^2$. Using this in (2.26), we obtain $\mu_{k+1}-\mu_k\leqq\sqrt{A\lambda_k}$. As above, summing this inequality for $j=0,1,\dots,k-1$, we obtain

$$\mu_k\leqq 1+\sqrt{A}\sum_{j=0}^{k-1}\sqrt{\lambda_j}.$$

Substituting $\mu_k=\beta_k^{-1/2}$ above proves the lower bound on β_k. □

From Corollary 2.2 and inequality (2.3), we obtain the following basic convergence rate result.

THEOREM 2.2. *For any feasible point $x\in\mathbf{R}^n$, the proximal point algorithm stated above has the global convergence rate estimate*

$$\begin{aligned} f(x_k)-f(x)+\frac{1}{2\lambda_{k-1}}\|y_{k-1}-x_k\|^2 &\leqq \frac{f(x_0)-f(x)+(A/2)\|x-x_0\|^2}{(1+(\sqrt{A}/2)\sum_{j=0}^{k-1}\sqrt{\lambda_j})^2} \\ &= O\left(\frac{1}{(\sum_{j=0}^{k-1}\sqrt{\lambda_j})^2}\right). \end{aligned} \tag{2.27}$$

Remark 2.2. The convergence rate estimate above is given in terms of the objective function gap $f(x_k)-f^*$. The convergence of the points $\{x_k\}$ is a future research topic.

Even if we can show that the sequence $\{x_k\}$ converges to an optimal solution x^*, it is unlikely that a convergence rate can be provided for $\{\|x_k - x^*\|\}$ without further assumptions on f. Of course, if f is strongly convex, then, using the standard properties of strongly convex functions, we can show that $\{x_k\}$ converges to the unique optimal solution of f and that

$$\|x_k - x^*\| = O\left(\frac{1}{\sum_{j=0}^{k-1} \sqrt{\lambda_j}}\right).$$

Also, in certain problems (for example, linear programming) we can prove that $\{x_k\}$ converges to the optimal set and provide an estimate of the convergence rate; see Güler [8].

Remark 2.3. The term $\|y_{k-1} - x_k\|^2/(2\lambda_{k-1})$ in (2.27) is not strictly necessary to obtain convergence estimates for the algorithm presented in this section. However, it will be crucial in the next section where we present a relaxed proximal point algorithm in which x_{k+1} is calculated only approximately:

$$x_{k+1} \approx J_{\lambda_k} y_k := \arg\min_{z \in \mathbf{R}^n} \left\{ f(z) + \frac{1}{2\lambda_k} \|z - y_k\|^2 \right\}.$$

It is also crucial in proving the finite termination of the augmented Lagrangian algorithm for linear programming in Güler [8], which is an application of the algorithm presented in this section.

The convergence rate of the proximal point algorithm is summarized below.

THEOREM 2.3. *Suppose f has a minimizer x^* and $f^* = f(x^*) = \min_{z \in \mathbf{R}^n} f(z)$. Denote the set of minimizers of f by X^*. The proximal point algorithm above possesses the global convergence rate estimate*

$$f(x_k) - f^* \leqq \frac{4}{A(\sum_{j=0}^{k-1} \sqrt{\lambda_j})^2} \left(f(x_0) - f^* + \frac{A}{2} \rho(x_0, X^*)^2 \right). \tag{2.28}$$

The algorithm converges, that is, $f(x_k) \to f^*$ if

$$\sum_{k=0}^{\infty} \sqrt{\lambda_k} = \infty. \tag{2.29}$$

In particular, if $\lambda_k \geqq \lambda > 0$, we have the convergence rate estimate

$$\begin{aligned} f(x_k) - f^* &\leqq \frac{4/(A\lambda)}{k^2} \left(f(x_0) - f^* + \frac{A}{2} \rho(x_0, X^*)^2 \right) \\ &= O\left(\frac{1}{k^2}\right). \end{aligned} \tag{2.30}$$

Remark 2.4. The convergence rate of our proximal point algorithm given in (2.26) compares favorably with the convergence rate estimate

$$f(x_k) - f^* \leqq \frac{\rho(x_0, X^*)^2}{2\sum_{j=0}^{k-1} \lambda_j} \tag{2.31}$$

obtained in Güler [7] for the classical proximal point algorithm. It is clear that the convergence rate estimate (2.28) is faster than (2.31). Moreover, it is shown in Güler [7] (Remark 2.1) that the condition $\sum_{k=0}^{\infty} \lambda_k = \infty$ is necessary and sufficient for the convergence of the classical proximal point algorithm. In contrast, the algorithm presented here converges under the weaker condition (2.29).

Further properties of the algorithm are given below. Note that (2.34) follows from (2.32) because of (2.8).

COROLLARY 2.3. *If* (2.29) *holds true in the proximal point algorithm, then*

$$\frac{\|x_{k+1}-y_k\|^2}{\lambda_k} \to 0. \tag{2.32}$$

If the sequence $\{\lambda_k\}$ *is bounded from above, then*

$$\|x_{k+1}-y_k\| \to 0. \tag{2.33}$$

Also,

$$\lambda_k \rho(0, \partial f(x_{k+1}))^2 \to 0. \tag{2.34}$$

If the sequence $\{\lambda_k\}$ *is bounded away from* 0, *then*

$$\rho(0, \partial f(x_k)) \to 0. \tag{2.35}$$

If f is differentiable, (2.35) *means that*

$$\|f'(x_k)\| \to 0. \tag{2.36}$$

Remark 2.5. Ekeland's ε-variational principle (see Aubin and Ekeland [1, Chap. 5]) can be used to prove that if f is bounded from below, that is, $f^* := \inf_{x\in\mathbf{R}^n} f(x) > -\infty$, then there exist x_k and $w_k \in \partial f(x_k)$ such that $f(x_k) \to f^*$ and $w_k \to 0$. A slight generalization of Corollary 2.3 shows that such x_k and w_k can be generated by our proximal point algorithm.

3. The algorithm with inexact minimization. In the proximal point algorithm presented in § 2, x_{k+1} is given by

$$x_{k+1} = J_{\lambda_k} y_k := \arg\min_{z\in\mathbf{R}^n} \left\{ f(z) + \frac{1}{2\lambda_k} \|z-y_k\|^2 \right\}. \tag{3.1}$$

The point x_{k+1} is thus the exact minimum of the augmented function

$$\phi_k(z) := f(z) + \frac{1}{2\lambda_k} \|z-y_k\|^2. \tag{3.2}$$

The calculation of x_{k+1} can be almost as difficult to solve as the original minimization problem (1.1). In this section, we show that a modification of the algorithm in § 2, which requires only an approximate minimization of ϕ_k, that is,

$$x_{k+1} \approx J_{\lambda_k} y_k := \arg\min_{z\in\mathbf{R}^n} \phi_k(z), \tag{3.3}$$

still yields a convergent proximal point algorithm.

DEFINITION 3.1. We will say that x_{k+1} is an approximation minimizer of ϕ_k if the following criterion **A**′ in Rockafellar [18, pp. 880] is satisfied:

$$\rho(0, \partial\phi_k(x_{k+1})) \leqq \frac{\varepsilon_k}{\lambda_k}. \tag{A′}$$

Note that if f is differentiable, condition **A**′ means that

$$\|\phi_k'(x_{k+1})\| \leqq \frac{\varepsilon_k}{\lambda_k}.$$

We will give conditions on the magnitude of the errors ε_k which are sufficient to obtain convergent algorithms.

The result below shows that x_{k+1} is in fact an approximate minimizer of ϕ_k. It will be needed later in this section.

