SIAM J. OPTIM.
Vol. 14, No. 1, pp. 1–34

ADAPTIVE APPROXIMATIONS AND EXACT PENALIZATION FOR THE SOLUTION OF GENERALIZED SEMI-INFINITE MIN-MAX PROBLEMS*

J. O. ROYSET†, E. POLAK†, AND A. DER KIUREGHIAN†

Abstract. We develop an implementable algorithm for the solution of a class of generalized semi-infinite min-max problems. To this end, first we use exact penalties to convert a generalized semi-infinite min-max problem into a finite family of semi-infinite min-max-min problems. Second, the inner min-function is smoothed and the semi-infinite max part is approximated, using discretization, to obtain a three-parameter family of finite min-max problems. Under a calmness assumption, we show that when the penalty is sufficiently large the semi-infinite min-max-min problems have the same solutions as the original problem, and that when the smoothing and discretization parameters go to infinity the solutions of the finite min-max problems converge to solutions of the original problem, provided the penalty parameter is sufficiently large.

Our algorithm combines tests for adjusting the penalty, the smoothing, and the discretization parameters and makes use of a min-max algorithm as a subroutine. In effect, the min-max algorithm is applied to a sequence of gradually better-approximating min-max problems, with the penalty parameter eventually stopping to increase, but with the smoothing and discretization parameters driven to infinity. A numerical example demonstrates the viability of the algorithm.

Key words. generalized minimax, semi-infinite optimization, nonsmooth optimization algorithms

AMS subject classifications. 49K35, 49M30, 90C34

PII. S1052623402406777

1. Introduction.

We consider the class of generalized semi-infinite min-max problems in the form

$$\mathbf{P} \qquad \min_{x\in\mathbb{R}^n} \psi(x), \tag{1.1}$$

where $\psi : \mathbb{R}^n \to \mathbb{R}$ is defined by

$$\psi(x) \stackrel{\triangle}{=} \max_{y\in Y}\{\phi(x,y) \mid f(x,y) \le 0\}, \tag{1.2}$$

with $\phi : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$, $f : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^{r_1}$, $Y \stackrel{\triangle}{=} \{y \in \mathbb{R}^m | g(y) \le 0\}$, $g : \mathbb{R}^m \to \mathbb{R}^{r_2}$, and $v \le 0$ meaning $v^1 \le 0, \ldots, v^q \le 0$, for any $v = (v^1, \ldots, v^q) \in \mathbb{R}^q$. We use superscripts to denote components of vectors.

This class of generalized semi-infinite min-max problems is of both theoretical and practical interest. In particular, generalized semi-infinite min-max problems occur in various engineering applications. For example, optimal design of civil and aerospace structures is frequently considered in a probabilistic framework, where uncertainties in material properties, loads, and boundary conditions are taken into account. Let $x \in \mathbb{R}^n$ be a vector of deterministic design variables, e.g., physical dimensions of the

*Received by the editors May 1, 2002; accepted for publication (in revised form) November 11, 2002; published electronically May 15, 2003. This work was partially supported by the National Science Foundation under grant ECS-9900985.

http://www.siam.org/journals/siopt/14-1/40677.html

†University of California, Berkeley, CA 94720 (joroyset@ce.berkeley.edu, polak@eecs.berkeley.edu, adk@ce.berkeley.edu). The first author was supported by the Norwegian Research Council and the Space Science Laboratory, University of California, Berkeley.

structure, or parameters in the probability distribution of the random quantities. The probability of failure of a structure $P_f : \mathbb{R}^n \to [0,1]$ for a given design vector x, is defined by (see [5])

$$P_f(x) \stackrel{\triangle}{=} \int_{\{y \in \mathbb{R}^m | h(x,y) \leq 0\}} \varphi(y) dy, \tag{1.3}$$

where $\varphi(\cdot)$ is the m-dimensional standard normal probability density function,[1] and $h : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ is a smooth real-valued limit-state function.