LEMMA 3.1. *Let* $\phi_k^* = \min_{z\in\mathbf{R}^n} \phi_k(z)$. *If* x_{k+1} *satisfies condition* $\mathbf{A}'$, *then*

$$\frac{1}{2\lambda_k}\|x_{k+1}-J_{\lambda_k}y_k\|^2 \leqq \phi_k(x_{k+1})-\phi_k^* \leqq \frac{\varepsilon_k^2}{2\lambda_k}. \tag{3.4}$$

Proof. We start by proving the first inequality. By definition, $J_{\lambda_k}y_k$ is the exact minimizer of ϕ_k. Since ϕ_k is strongly convex with modulus $1/\lambda_k$ and $0\in\phi_k(J_{\lambda_k}y_k)$, it follows from Proposition 6(c) of Rockafellar [18] that

$$\begin{aligned}\phi_k(x_{k+1})-\phi_k^* &= \phi_k(x_{k+1})-\phi_k(J_{\lambda_k}y_k)\\ &\geqq \langle 0, x_{k+1}-J_{\lambda_k}y_k\rangle+\frac{1}{2\lambda_k}\|x_{k+1}-J_{\lambda_k}y_k\|^2\\ &=\frac{1}{2\lambda_k}\|x_{k+1}-J_{\lambda_k}y_k\|^2.\end{aligned}$$

This proves the first inequality.

It remains to prove the second inequality. Let $w_k\in\partial\phi_k(x_{k+1})$ be such that $\|w_k\|\leqq \varepsilon_k/\lambda_k$. Since ϕ_k is strongly convex with modulus $1/\lambda_k$, and $\|w_k\|\leqq\varepsilon_k/\lambda_k$, we have

$$\begin{aligned}\phi_k(J_{\lambda_k}y_k)-\phi_k(x_{k+1}) &\geqq \langle w_k, J_{\lambda_k}y_k-x_{k+1}\rangle+\frac{1}{2\lambda_k}\|J_{\lambda_k}y_k-x_{k+1}\|^2\\ &\geqq -\|w_k\|\,\|J_{\lambda_k}y_k-x_{k+1}\|+\frac{1}{2\lambda_k}\|J_{\lambda_k}y_k-x_{k+1}\|^2\\ &\geqq -\frac{\varepsilon_k}{\lambda_k}\|J_{\lambda_k}y_k-x_{k+1}\|+\frac{1}{2\lambda_k}\|J_{\lambda_k}y_k-x_{k+1}\|^2\\ &\geqq \frac{1}{\lambda_k}\min_{t\in\mathbf{R}}\left\{\frac{1}{2}t^2-\varepsilon_k t\right\}\\ &= -\frac{\varepsilon_k^2}{2\lambda_k},\end{aligned}$$

where the first inequality again follows from Proposition 6 in Rockafellar [18]. This proves the lemma. □

COROLLARY 3.1. *If* x_{k+1} *is chosen according to criterion* $\mathbf{A}'$, *then*

$$\|x_{k+1}-J_{\lambda_k}y_k\|\leqq\varepsilon_k. \tag{3.5}$$

Corollary 3.1 is proved in a more general context in Rockafellar [18, Prop. 3].

We will need the following slight generalization of Theorem 2.1. Its proof is similar to that of the original Theorem 2.1.

LEMMA 3.2. *If, for some* $k\geqq 0$, x_k *satisfies the inequality*

$$f(x_k)\leqq\varphi_k^*+\delta_k, \tag{3.6}$$

then for any $y_k\in\mathbf{R}^n$, $\lambda_k>0$, *and* $\alpha_k\in[0,1)$, *the following inequality holds true*:

$$\begin{aligned}\varphi_{k+1}^*+(1-\alpha_k)\delta_k &\geqq f(J_{\lambda_k}y_k)+\frac{1}{2\lambda_k}\left(2-\frac{\alpha_k^2}{A_{k+1}\lambda_k}\right)\|y_k-J_{\lambda_k}y_k\|^2\\ &\quad+\frac{1}{\lambda_k}\langle y_k-J_{\lambda_k}y_k, (1-\alpha_k)x_k+\alpha_k v_k-y_k\rangle.\end{aligned} \tag{3.7}$$

Also, Corollary 2.2 generalizes to Corollary 3.2.

COROLLARY 3.2. *If y_k is chosen as in* (2.18) *and α_k is chosen as in* (2.21), *then*

$$\varphi_{k+1}^* + (1-\alpha_k)\delta_k \geqq f(J_{\lambda_k} y_k) + \frac{1}{2\lambda_k}\|J_{\lambda_k} y_k - y_k\|^2 = \phi_k^*. \tag{3.8}$$

The following result estimates how the individual errors $\{\varepsilon_j\}_{j=0}^{k-1}$ at each step accumulate to a total error δ_k at step k.

THEOREM 3.1. *If, in the algorithm in* § 2, *x_{k+1} is calculated according to criterion* A′ *instead of* (2.20), *then*

$$f(x_k) \leqq \varphi_k^* + \delta_k, \tag{3.9}$$

where $\{\delta_k\}_{k=0}^{\infty}$ satisfies the difference equation

$$\delta_0 = 0, \quad \delta_{k+1} = (1-\alpha_k)\delta_k + \frac{\varepsilon_k^2}{2\lambda_k}, \quad k = 0, 1, \ldots. \tag{3.10}$$

Proof. We prove (3.9) and (3.10) by induction. Since $f(x_0) = \varphi_0^*$, they are true for $k=0$. Suppose (3.9) and (3.10) hold true for k. We will show that they also hold true for $k+1$. We have

$$\begin{aligned} \varphi_{k+1}^* + (1-\alpha_k)\delta_k &\geqq \phi_k^* \quad \text{(from (3.8))} \\ &\geqq \phi_k(x_{k+1}) - \frac{\varepsilon_k^2}{2\lambda_k} \quad \text{(from Lemma 3.1)} \\ &= f(x_{k+1}) + \frac{1}{2\lambda_k}\|x_{k+1} - y_k\|^2 - \frac{\varepsilon_k^2}{2\lambda_k}, \end{aligned}$$

which implies

$$f(x_{k+1}) + \frac{1}{2\lambda_k}\|x_{k+1} - y_k\|^2 \leqq \varphi_{k+1}^* + \delta_{k+1}.$$

This proves the theorem. □

Note that Lemma 2.1 still holds true, so that (2.3) is valid. Combining (2.3) and Theorem 3.1 results in the following theorem.

THEOREM 3.2. *In the modified proximal point algorithm in which the point x_k is calculated according to criterion* A′, *we have for any $x \in \mathbf{R}^n$, the convergence rate estimate*

$$f(x_k) - f(x) \leqq \beta_k(\varphi_0(x) - f(x)) + \delta_k,$$

where $\{\delta_k\}$ satisfies the difference equation (3.10). *In particular, we have the convergence rate estimate*

$$f(x_k) - f^* \leqq \beta_k\left(f(x_0) - f^* + \frac{A}{2}\rho(x_0, X^*)^2\right) + \delta_k. \tag{3.11}$$

From (3.11) we see that in order for the modified algorithm to converge, we must have $\delta_k \to 0$. In the next result, we obtain bounds on δ_k.

LEMMA 3.3. *The solution to the difference equation* (3.10) *is given by*

$$\delta_k = \frac{\beta_k}{2} \cdot \sum_{j=0}^{k-1} \frac{\varepsilon_j^2/\lambda_j}{\beta_{j+1}}. \tag{3.12}$$

Moreover,

$$\delta_k \leqq 2\sum_{j=0}^{k-1} \frac{\varepsilon_j^2}{\lambda_j}\left(\frac{1+\sqrt{A}\sum_{i=0}^{j}\sqrt{\lambda_i}}{1+\sqrt{A}\sum_{j=0}^{k-1}\sqrt{\lambda_j}}\right)^2. \tag{3.13}$$

Assume $\{\lambda_k\}_{k=0}^{\infty}$ is an increasing sequence or, more generally, that there exists a constant $M>0$ such that

$$\lambda_i \leqq M\lambda_j \quad \textit{whenever } i \leqq j, \tag{3.14}$$

and that for some $\sigma>0$

$$\varepsilon_k = O(1/k^{\sigma}), \qquad k=1,2,\dots; \tag{3.15}$$

that is, there is a constant $c>0$ such that $\varepsilon_k \leqq c/k^{\sigma}$ for all $k \geqq 1$. Then

$$\delta_k = O\left(\frac{1}{k^{2\sigma-1}}\right). \tag{3.16}$$

Proof. Since $1-\alpha_j = \beta_{j+1}/\beta_j$, for any $j \geqq 0$, (3.10) can be written as $\delta_{j+1} = (\beta_{j+1}/\beta_j)\delta_j + \varepsilon_j^2/(2\lambda_j)$. Dividing this equality by β_{j+1}, and rearranging its terms, we obtain

$$\frac{\delta_{j+1}}{\beta_{j+1}} - \frac{\delta_j}{\beta_j} = \frac{\varepsilon_j^2/\lambda_j}{2\beta_{j+1}}. \tag{3.17}$$

Summing (3.17) for $j=0,1,\dots,k-1$, and noting $\delta_0=0$, we obtain (3.12).