The optimal design problem is typically in the form

$$\min_{x \in \mathbb{R}^n} \{\ c^0(x) + c^1(x) P_f(x)\ \}, \tag{1.4}$$

where $c^0 : \mathbb{R}^n \to \mathbb{R}$ is the initial cost of the structure and $c^1 : \mathbb{R}^n \to \mathbb{R}$ is the cost of structural failure. The evaluation of $P_f(\cdot)$ is computationally expensive, and the mathematical properties of $P_f(\cdot)$ are not easily available. Hence, a first-order approximation to the probability of failure is usually considered acceptable. Based on such approximations, it can be shown (see [22]) that (1.4) can be approximated by

$$\min_{x \in \mathbb{R}^n} \max_{y \in \Gamma(x)} \left\{ c^0(x) + c^1(x) \Phi \left[\frac{\beta h(x,0)}{h(x,y) - h(x,0)} \right] \right\}, \tag{1.5}$$

where $\Phi(\cdot)$ is the standard normal cumulative distribution function and $\Gamma(x) = \{y \in \mathbb{R}^m \mid h(x,y) - h(x,0) \leq -\alpha, \|y\|^2 \leq \beta^2\}$, with $\alpha, \beta > 0$. Hence, the optimal design problem (1.4) can approximately be solved by solving a generalized semi-infinite min-max problem in the form (1.1), with $f(x,y) = h(x,y) - h(x,0) + \alpha$ and $g(y) = \|y\|^2 - \beta^2$.

There is a nontrivial literature dealing with the existence of and formulas for directional derivatives of generalized max-functions, such as the one in (1.2) (e.g., [2, 21]), and with first-order optimality conditions for generalized semi-infinite optimization problems of the form

$$\min_{x \in \mathbb{R}^n} \{f^0(x) \mid \psi(x) \leq 0\}, \tag{1.6}$$

where $f^0 : \mathbb{R}^n \to \mathbb{R}$ is smooth and $\psi(\cdot)$ is as in (1.2) [9, 23, 24, 25, 26, 28]. Just as we do in Assumption 2.6 below, [28] also assumes that the linear independence constraint qualification for the "inner problem" (1.2) is satisfied. Under this assumption, [28] shows that the problem in (1.6) is equivalent to a standard semi-infinite optimization problem, i.e., a problem in the form $\min_{x \in \mathbb{R}^n} \{f^0(x) \mid \phi(x,\omega) \leq 0,\ \omega \in \Omega\}$, with $\phi(\cdot,\cdot)$ smooth and Ω of infinite cardinality. However, it is not clear how to implement a procedure for constructing the equivalent problem.

There are only a few papers dealing with numerical methods for problems in the form (1.6). In [7], an algorithm is presented, without a convergence proof, for a special class of problems, with $f^0(x) = x \in \mathbb{R}$, arising in the evaluation of the acceleration radius of manipulator positioning systems. Other basic ideas for solving problems of the form (1.6) in robotics (maneuverability problems) can be found in [8]. A special case of (1.6) arising in robotics and minimum time optimal control problems is considered in [10], where $y \in \mathbb{R}, m = 1$. In [13], we find an algorithm for the

[1] When the uncertainties are *not* described by standard normal random variables, such variables can always be obtained by a nonlinear transformation; see [5].

solution of the special case with $\phi(x,y) = \frac{1}{2}\langle y, Gy\rangle + \langle a, y\rangle + \langle y, Hx\rangle$, G, H matrices, $f^k(x,y) = \langle p^k, y\rangle + q^k(x)$, $a, p^k \in \mathbb{R}^m$, and convex functions q^k. In [26, 27] we find a conceptual algorithm for solving the problem (1.6). In these papers it is assumed that the LICQ, second-order sufficient conditions, and strict complementary slackness for the "inner problem" in (1.2) hold. The algorithm in [26, 27] applies a globally convergent Newton-type method to the Karush–Kuhn–Tucker system for a locally reduced problem. In addition, a conceptual algorithm, based on discretization, is presented in [27]. In the still unpublished paper [12], Levitin employs a differentiable penalty function to remove the constraints $f(x,y) \leq 0$, and shows that the sequence of global solutions of the penalized problem converges to a global solution of (1.6), as the penalty goes to infinity. Thus, in spirit, his approach is close to ours. To the authors' knowledge there exists no implementable algorithm for solving general forms of **P**.

In this paper we present an implementable algorithm for solving general forms of **P** under a calmness assumption. We use an exact penalty function to eliminate the inequalities in (1.2) that depend on x, i.e., $f(x,y) \leq 0$, and as a result convert the generalized semi-infinite min-max problem into a standard semi-infinite min-max problem with an unknown penalty parameter. In principle, we could have picked any one of the existing exact penalty or augmented Lagrangian functions for this purpose; see, e.g., [16, 17]. However, the use of augmented Lagrangians, together with differentiable multiplier estimates as in [6], is unattractive because it would require a second-order sufficient condition to hold at solutions of the "inner problem" in (1.2), evaluation of second-order derivatives even by a first-order algorithm, and the linear independence assumption on the gradients $\nabla_y f^k(x,y)$ and $\nabla g^k(y)$ at every $x \in \mathbb{R}^n$ and $y \in Y$. Hence, we opted for a standard nondifferentiable exact penalty function, which avoids the need for an assumption about a second-order sufficient condition and second-order derivative evaluations and requires only the linear independence assumption on the gradients $\nabla_y f^k(x,y)$ and $\nabla g^k(y)$ at points $y \in Y$, which are solutions to the "inner problem" (1.2). The selected approach leads to an algorithm that generates sequences converging to weaker stationary points than the ones given in [24]; see Appendix A. It is unknown whether a different penalty function would have resulted in an algorithm converging to stronger stationary points.

Since a penalty function of the form $\phi(x,y) - \pi\|f(x,y)_+\|_\infty$ is in fact a min-function, use of a nondifferentiable exact penalty function results in a semi-infinite min-max-min problem with an unknown penalty parameter. This problem can be approximated by a finite min-max problem obtained by discretizing the semi-infinite part and smoothing the min-function. This adds two more parameters to the resulting min-max problem. In view of this, our algorithm combines tests for adjusting the three parameters with the Pironneau–Polak–Pshenichnyi min-max algorithm [17, 20]. Under mild assumptions, we show that if the algorithm generates a bounded sequence, then the penalty parameter remains bounded and that there exists an accumulation point which satisfies a first-order optimality condition.

Along the way, we needed a few results from [19], such as a new optimality condition for min-max-min problems and tests for adjusting discretization and smoothing parameters. For completeness, we have duplicated those results.

In section 2 we define the penalized problem and establish its relation to **P**. In the process we obtain a new first-order optimality condition for **P**. Approximations for the solution of the penalized problem are defined in section 3. Section 4 presents the algorithm and the proof of its convergence. The paper ends with a numerical example and concluding remarks.

2. Exact penalization. As described in the introduction, we introduce exact penalization for the violation of the constraints $f(x,y) \leq 0$ in (1.2). Let π denote this penalty. Hence, for any $\pi > 0$ we define a family of related problems by

$$\mathbf{P}_\pi \qquad \min_{x \in \mathbb{R}^n} \psi_\pi(x), \tag{2.1}$$

where $\psi_\pi : \mathbb{R}^n \to \mathbb{R}$ is defined by

$$\psi_\pi(x) \stackrel{\triangle}{=} \max_{y \in Y} \{\phi(x,y) - \pi \|f(x,y)_+\|_\infty\}, \tag{2.2}$$

with $\|v_+\|_\infty \stackrel{\triangle}{=} \max\{\max\{v^1, 0\}, \ldots, \max\{v^q, 0\}\}$.

At first glance (2.1) looks like an ordinary min-max problem. However, $\|f(x,y)_+\|_\infty$ is a max-function, and hence we see that, with $\mathbf{r} \stackrel{\triangle}{=} \{1, \ldots, r\}$ and $r \stackrel{\triangle}{=} r_1 + 1$,

$$\psi_\pi(x) = \max_{y \in Y} \{\phi(x,y) - \pi \|f(x,y)_+\|_\infty\} = \max_{y \in Y} \min_{k \in \mathbf{r}} \phi_\pi^k(x,y), \tag{2.3a}$$

where

$$\phi_\pi^k(x,y) \stackrel{\triangle}{=} \phi(x,y) - \pi f^k(x,y), \quad k \in \mathbf{r_1} \stackrel{\triangle}{=} \{1, \ldots, r_1\}, \tag{2.3b}$$