Inequality (3.13) is obtained from (3.12) by using the lower bound on β_{j+1} given in Lemma 2.2.

It remains to prove (3.16). If (3.14) is true, then $\sum_{k=0}^{\infty}\sqrt{\lambda_k}=\infty$. Thus there exists a constant $c>0$ such that $1+\sqrt{A}\sum_{i=0}^{j}\sqrt{\lambda_i} \leqq c\sqrt{A}\sum_{i=0}^{j}\sqrt{\lambda_i}$. We deduce from (3.13) that there are constants $c>0$ (not the same constant c above) and $\bar{c}>0$ such that

$$\delta_k \leqq c\sum_{j=0}^{k-1}\varepsilon_j^2\left(\frac{\sum_{i=0}^{j}\sqrt{\lambda_i/\lambda_j}}{\sum_{j=0}^{k-1}\sqrt{\lambda_j}}\right)^2 \leqq \bar{c}\cdot\frac{\sum_{j=0}^{k-1}(j\varepsilon_j)^2}{k^2}. \tag{3.18}$$

If ε_k satisfies (3.15), there exists a constant $\tilde{c}>0$ such that

$$\sum_{j=0}^{k-1}(j\varepsilon_j)^2 \leqq \tilde{c}\int_0^k t^{2-2\sigma}\,dt = \frac{\tilde{c}}{3-2\sigma}k^{3-2\sigma}.$$

Using this estimate in (3.18) proves (3.16). □

The theorem below, which summarizes the results of this section, gives the convergence rate estimates for the proximal point with errors. It is obtained from Theorem 3.2 and Lemmas 2.2 and 3.3.

THEOREM 3.3. *Consider a proximal point algorithm that differs from the one stated in § 2 only in that the point x_k is approximately calculated according to criterion* $\mathbf{A}'$ *with an error ε_k. Assume that errors $\{\varepsilon_k\}$ satisfy condition* (3.15) *for some $\sigma>\frac{1}{2}$, and that parameters $\{\lambda_k\}$ are chosen according to condition* (3.14). *Then, for any feasible $x \in \mathbf{R}^n$,*

$$f(x_k)-f(x) \leqq O\left(\frac{1}{k^2}\right)+O\left(\frac{1}{k^{2\sigma-1}}\right) \to 0.$$

In particular,

$$f(x_k)-f^* = O\left(\frac{1}{k^2}\right)+O\left(\frac{1}{k^{2\sigma-1}}\right),$$

and if $\sigma \geqq \frac{3}{2}$,

$$f(x_k)-f^* = O\left(\frac{1}{k^2}\right).$$

Remark 3.1. Theorem 3.3 can be compared with results in Rockafellar [18] (see also Brézis and Lions [3, pp. 343]). Rockafellar proves that under condition A′ or condition (3.5) (which he calls condition A) together with the condition

$$\sum_{k=0}^{\infty} \varepsilon_k < \infty, \tag{3.19}$$

the classical proximal point algorithm converges for a maximal monotone operator. In [3] and [18] convergence means that $x_k \to x^*$ to some solution of the maximal monotone operator. Rockafellar shows that (3.19) is a necessary and sufficient condition for convergence. Our sense of convergence is different from the one in [3] and [18] in that we require only that $f(x_k) \to f^*$. However, our condition (3.15) on $\{\varepsilon_k\}$ is somewhat weaker than (3.19) and we are able to prove the convergence rates in Theorem 3.3. It is interesting to note that such convergence rates for $f(x_k) - f^*$ are not currently available for the inexact minimization version of the classical proximal minimization algorithm.

4. Further properties of the algorithms. In this section, we develop monotonic versions of the algorithms presented in §§ 2 and 3. We also show that all algorithms minimize f even if f has no minimizers or is unbounded from below.

The proximal point algorithm developed in the previous sections need not be monotonic, that is, we may have $f(x_{k+1}) > f(x_k)$. Here we present monotonic versions of the algorithms and discuss their convergence properties.

We obtain the monotonic version of the algorithm in § 2 simply by replacing the equation defining x_{k+1} in (2.24) with the following:

$$\bar{x}_{k+1} = J_{\lambda_k} y_k, \qquad x_{k+1} = \arg\min \{f(\bar{x}_{k+1}), f(x_k)\}.$$

THE MONOTONIC PROXIMAL POINT ALGORITHM.

Initialization. Choose a feasible starting point $x_0 \in \mathbf{R}^n$ ($f(x_0) < \infty$), and constants $\lambda_0 > 0$ and $A > 0$. Define $v_0 := x_0$, $A_0 := A$.

Step k, $k \geqq 0$:

(a) Choose $\lambda_k > 0$, and set

$$\alpha_k = \frac{\sqrt{(A_k\lambda_k)^2 + 4A_k\lambda_k} - A_k\lambda_k}{2}.$$

(b) Define

$$y_k = (1-\alpha_k)x_k + \alpha_k v_k,$$

$$\bar{x}_{k+1} = J_{\lambda_k} y_k := \arg\min_{z \in \mathbf{R}^n} \left\{ f(z) + \frac{1}{2\lambda_k} \|z - y_k\|^2 \right\},$$

$$x_{k+1} = \arg\min \{f(\bar{x}_{k+1}), f(x_k)\},$$

$$v_{k+1} = v_k + \frac{1}{\alpha_k}(x_{k+1} - y_k),$$

$$A_{k+1} = (1-\alpha_k)A_k.$$

It is easy to verify that the algorithm stated above possesses the same global convergence rate estimates as the original version in § 2. The statement and the properties of the algorithm of § 3 are similar.

The next result shows that the algorithms in this paper minimize f in the case when f has no minimizers or is even unbounded from below.

THEOREM 4.1. *Suppose f has no minimizers or is unbounded from below. The original proximal point algorithms in* §§2 *and* 3, *and their monotonic versions discussed above, minimize f, that is,*

$$\lim_{k\to\infty} f(x_k) = \inf_{x\in\mathbf{R}^n} f(x). \tag{4.1}$$

Proof. Since the algorithm in § 2 is a special case of the algorithm in § 3, we prove (4.1) only for the latter. The proofs of (4.1) for the monotonic versions are similar.

We first consider the case $f^* > -\infty$. Suppose $\varepsilon > 0$ is given. Let x^ε be a point satisfying $f(x^\varepsilon) - f^* \leqq \varepsilon/2$. If k is large enough, from (3.23) we obtain $f(x_k) - f(x^\varepsilon) \leqq \varepsilon/2$. Thus $f(x_k) - f^* \leqq \varepsilon$, and (4.1) holds true.

If $f^* = -\infty$, let M be an arbitrary number and x^M be a point satisfying $f(x^M) \leqq M$. If k is large enough, from (3.23) we have $f(x_k) - f(x^M) \leqq \varepsilon$. Thus $f(x_k) \leqq M + \varepsilon$ and (4.1) holds true. □

5. Concluding remarks. In this paper, we presented new proximal point algorithms for the convex minimization problem (1.1). We presented an exact minimization algorithm in § 2, and an inexact minimization algorithm in § 3. The algorithm in § 3 is important in practice, since the exact minimization of the auxiliary function that occurs at each step is impractical, and may in fact be almost as difficult to solve as the original minimization problem. We demonstrated the convergence of our algorithms and supplied global convergence rates for them. These rates are faster than the rates the author obtained [7] for the classical proximal point algorithm. Thus our algorithms accelerate the classical proximal point algorithm.

The algorithms developed here are general enough to solve the general convex program (1.4). When applied to the dual program (1.5), they give rise to the so-called *augmented Lagrangian* methods discussed in Bertsekas [2], Rockafellar [19], and others. In [8], the author applies the algorithm in § 2 to linear programming and obtains an algorithm that accelerates the augmented Lagrangian method of Polyak and Tret'iakov [15]. As is true of the algorithm of Polyak and Treti'akov [15], the application of the exact minimization algorithm in § 2 to linear programming terminates in finitely many iterations.