$$\phi_\pi^r(x,y) \stackrel{\triangle}{=} \phi(x,y). \tag{2.3c}$$

We need the following notation: Let $\mathbb{B}(x,\rho) \stackrel{\triangle}{=} \{x' \in \mathbb{R}^n | \|x - x'\| \leq \rho\}$, and let $\omega_\pi : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$, $\hat{Y} : \mathbb{R}^n \to 2^{\mathbb{R}^m}$ and $\hat{Y}_\pi : \mathbb{R}^n \to 2^{\mathbb{R}^m}$ be defined by

$$\omega_\pi(x,y) \stackrel{\triangle}{=} \min_{k \in \mathbf{r}} \phi_\pi^k(x,y), \tag{2.3d}$$

$$\hat{Y}(x) \stackrel{\triangle}{=} \arg\max_{y \in Y} \{\phi(x,y) \mid f(x,y) \leq 0\}, \tag{2.4a}$$

$$\hat{Y}_\pi(x) \stackrel{\triangle}{=} \arg\max_{y \in Y} \{\phi(x,y) - \pi \|f(x,y)_+\|_\infty\} = \{y \in Y \mid \omega_\pi(x,y) = \psi_\pi(x)\}. \tag{2.4b}$$

Note that (2.3a,d) imply that $\psi_\pi(x) = \max_{y \in Y} \omega_\pi(x,y)$.

Assumption 2.1. *We assume that*

(i) $\phi(\cdot,\cdot), f^k(\cdot,\cdot), k \in \mathbf{r_1} = \{1, \ldots, r_1\}$, *and* $g^k(\cdot), k \in \mathbf{r_2} \stackrel{\triangle}{=} \{1, \ldots, r_2\}$, *are continuously differentiable, and that*

(ii) $Y \subset \mathbb{R}^m$ *is compact, and that* $\{y \in Y | f(x,y) \leq 0\} \neq \emptyset$ *for all* $x \in \mathbb{R}^n$. ∎

The notion of calmness (see [4, 3]) can be used to show the local equivalence of $\mathbf{P}_\pi$ and $\mathbf{P}$ for π sufficiently large. For any $x \in \mathbb{R}^n$ and $u \in \mathbb{R}^{r_1}$, consider the perturbed "inner problem" (see (1.2)) defined by

$$\mathbf{IP}(x,u) \qquad \max_{y \in Y} \{\phi(x,y) \mid f(x,y) \leq u\}. \tag{2.5}$$

Let the value function $v : \mathbb{R}^n \times \mathbb{R}^{r_1} \to \mathbb{R} \bigcup \{-\infty\}$ of $\mathbf{IP}(x,u)$ be defined by

$$v(x,u) \stackrel{\triangle}{=} \max_{y \in Y} \{\phi(x,y) \mid f(x,y) \leq u\}, \tag{2.6}$$

where $v(x,u) = -\infty$ if $f(x,y) > u$ for all $y \in Y$.

We now define local calmness. A sufficient condition for local calmness will be given at the end of the section.

DEFINITION 2.2. *We say that* $\mathbf{IP}(\hat{x},0)$ *is* locally calm *at* $\hat{x}\in\mathbb{R}^n$ *if there exist* $\hat{\rho}>0$ *and* $\hat{\alpha}<\infty$ *such that*

$$v(x,u)-v(x,0)\le\hat{\alpha}\|u\|_\infty \tag{2.7}$$

for every $x\in\mathbb{B}(\hat{x},\hat{\rho})$ *and* $u\in\mathbb{R}^{r_1}$. ∎

THEOREM 2.3. *Suppose that Assumption* 2.1 *holds and that* $\mathbf{IP}(\hat{x},0)$ *is locally calm at* $\hat{x}\in\mathbb{R}^n$. *Then there exist a* $\hat{\pi}<\infty$ *and a* $\hat{\rho}>0$ *such that* $\psi(x)=\psi_{\hat{\pi}}(x)$ *for all* $x\in\mathbb{B}(\hat{x},\hat{\rho})$, *and hence* $\hat{x}$ *is a local minimizer for* $\mathbf{P}$ *if and only if* $\hat{x}$ *is a local minimizer for* $\mathbf{P}_{\hat{\pi}}$.

Proof. Let $\hat{\rho}>0$ and $\hat{\alpha}<\infty$ be as in Definition 2.2. Now, let $x\in\mathbb{B}(\hat{x},\hat{\rho})$ and $y\in\hat{Y}(x)$ be arbitrary. We will show that $y\in\hat{Y}_{\hat{\pi}}(x)$, with $\hat{\pi}=\hat{\alpha}$. For the sake of a contradiction, suppose that $y\notin\hat{Y}_{\hat{\pi}}(x)$. Then there exists $y'\in Y$ such that

$$\phi(x,y')-\hat{\pi}\|f(x,y')_+\|_\infty>\phi(x,y)-\hat{\pi}\|f(x,y)_+\|_\infty. \tag{2.8a}$$

Hence,

$$\begin{aligned}\phi(x,y')-\phi(x,y)&>\hat{\pi}\|f(x,y')_+\|_\infty-\hat{\pi}\|f(x,y)_+\|_\infty\\&=\hat{\pi}\|f(x,y')_+\|_\infty.\end{aligned} \tag{2.8b}$$

Next, $\phi(x,y')\le v(x,f(x,y')_+)$ and $\phi(x,y)=v(x,0)$. Hence, by (2.7)

$$\begin{aligned}\phi(x,y')-\phi(x,y)&\le v(x,f(x,y')_+)-v(x,0)\\&\le\hat{\pi}\|f(x,y')_+\|_\infty,\end{aligned} \tag{2.8c}$$

which is a contradiction. Hence, $y\in\hat{Y}(x)$, $y\in\hat{Y}_{\hat{\pi}}(x)$, and

$$\begin{aligned}\psi_{\hat{\pi}}(x)&=\phi(x,y)-\hat{\pi}\|f(x,y)_+\|_\infty\\&=\psi(x).\end{aligned} \tag{2.8d}$$

Hence, $\psi(x)=\psi_{\hat{\pi}}(x)$ for all $x\in\mathbb{B}(\hat{x},\hat{\rho})$, and the result follows. ∎

Optimality conditions can be expressed in terms of continuous, nonpositive valued optimality functions, which vanish at local minimizers; see [17].

THEOREM 2.4. *Suppose that Assumption* 2.1 *holds, and for any* $\pi>0$ *let* $\theta_\pi:\mathbb{R}^n\to\mathbb{R}$ *be an optimality function defined by*

$$\theta_\pi(x)\stackrel{\triangle}{=}-\min_{\bar{\zeta}\in\bar{G}\psi_\pi(x)}\zeta^{-1}+\zeta^0+\frac{1}{2}\|\zeta\|^2, \tag{2.9}$$

$$\bar{G}\psi_\pi(\hat{x})\stackrel{\triangle}{=}\operatorname*{conv}_{y\in Y}\operatorname*{conv}_{k\in\mathbf{r}}\left\{\begin{pmatrix}\phi^k_\pi(\hat{x},y)-\omega_\pi(\hat{x},y)\\ \psi_\pi(\hat{x})-\omega_\pi(\hat{x},y)\\ \nabla_x\phi^k_\pi(\hat{x},y)\end{pmatrix}\right\}, \tag{2.10}$$

where elements of $\bar{G}\psi_\pi(\hat{x})\subset\mathbb{R}^{n+2}$ *are denoted by* $\bar{\zeta}=(\zeta^{-1},\zeta^0,\zeta)$, *with* $\zeta\in\mathbb{R}^n$. *Then* (i) $\theta_\pi(\cdot)$ *is continuous and nonpositive valued, and* (ii) *if* $\hat{x}$ *is a local minimizer for* $\mathbf{P}_\pi$, *then* $\theta_\pi(\hat{x})=0$.

Proof. (i) By Corollaries 5.3.9 and 5.4.2 in [17], $\theta_\pi(\cdot)$ is continuous and nonpositive valued.