For simplicity's sake we have kept our discussion to finite-dimensional Euclidean spaces $\mathbf{R}^n$, however our results and algorithms are valid in any Hilbert space. Thus our algorithms may be applied to infinite-dimensional variational problems; see [4], [5], [6] and [10].

6. Appendix. Another proximal point algorithm. In this appendix, we present a second proximal point algorithm for problem (1.1). This algorithm uses ideas from Nesterov [13], where the first optimal algorithm for smooth convex programming is introduced.

The algorithm generates a sequence $\{x_k\}_{k=0}^\infty$ of approximations to an optimal point $x^* \in \mathbf{R}^n$ of problem (1.1), as well as an auxiliary sequence of points $\{y_k\}_{k=1}^\infty$.

THE SECOND PROXIMAL POINT ALGORITHM.

Initialization. Choose a point $x_0 \in \mathbf{R}^n$, and a constant $\lambda > 0$. Define $y_1 := x_0$, $\lambda_1 := \lambda$, and $\beta_1 := 1$.

Step **k**, $k \geqq 1$. Choose $\lambda_k \geqq \lambda_{k-1}$ and define

$$\beta_{k+1} = \frac{1+\sqrt{1+4\beta_k^2}}{2}, \tag{6.1}$$

$$x_k = J_{\lambda_k} y_k := \arg\min_{x \in \mathbf{R}^n} \left\{ f(x) + \frac{1}{2\lambda_k} \|x - y_k\|^2 \right\}, \tag{6.2}$$

$$y_{k+1} = x_k + \frac{\beta_k - 1}{\beta_{k+1}} (x_k - x_{k-1}) + \frac{\beta_k}{\beta_{k+1}} (x_k - y_k). \tag{6.3}$$

THEOREM 6.1. *The proximal point algorithm stated above possesses the global convergence rate estimate*

$$f(x_k) - \min_{x \in \mathbf{R}^n} f(x) \leqq \frac{1}{\lambda (k+1)^2} \rho(x_0, X^*)^2, \tag{6.4}$$

where X^* is the set of minimizers of f.

Proof. From (2.8) and (6.2), we have

$$\frac{y_{i+1} - x_{i+1}}{\lambda_{i+1}} \in \partial f(x_{i+1}), \qquad i = 0, 1, \ldots .$$

Since f is convex, we have

$$f(x_i) - f(x_{i+1}) \geqq \frac{1}{\lambda_{i+1}} \langle y_{i+1} - x_{i+1}, x_i - x_{i+1} \rangle, \tag{6.5}$$

$$f(x^*) - f(x_{i+1}) \geqq \frac{1}{\lambda_{i+1}} \langle y_{i+1} - x_{i+1}, x^* - x_{i+1} \rangle. \tag{6.6}$$

Note that (6.1) implies

$$\beta_{i+1}(\beta_{i+1} - 1) = \beta_i^2. \tag{6.7}$$

For brevity, we define $W_i := f(x_i) - f(x^*)$. Multiplying (6.5) by $\beta_i^2 = \beta_{i+1}(\beta_{i+1} - 1)$ and (6.6) by β_{i+1}, and using (6.7), we obtain

$$\beta_i^2 (W_i - W_{i+1}) \geqq \frac{1}{\lambda_{i+1}} \langle \beta_{i+1}(y_{i+1} - x_{i+1}), (\beta_{i+1} - 1)(x_i - x_{i+1}) \rangle, \tag{6.8}$$

$$-\beta_{i+1} W_{i+1} \geqq \frac{1}{\lambda_{i+1}} \langle \beta_{i+1}(y_{i+1} - x_{i+1}), x^* - x_{i+1} \rangle. \tag{6.9}$$

Adding (6.8) and (6.9), and using (6.7), we obtain

$$\beta_i^2 W_i - \beta_{i+1}^2 W_{i+1} \geqq \frac{1}{\lambda_{i+1}} \langle \beta_{i+1}(y_{i+1} - x_{i+1}), \beta_{i+1}(x_i - x_{i+1}) + x^* - x_i \rangle. \tag{6.10}$$

Using the polarization identity $4\langle x, y \rangle = \|x + y\|^2 - \|x - y\|^2$, the scalar product term in (6.10) can be expressed as

$$\tfrac{1}{4} \|\beta_{i+1}(x_i + y_{i+1} - 2x_{i+1}) + x^* - x_i\|^2 - \tfrac{1}{4} \|\beta_{i+1}(x_i - y_{i+1}) + x^* - x_i\|^2. \tag{6.11}$$

Let us define

$$\theta_i := \beta_{i+1}(x_i - y_{i+1}) + x^* - x_i, \qquad i = 0, 1, \ldots . \tag{6.12}$$

Using (6.3), it is easy to show that

$$\theta_i = \beta_i (x_{i-1} + y_i - 2x_i) + x^* - x_{i-1}, \qquad i = 1, 2, \ldots . \tag{6.13}$$

From (6.12) and (6.13), respectively, we see that the second term in (6.11) equals θ_i and the first term in (6.11) equals θ_{i+1}. Using these facts and the fact that $\lambda_{i+1} \geqq \lambda_i$ in (6.10), we obtain

$$\beta_i^2 W_i - \beta_{i+1}^2 W_{i+1} \geqq \frac{1}{4\lambda_{i+1}} \|\theta_{i+1}\|^2 - \frac{1}{4\lambda_i} \|\theta_i\|^2. \tag{6.14}$$

Summing (6.14) for $i = 1, \dots, k-1$, we obtain

$$\beta_1^2 W_1 - \beta_k^2 W_k \geqq \frac{1}{4\lambda_k} \|\theta_k\|^2 - \frac{1}{4\lambda_1} \|\theta_1\|^2 \geqq -\frac{1}{4\lambda_1} \|\theta_1\|^2.$$

Since $\lambda_1 = \lambda$ and $\beta_1 = 1$, we obtain from the last inequality,

$$\beta_k^2 W_k \leqq W_1 + \frac{1}{4\lambda} \|\theta_1\|^2. \tag{6.15}$$

Using (6.6) with $i = 0$, and noting $x_0 = y_1$, we have

$$\begin{aligned} -W_1 &\geqq \frac{1}{\lambda} \langle y_1 - x_1, x^* - x_1 \rangle \\ &= \frac{1}{4\lambda} \|x^* + y_1 - 2x_1\|^2 - \frac{1}{4\lambda} \|x^* - x_0\|^2 \qquad (6.16) \\ &= \frac{1}{4\lambda} \|\theta_1\|^2 - \frac{1}{4\lambda} \|x^* - x_0\|^2 \quad \text{(using (6.13))}. \end{aligned}$$

Thus, from (6.15) and (6.16), we obtain

$$\beta_k^2 W_k \leqq \frac{1}{4\lambda} \|x_0 - x^*\|^2. \tag{6.17}$$

It is easy to show by induction that $\beta_k \leqq (k+1)/2$. Since $x^* \in X^*$ is arbitrary, the theorem follows from (6.17). □

Acknowledgments. The author is grateful to Dr. Nesterov for providing him with a copy of [14]. The author also thanks a referee for helpful remarks.

REFERENCES

[1] J. P. Aubin and I. Ekeland, *Applied Nonlinear Analysis*, Interscience Publications, John Wiley, New York, 1984.

[2] D. P. Bertsekas, *Constrained Optimization and Lagrange Multiplier Methods*, Academic Press, New York, 1982.

[3] H. Brézis and P. L. Lions, *Produits infinis de résolvantes*, Israel J. Math., 29 (1978), pp. 329-345.

[4] M. Fortin and R. Glowinski, *Augmented Lagrangian Methods: Applications to Numerical Solutions of Boundary Value Problems*, North-Holland, Amsterdam, 1983.

[5] R. Glowinski, *Numerical Methods for Nonlinear Variational Problems*, Springer-Verlag, New York, 1984.

[6] R. Glowinski, J. L. Lions, and R. Tremolieres, *Numerical Analysis of Variational Inequalities*, North-Holland, Amsterdam, 1981.

[7] O. Güler, *On the convergence of the proximal point algorithm for convex minimization*, SIAM J. Control Optim., 29 (1991), pp. 403-419.

[8] ———, *Augmented Lagrangian algorithms for linear programming*, Working Paper Series 91-3, Dept. of Management Sciences, The Univ. of Iowa, Iowa City, IA, 1991; J. Optim. Theory Appl., to appear.

[9] M. R. Hestenes, *Multiplier and gradient methods*, J. Optim. Theory Appl., 4 (1969), pp. 303-320.

[10] K. Ito and K. Kunisch, *The augmented Lagrangian method for parameter estimation in elliptic systems*, SIAM J. Control Optim., 28 (1990), pp. 113-136.

[11] B. Martinet, *Regularisation, d'inéquations variationelles par approximations succesives*, Rev. Française d'Inform. Recherche Oper., 4 (1970), pp. 154-159.

[12] J. J. Moreau, *Proximité et dualité dans un espace Hilbertien*, Bull. Soc. Math. France, 93 (1965), pp. 273-299.

[13] Yu. E. Nesterov, *A method of solving a convex programming problem with convergence rate $O(1/k^2)$*, Dokl. Akad. Nauk, 269 (1983), pp. 543-547. (In Russian.) (Translated in Soviet Math. Dokl., 27 (1983), pp. 372-376.)

[14] ———, *On an approach to the construction of optimal methods of minimization of smooth convex functions*, Ekonom. i Mat. Metody, 24 (1988), pp. 509-517.

[15] B. T. Polyak and N. V. Tret'iakov, *An iterative method for linear programming and its economic interpretation*, Ekonom. i Mat. Metody, 8 (1972), pp. 740-751. (In Russian.) (Translated in Matekon, 8 (1972), pp. 81-100.)

[16] M. J. D. Powell, *A method for nonlinear constraints in minimization problems*, in Optimization, R. Fletcher, ed., Academic Press, New York, 1969, pp. 283-298.

[17] R. T. Rockafellar, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.

[18] ———, *Monotone operators and the proximal point algorithm*, SIAM J. Control Optim., 14 (1976), pp. 877-898.

[19] ———, *Augmented Lagrangians and applications of the proximal point algorithm in convex programming*, Math. Oper. Res., 1 (1976), pp. 97-116.

SIAM J. OPTIMIZATION
Vol. 2, No. 4, pp. 665-667, November 1992

008

A NECESSARY AND SUFFICIENT CONDITION FOR A CONSTRAINED MINIMUM*

J. WARGA†

Abstract. Let U be an open subset of $\mathbb{R}^n$, X a compact semi-analytic subset of U, $(f_0,f): U \to \mathbb{R}\times\mathbb{R}^n$ analytic, and $0\in f(X)$. It is proven that a point $x_0 \in X$ minimizes $f_0(x)$ subject to $f(x)=0$ if and only if $x_0\in X$ minimizes $f_0(x)+c|f(x)|^{1/N}$ for all sufficiently large c and N. This reduces the constrained minimization problem to a finite number of unconstrained problems.

Key words. constrained minimization, real analytic functions, semi-analytic sets, numerical procedures

AMS(MOS) subject classifications. 26E05, 90C30

We consider the problem of minimizing $f_0(x)$ on X subject to the constraint $f(x)=0$, where X is a compact subset of $\mathbb{R}^n$, $f_0: X\to\mathbb{R}$, and $f: X\to\mathbb{R}^m$. Let $|\cdot|$ denote the euclidean norm in $\mathbb{R}^k$. It is easy to verify that, if (f_0,f) is continuous, then a point x_0 yields this constrained minimum if and only if $f(x_0)=0$ and there exists a nondecreasing continuous "penalty" function $h: [0,\infty)\to[0,\infty)$ such that $h(0)=0$ and x_0 minimizes the scalar function $f_0(x)+h(|f(x)|)$. As suggested by a referee, this assertion follows directly from the choice of

$$h(r)=\sup\{f_0(x_0)-f_0(x) \mid |f(x)|\leqq r\} \quad \forall r \geqq 0.$$

In the general case, the determination of an appropriate function $h(\cdot)$ may be more difficult than solving the original problem. It turns out, however, that if the functions f_0 and f are analytic and the set X is defined by a finite number of analytic equalities and inequalities, which we shall henceforth assume to be the case, then we can define a "universal" form of a "penalty" function $h(\cdot)$ independent of either (f_0,f) or x_0.

Let $\bar{B}(x, r)$ denote the closed euclidean ball of center x and radius r. A semi-analytic set in $\mathbb{R}^n$ is one defined by an analytic equality or inequality or obtained from such sets by applying a finite number of elementary set operations (unions, intersections, and set-differences). Theorem 1 below provides a global necessary and sufficient condition for constrained minimum and justifies a global optimization procedure for reducing the constrained minimization problem to a *finite* sequence of unconstrained minimization problems. The Corollary of Theorem 1 defines a "universal" exact penalty function for finite-dimensional local optimization problems defined by analytic functions.

THEOREM 1. *Let U be an open subset of $\mathbb{R}^n$, X a compact semi-analytic subset of U, $(f_0,f): U\to\mathbb{R}\times\mathbb{R}^m$ analytic, and $0\in f(X)$. Then there exist $N^*\in\{1,2,\dots\}$ and $c^*>0$ such that, for every choice of real numbers $N\geqq N^*$ and $c\geqq c^*$, x_0 minimizes $f_0(x)$ on X subject to $f(x)=0$ if and only if x_0 minimizes $f_0(x)+c|f(x)|^{1/N}$ on X. Furthermore, if ν, $\gamma>0$, $f(x^*)=0$, and x^* minimizes $f_0(x)+\gamma|f(x)|^{1/\nu}$, then x^* minimizes $f_0(x)$ on X subject to $f(x)=0$.*

A global optimization procedure. We assume that, for $\nu=1, 2, \dots$, we have a procedure for finding a set $\mathcal{M}_\nu$ of points in X that minimize the function

$$x\to\varphi_\nu(x):=f_0(x)+\nu|f(x)|^{1/\nu}: X\to\mathbb{R}.$$

* Received by the editors January 14, 1991; accepted for publication (in revised form) September 10, 1991. This work was partially supported by National Science Foundation grant DMS 8619002.

† Department of Mathematics, Northeastern University, Boston, Massachusetts 02115.

We determine $\mathcal{M}_\nu$ for $\nu = 1, 2, \ldots$ until, for some integer μ and $\nu = \mu$, each point $x' \in \mathcal{M}_\nu$ yields $f(x') = 0$. At this point we terminate the iteration.

It follows from Theorem 1 that the iteration will terminate for some integer μ (not exceeding $\max\{N^*, c^*\} + 1$), and that each point $x' \in \mathcal{M}_\mu$ also yields the minimum of $f_0(x)$ subject to $f(x) = 0$. Furthermore, if each $\mathcal{M}_\nu$ contains all the minima of $\varphi_\nu(x)$, then $\mathcal{M}_\mu$ is the set of all the points in X that yield the constrained minimum.

COROLLARY. *Let $U \subset \mathbb{R}^n$ be open, $(f_0, f): U \to \mathbb{R}^m$ analytic, and $f(x_0) = 0$. Then $x_0 \in U$ minimizes $f_0(x)$ locally subject to $f(x) = 0$ if and only if x_0 minimizes $f_0(x) - 1/\log(|f(x)|)$ locally, where $1/\log 0$ is defined as 0. More precisely, there exists $r_1 > 0$ such that*

$$f_0(x_0) = \min\{f_0(x) \mid x \in \bar{B}(x_0, r_1), f(x) = 0\}$$

if and only if there exists $r_2 > 0$ such that

$$f_0(x_0) = \min\{f_0(x) - 1/\log(|f(x)|) | x \in \bar{B}(x_0, r_2)\}.$$

Remarks. (i) The assumption that (f_0, f) is analytic cannot be replaced by $(f_0, f) \in C^\infty$. Let

$$X = [-1, 1]^2, \quad f_0(x_1, x_2) = -x_1^2, \quad f(x_1, x_2) = e^{-1/x_1^6} + x_2^2.$$

Then, for any $c > 0$ and $N \in \{1, 2, \ldots\}$,

$$f_0(x_1, 0) = 1/[\log(f(x_1, 0))]^{1/3} < 1/\log(f(x_1, 0)) < -c[f(x_1, 0)]^{1/N}$$

for all sufficiently small $|x_1|$.

(ii) The assumption that X is compact cannot be dropped. Let

$$X = \mathbb{R}^2, \quad f_0(x_1, x_2) = x_1 x_2, \quad f(x_1, x_2) = x_1.$$

Then 0 minimizes $f_0(x_1, x_2)$ subject to $f(x_1, x_2) = 0$, but $(f_0, f)(\mathbb{R}^2)$ is dense in $\mathbb{R}^2$.

The proof of Theorem 1 follows from a generalization by Hironaka [1] of an inequality of Łojasiewicz. We refer to a subset of $\mathbb{R}^n$ as *subanalytic* if it is the image of a semi-analytic set under a proper analytic map.

INEQUALITY III (see [1, Ineq. III, p. 9.5]). *Let A and B be closed subanalytic subsets of $\mathbb{R}^n$ such that $A \cap B \neq \emptyset$. Then, for each compact subset K of $\mathbb{R}^n$, we can find $N \in \{1, 2, \ldots\}$ and $C > 0$ such that, for all $x \in K$,*

$$C(\operatorname{dist}(x, A) + \operatorname{dist}(x, B)) \geqq \operatorname{dist}(x, A \cap B)^N.$$

Proof of Theorem 1. Let

$$a := \inf\{f_0(x) | x \in X, f(x) = 0\},$$
$$\mathcal{F} = \{x \in X | f(x) = 0, f_0(x) = a\}.$$

Since X is compact, $0 \in f(X)$, and (f_0, f) is continuous, the set $\mathcal{F}$ is nonempty. Let

$$A = (f_0, f)(X), \quad B = (-\infty, a] \times \{0\} \subset \mathbb{R} \times \mathbb{R}^m, \quad K = [\min f_0(X), a] \times \{0\}.$$

Then A and B (which are images of compact semi-analytic sets under analytic maps) are closed subanalytic sets in $\mathbb{R} \times \mathbb{R}^m$ and $A \cap B = (a, 0)$. It follows, by Inequality III, that there exist $C > 0$ and $N' \in \{1, 2, \ldots\}$ such that

$$(*) \qquad C(\operatorname{dist}(y, A) + \operatorname{dist}(y, B)) \geqq \operatorname{dist}(y, A \cap B)^{N'} \quad \forall y \in K.$$

Now let $x \in X$ be such that $f_0(x) < a$, and let $y = (f_0(x), 0)$. Then

$$\operatorname{dist}(y, A) \leqq |f(x)|, \quad \operatorname{dist}(y, B) = 0, \quad \operatorname{dist}(y, A \cap B) = a - f_0(x),$$

and, by (*),

$$C|f(x)| \geqq C(\mathrm{dist}\,(y, A)+\mathrm{dist}\,(y, B)) \geqq \mathrm{dist}\,(y, A\cap B)^{N'} = [a-f_0(x)]^{N'}.$$

Let $x_1 \in \mathscr{F}$. It follows, setting $c' = C^{1/N'}$, that

$$f_0(x_1)-f_0(x) \leqq c'|f(x)|^{1/N'}$$

if $f_0(x) < f_0(x_1)$, and thus

$$(**) \qquad f_0(x)+c'|f(x)|^{1/N'} \geqq f_0(x_1) = f_0(x_1)+c'|f(x_1)|^{1/N'} \quad \forall\, x\in X.$$

Let

$$\psi_1(x) := f_0(x)+c'|f(x)|^{1/N'},$$

$M = 1+\max\,\{|f(x)|\,|\,x\in X\}$, $c^* = c'M^{1/N'}$, $N^* = N'$, $c \geqq c^*$, and $N \geqq N^*$. Then

$$c'|f(x)|^{1/N'} = c'M^{1/N'}|f(x)/M|^{1/N'} \leqq c|f(x)|^{1/N} \quad \forall\, x\in X.$$

It follows that

$$\psi_1(x) \leqq \psi(x) := f_0(x)+c|f(x)|^{1/N} \quad \forall\, x\in X.$$

Thus, by (**), each point $x_1 \in \mathscr{F}$ minimizes both $\psi_1(x)$ and $\psi(x)$ on X, proving the "only if" part of our assertion.

Now let x_0 minimize $\psi(x)$ on X, and let $x_1 \in \mathscr{F}$. Since x_1 also minimizes $\psi(x)$, we have

$$f_0(x_1) = \psi(x_1) = \psi(x_0) = f_0(x_0)+c|f(x_0)|^{1/N}.$$

Now x_1 minimizes $\psi_1(x)$ and therefore

$$f_0(x_1) = \psi_1(x_1) \leqq \psi_1(x_0) \leqq \psi(x_0) = f_0(x_1);$$

hence $f_0(x_1) = \psi_1(x_0)$ and therefore

$$f_0(x_1) = f_0(x_0)+c'|f(x_0)|^{1/N} = f_0(x_0)+c|f(x_0)|^{1/N}.$$

If $f(X) = \{0\}$, then our theorem is trivially satisfied. Otherwise, $M > 1$ and therefore $c > c'$. It follows then from the last relation that $f(x_0) = 0$ and $f_0(x_0) = f_0(x_1)$. Thus x_0 minimizes $f_0(x)$ on X subject to $f(x) = 0$.

Finally, assume that $\nu, \gamma > 0$, $f(x^*) = 0$, and x^* minimizes $\varphi(x) := f_0(x)+\gamma|f(x)|^{1/\nu}$ on X. Then, if $x' \in X$ and $f(x') = 0$, we have

$$f_0(x^*) = \varphi(x^*) \leqq \varphi(x') = f_0(x').$$

Thus x^* minimizes $f_0(x)$ subject to $f(x) = 0$. □

Proof of the corollary. The "if" part is obvious. Now assume that there exists $r_1 > 0$ such that

$$f_0(x_0) = \min\,\{f_0(x)\,|\,x\in \bar{B}(x_0, r_1), f(x) = 0\}.$$

Then, by Theorem 1, there exist $c > 0$ and $N \in \{1, 2, \dots\}$ such that

$$f_0(x)+c|f(x)|^{1/N} \geqq f_0(x_0) \quad \forall x\in X := \bar{B}(x_0, r_1).$$

Since $\lim_{y\to 0}|y|^{1/N}\log(|y|) = 0$ and f is continuous, we may find r', $r_2 > 0$ such that $-c|y|^{1/N} \geqq 1/\log(|y|)$ if $|y| \leqq r'$, and $|f(x)| \leqq r'$ if $x \in \bar{B}(x_0, r_2)$. Then

$$f_0(x)-1/\log(|f(x)|) \geqq f_0(x_0) = f_0(x_0)-1/\log(|f(x_0)|) \quad \forall\, x\in \bar{B}(x_0, r_2). \qquad \square$$

REFERENCE

[1] H. Hironaka, *Introduction to real-analytic sets and real-analytic maps*, Istituto Matematico "L. Tonelli" dell'Universitá de Pisa, Italy, 1973.

SIAM J. OPTIMIZATION
Vol. 2, No. 4, pp. 668-672, November 1992

DIAGONAL MATRIX SCALING AND LINEAR PROGRAMMING*

LEONID KHACHIYAN† AND BAHMAN KALANTARI‡

Abstract. A positive semidefinite symmetric matrix either has a nontrivial nonnegative zero or can be scaled by a positive diagonal matrix into a doubly quasi-stochastic matrix. This paper describes a simple path-following Newton algorithm of the complexity $O(\sqrt{n}\,L)$ iterations to either scale an $n\times n$ matrix or give a nontrivial nonnegative zero. The latter problem is well known to be equivalent to linear programming.

Key words. diagonal matrix scaling, linear programming, path-following Newton's methods

AMS(MOS) subject classification. 90C05

1. Introduction.

1.1. Scaling. We consider the following problem of diagonal matrix scaling: Given an $n\times n$ symmetric positive semidefinite matrix A, either find a positive diagonal matrix X, which scales A into a doubly quasi-stochastic matrix

$$XAXe=e, \qquad X=\operatorname{diag}(x_1,\dots,x_n)>0,$$

or prove that A is not scalable. Here $e=(1,\dots,1)^T\in\mathbb{R}^n$.

Letting $x=(x_1,\dots,x_n)^T$, $x^{-1}=(1/x_1,\dots,1/x_n)^T$, the problem can be written as

$$(1.1)\qquad Ax-x^{-1}=0, \qquad x>0.$$

1.2. Bounds on solutions. Scaling and linear programming. Let

$$(1.2)\qquad \mu=\min\{x^TAx\,|\,x\in S_+\},$$

where $S_+=\{x\in\mathbb{R}^n\,|\,x\geqq 0,\ \|x\|=(x_1^2+\cdots+x_n^2)^{1/2}=1\}$ is the intersection of the unit n-dimensional Euclidean sphere with the nonnegative orthant. It is known [1], [2] that

(1.3) *a positive semidefinite matrix A is scalable if and only if $\mu>0$.*

In fact, as we prove in the Appendix, the "only if" part of this statement can be strengthened as follows: *If $x>0$ scales a positive semidefinite matrix A (see (1.1)) then*

$$(1.4)\qquad n^{-1}\leqq \|x\|^2\mu\leqq n.$$

The "if" part will follow from the algorithm to be described.

Without loss of generality we assume henceforth that $n\geqq 4$ and

$$(1.5)\qquad \|Ae\|^2\leqq n.$$

In §§ 2 and 3 we describe a path-following Newton's method for solving the scaling problem (1.1). The method generates a sequence of positive points $\hat{x}_0,\dots,\hat{x}_k$ with the following properties.

* Received by the editors June 4, 1990; accepted for publication (in revised form) September 12, 1991.

† Department of Computer Science, Rutgers University, New Brunswick, New Jersey and Computing Center, Russian Academy of Sciences, 117418 Moscow, Russia. This research was performed while the author was visiting the Department of Computer Science and the Center for Discrete Mathematics and Theoretical Computer Science (DIMACS), Rutgers University, New Brunswick, New Jersey 08903.

‡ Department of Computer Science, Rutgers University, New Brunswick, New Jersey 08903. This research was supported in part by National Science Foundation grant CCR-88-07518.

If A is scalable, then in at most $k=(4\sqrt{n}+1)\ln(10n^2/\mu)$ iterations we determine the scalability of A and in an additional s iterations scale A with accuracy 2^{s-2} digits:

$$\|\hat{X}_{k+s}A\hat{X}_{k+s}e-e\| \leqq (\tfrac{3}{4})^{2^s}. \tag{1.6}$$

If A is not scalable, the sequence of projected points $\hat{x}_0/\|\hat{x}_0\|, \ldots, \hat{x}_k/\|\hat{x}_k\| \in S_+$ converges to a zero of the quadratic form x^TAx over S_+:

$$\left(\frac{\hat{x}_k}{\|\hat{x}_k\|}\right)^T A\left(\frac{\hat{x}_k}{\|\hat{x}_k\|}\right) \leqq 10n^2 \exp\left(-\frac{k}{4\sqrt{n}+1}\right). \tag{1.7}$$

As usual, if the entries of A are rational and $\mu>0$, then $\mu \geqq 2^{-L}$ where L is the binary length of the input. So the number of Newton's iterations of the method does not exceed $O(\sqrt{n}\,L)$. The computational cost of each iteration is $O(n(\operatorname{rank} A)^2)$ arithmetic operations.

Observe that the problem of computing a zero of a positive semidefinite quadratic form over S_+ is well known to be equivalent to the general linear programming problem (see, e.g., [1]). Therefore, the scaling problem (1.1) is somewhat more general than linear programming, and for the latter problem we obtain the same bound as in [5].

2. Newton's system and region of quadratic convergence. Instead of (1.1), it is convenient to consider a slightly more general problem of computing a positive zero of the mapping

$$F(x)=b+Ax-x^{-1},$$

where b is a fixed n-dimensional vector. Since $F(x+y)=F(x)+Ay+X^{-2}y+$higher order terms in y, the following linear system

$$(X^{-2}+A)y=x^{-1}-Ax-b \tag{2.1}$$

defines Newton's vector y for a point $x>0$. Multiplying this system by X, we get

$$(E+XAX)X^{-1}y=e-XAXe-Xb, \tag{2.2}$$

where $E=\operatorname{diag}(e)$ is the identity matrix. Letting

$$z=X^{-1}y, \quad A_x=XAX, \quad b_x=Xb, \tag{2.3}$$

(2.2) can be written as

$$(E+A_x)z=e-A_xe-b_x. \tag{2.4}$$

Comparing (2.4) to the system (2.1) written at $x=e$, we see that (2.4) coincides with the latter system after the transformation (2.3).

Let

$$x'=\text{Newton}\,(F,x)=x+y=X(e+z)$$

be the vector obtained as a result of one Newton's iteration at a point $x>0$. To guarantee the positiveness of x' it suffices to require $\|z\|<1$. On the other hand, since A_x is positive semidefinite, it follows from (2.4) that

$$\|z\|=\|(E+A_x)^{-1}(e-A_xe-b_x)\| \leqq \|e-A_xe-b_x\|. \tag{2.5}$$

Therefore, the *conditions*

$$x>0, \qquad \|e-A_xe-b_x\|<1 \tag{2.6}$$

imply that $x'=$ Newton (F,x) is positive.

It is also easy to see that *under the assumption* (2.6) *each Newton's iteration quadratically decreases the norm of the right-hand side of* (2.4):

$$\|e-A_{x'}e-b_{x'}\| \leqq \|e-A_xe-b_x\|^2. \tag{2.7}$$

Indeed, letting $Z=\operatorname{diag}(z)$, we have

$$X'=\operatorname{diag}(X(e+z))=(E+Z)X=X(E+Z),$$

$$A_{x'}e=X'AX'e=(E+Z)A_x(E+Z)e=(E+Z)A_x(e+z).$$

From (2.4), $A_x(e+z)=e-z-b_x$. Thus

$$A_{x'}e=(E+Z)(e-z-b_x)=e-Zz-(E+Z)b_x=e-Zz-b_{x'}.$$

Hence

$$\|e-A_{x'}e-b_{x'}\|=\|Zz\|=(z_1^4+\cdots+z_n^4)^{1/2}=\|z\|_4^2.$$

But $\|z\|_4 \leqq \|z\|_2=\|z\|$, so that (2.7) follows from (2.5).

It follows from (2.5) and (2.7) that for an arbitrary starting point x_0 from the region (2.6) the norm of Newton's vectors $z_0, z_1, \ldots, z_k$ quadratically converges to 0. Hence the sequence of Newton's iterates $x_k=(E+Z_k)(E+Z_{k-1})\cdots(E+Z_0)x_0$ quadratically converges. Clearly, the limit x of this sequence is nonnegative and satisfies the system of equations $XAXe+Xb=e$. Now it is easily seen from this system that all the components of x are positive. Thus, *in the region* $x>0$, $\|A_xe+b_x-e\|<1$, *Newton's method is well defined and quadratically converges to a positive zero of the system of equations* $b+Ax-x^{-1}=0$.

In particular, *if the region* $x>0$, $\|A_xe-e\|<1$, *is nonempty, a positive semidefinite matrix A can be scaled into a doubly quasi-stochastic matrix.*

3. Path-following method for scaling.

Consider the family of mappings

$$F_t(x)=tb+tAx-x^{-1}, \qquad x>0,$$

defined for each value of t from 1 to 0. Set

$$t_0=1, \quad x_0=e \quad \text{and} \quad b=e-Ae, \quad \delta=\tfrac{1}{2}.$$

Clearly,

$$\|t_kX_kAX_ke+t_kX_kb-e\| \leqq \delta \tag{3.1}$$

for $k=0$, since in this case the left-hand side of (3.1) equals zero. Suppose that (3.1) holds for some $k \geqq 0$ and $t_k \in (0,1]$. By (2.7), we can apply Newton's method to F_{t_k} at x_k, and in one iteration compute a new positive vector x_{k+1} satisfying

$$\|t_kX_{k+1}AX_{k+1}e+t_kX_{k+1}b-e\| \leqq \delta^2. \tag{3.2}$$

Now we want to decrease t and obtain (3.1) for $k+1$ with a smaller value $t_{k+1}=t_k-\nu_k$. This can be done if

$$\nu_k\|X_{k+1}AX_{k+1}e+X_{k+1}b\| \leqq \delta-\delta^2.$$

But (3.2) implies

$$\|X_{k+1}AX_{k+1}e+X_{k+1}b\| \leqq \frac{\|e\|+\delta^2}{t_k},$$

so we can put

$$t_{k+1}=t_k\left(1-\frac{\delta-\delta^2}{\|e\|+\delta^2}\right)=t_k\left(1-\frac{1}{4\sqrt{n}+1}\right).$$

The last equality follows from $\delta=\frac{1}{2}$. Letting

$$\hat{x}_k=t_k^{1/2}x_{k+1},$$

we obtain a sequence of positive points $\hat{x}_0, \hat{x}_1, \dots, \hat{x}_k$ satisfying the condition (3.2)

$$\|\hat{X}_kA\hat{X}_ke+t_k^{1/2}\hat{X}_kb-e\|\leqq\delta^2=\tfrac{1}{4} \tag{3.3}$$

with

$$t_k\leqq\exp\left(-\frac{k}{4\sqrt{n}+1}\right). \tag{3.4}$$

At each iteration of the path-following method we check the condition

$$\|\hat{X}_kA\hat{X}_ke-e\|\leqq\tfrac{3}{4}. \tag{3.5}$$

If (3.5) holds, we conclude that A is scalable. In this case, using $\hat{x}_k$ as a starting point for solving with Newton's method $F(x)=Ax-x^{-1}=0$, we can scale A in s iterations with accuracy (1.6). This completes the description of the method.

4. Convergence. Suppose that (3.5) does not hold, i.e., we cannot scale A until at least the kth iteration. Then it follows from (3.3) that $t_k^{1/2}\|\hat{X}_kb\|\geqq\frac{1}{2}$, and so the norm of $\hat{x}_k$ is large:

$$\|\hat{x}_k\|\geqq(2\|b\|t_k^{1/2})^{-1}. \tag{4.1}$$

Let

$$\xi_k=\hat{X}_kA\hat{X}_ke+t_k^{1/2}\hat{X}_kb-e.$$

From (3.3) we know that $\|\xi_k\|\leqq\frac{1}{4}$. Hence

$$e^T\xi_k=\hat{x}_k^TA\hat{x}_k+t_k^{1/2}\hat{x}_k^Tb-n\leqq\|e\|\,\|\xi_k\|\leqq\frac{\sqrt{n}}{4}. \tag{4.2}$$

This implies (1.7):

$$\left(\frac{\hat{x}_k}{\|\hat{x}_k\|}\right)^TA\left(\frac{\hat{x}_k}{\|\hat{x}_k\|}\right)\leqq\frac{n+\sqrt{n}/4}{\|\hat{x}_k\|^2}+\frac{t_k^{1/2}\|b\|}{\|\hat{x}_k\|}\leqq(4n+\sqrt{n}+2)\|b\|^2t_k$$
$$\leqq 5n\|b\|^2t_k=5n\|e-Ae\|^2t_k\leqq10n^2t_k\leqq10n^2\exp\left(-\frac{k}{4\sqrt{n}+1}\right).$$

The above inequalities follow from (4.2), (4.1), the assumption that $n\geqq4$, (1.5), and (3.4), respectively. Thus, if $\mu=\min\{x^TAx\,|\,x\in S_+\}>0$, in at most $(4\sqrt{n}+1)\ln(10n^2/\mu)$ iterations we obtain (3.4) and scale A.

5. Concluding comments. The scaling problem (1.1) for matrices with nonnegative entries has been a problem of interest since at least the early sixties (see, e.g., [6]). The relevance of positive semidefinite matrix scaling to linear programming via the characterization (1.3) was given in [1] and [2]. As shown above, the posititive semidefinite case provides a convenient elementary format for describing an $O(\sqrt{n}\,L)$ Newton iteration method for linear programming.

As was observed in [3], any solution to the scaling probem (1.1) is a stationary point of the logarithmic barrier function $f(x)=\frac{1}{2}x^TAx-\sum_{i=1}^n \ln x_i$. An alternative way to scale a positive semidefinite matrix is to apply Newton's method (say, with line search) directly to minimize $f(x)$. In this case it is easy to show (and it also follows from the one-self-concordance of $f(x)$; see [4, § 1]) that outside the region of quadratic convergence each Newton's iteration decreases f by a constant factor. If $f(x)$ attains its minimum value f_* over $x \geqq 0$, then A is scalable and (1.4), (1.5) imply $f(e)-f_* \leqq n \ln(1/\mu)=O(nL)$. In this case, starting the iterations from e, we enter the region of quadratic convergence in $O(nL)$ iterations. Otherwise, from $2 \min\{f(tx) \mid t \in (0,\infty)\} = n \ln k(x)+n(1-\ln n)$, where $k(x)=x^TAx/(x_1 \cdots x_n)^{2/n}$ is Karmarkar's potential function, we get $k(x)/n \leqq \exp(-1+2f(x)/n)$. Since $f(x)$ is reduced by a constant factor at each iteration and $x^TAx/\|x\|^2 \leqq k(x)/n$, it follows that, by projecting the iterates of such a potential reduction method onto the unit sphere, we can compute in $O(nL)$ iterations a zero of x^TAx over S_+ with an accuracy of 2^{-L}.

Appendix. Suppose that $x>0$ scales A (see (1.1)). Multiplying (1.1) by x^T, we get $x^TAx=x^Tx^{-1}=n$. Since $x/\|x\| \in S_+$, the second of the inequalities (1.4) follows from the definition (1.2)

$$\mu \leqq \left(\frac{x}{\|x\|}\right)^T A\left(\frac{x}{\|x\|}\right)=\frac{n}{\|x\|^2}.$$

To prove the first of the inequalities (1.4), observe that if B is a positive semidefinite and doubly quasi-stochastic matrix, then from $Be=e$ it follows that the minimum of the quadratic form y^TBy over the simplex $\{y \in \mathbb{R}^n \mid y \geqq 0,\ e^Ty=n\}$ is attained at the point $y=e$ and equals n. Hence $\min\{y^TBy \mid y \in S_+\} \geqq n^{-1}$. Applying the last inequality to the doubly quasi-stochastic matrix $B=XAX$, we get

$$n^{-1} \leqq \min\{y^TXAXy \mid y \in S_+\} \leqq \|x\|_\infty^2 \min\{y^TAy \mid y \in S_+\} = \|x\|_\infty^2 \mu \leqq \|x\|^2 \mu.$$

REFERENCES

[1] B. KALANTARI (1990), *Canonical problems for quadratic programming and projective methods for their solution*, Proc. AMS Conference on Mathematical Problems Arising from Linear Programming, 1988, in Contemp. Math., Vol. 114, pp. 243-263.

[2] ——— (1989), *Derivation of a generalized and strengthened Gordan theorem from generalized Karmarkar potential and logarithmic barrier functions*, Tech. Report LCSR-TR-121, Dept. of Computer Science, Rutgers Univ., New Brunswick, New Jersey.

[3] A. W. MARSHALL AND I. OLKIN (1968), *Scaling of matrices to achieve specified row and column sums*, Numer. Math., 12, pp. 83-90.

[4] JU. E. NESTEROV AND A. S. NEMIROVSKY (1989), *Self-concordant functions and polynomial-time methods in convex programming*, Central Economic and Math. Institute, USSR Academy of Sciences, Moscow, Russia.

[5] J. RENEGAR (1988), *A polynomial-time algorithm based on Newton's method for linear programming*, Math. Programming, 40, pp. 59-93.

[6] R. SINKHORN (1964), *A relationship between arbitrary positive matrices and doubly stochastic matrices*, Ann. Math. Statist., 35, pp. 876-879.