(ii) If $\hat{x}$ is a local minimizer for $\mathbf{P}_\pi$, then

$$d_-\psi_\pi(x;h) \geq 0 \quad \forall\, h \in \mathbb{R}^n, \tag{2.11a}$$

where $d_-\psi_\pi(x;h)$ is the lower Dini directional derivatives of $\psi_\pi(\cdot)$ at a point x in a direction h, i.e.,

$$d_-\psi_\pi(x;h) \stackrel{\triangle}{=} \liminf_{t \downarrow 0} \frac{\psi_\pi(x+th)-\psi_\pi(x)}{t}. \tag{2.11b}$$

Next, for any $x \in \mathbb{R}^n$ and $y \in Y$, let $\hat{\mathbf{r}}_\pi(x,y) \stackrel{\triangle}{=} \{k \in \mathbf{r} \mid \phi_\pi^k(x,y) = \omega_\pi(x,y)\}$. By using (2.4b), the facts that for any $y \in Y$, $-\psi_\pi(x) \leq -\omega_\pi(x,y)$ and that $\hat{\mathbf{r}}_\pi(x,y) \subset \mathbf{r}$, and the definition of $\hat{\mathbf{r}}_\pi(x,y)$, we obtain that for any $x, h \in \mathbb{R}^n$ and $t > 0$,

$$\begin{aligned}
\frac{\psi_\pi(x+th)-\psi_\pi(x)}{t} &= \max_{y \in \hat{Y}_\pi(x+th)} \min_{k \in \mathbf{r}} \frac{\phi_\pi^k(x+th,y)-\psi_\pi(x)}{t} \\
&\leq \max_{y \in \hat{Y}_\pi(x+th)} \min_{k \in \hat{\mathbf{r}}_\pi(x,y)} \frac{\phi_\pi^k(x+th,y)-\omega_\pi(x,y)}{t} \\
&= \max_{y \in \hat{Y}_\pi(x+th)} \min_{k \in \hat{\mathbf{r}}_\pi(x,y)} \langle \nabla_x \phi_\pi^k(x+sth,y), h\rangle,
\end{aligned} \tag{2.11c}$$

where $s \in [0,1]$. Hence, since $\hat{Y}_\pi(\cdot)$ is outer semicontinuous in the sense of Kuratowski–Painlevé (see [21, 17]), we have that

$$\liminf_{t \downarrow 0} \frac{\psi_\pi(x+th)-\psi_\pi(x)}{t} \leq \max_{y \in \hat{Y}_\pi(x)} \min_{k \in \hat{\mathbf{r}}_\pi(x,y)} \langle \nabla_x \phi_\pi^k(x,y), h\rangle. \tag{2.11d}$$

Next, we proceed by contraposition. Suppose that $0 \notin \bar{G}\psi_\pi(\hat{x})$. Then there exists a nonzero vector $h \in \mathbb{R}^n$ such that $\langle \nabla_x \phi_\pi^k(\hat{x},y), h\rangle < 0$ for all $y \in \hat{Y}_\pi(\hat{x})$ and all $k \in \hat{\mathbf{r}}_\pi(\hat{x},y)$. Hence, by (2.11d), $d_-\psi_\pi(\hat{x};h) < 0$. Therefore, (2.11a) implies that $0 \in \bar{G}\psi_\pi(\hat{x})$ and $\theta_\pi(\hat{x}) = 0$. ∎

In view of Theorem 2.3, we can formulate the following optimality condition for $\mathbf{P}$.

THEOREM 2.5. *Suppose that Assumption* 2.1 *holds and that* $\mathbf{IP}(\hat{x},0)$ *is locally calm at* $\hat{x} \in \mathbb{R}^n$. *If* $\hat{x}$ *is a local minimizer for* $\mathbf{P}$, *then there exists a* $\hat{\pi} < \infty$ *such that* $\theta_{\hat{\pi}}(\hat{x}) = 0$ *and* $\psi(\hat{x}) = \psi_{\hat{\pi}}(\hat{x})$. ∎

The optimality condition for $\mathbf{P}$ in Theorem 2.5 can be related to an optimality condition in [24]; see Appendix A.

In the remainder of the section, we derive results leading to the conclusion that Assumption 2.1, together with Assumption 2.6, are sufficient conditions for local calmness.

ASSUMPTION 2.6. *We assume that for any* $x \in \mathbb{R}^n$ *and* $y \in \hat{Y}(x)$, *the vectors* $\nabla_y f^k(x,y), k \in \mathbf{r}_\mathbf{1}^*(x,y)$, *together with the vectors* $\nabla g^k(y), k \in \mathbf{r}_\mathbf{2}^*(y)$, *are linearly independent, where* $\mathbf{r_1} = \{1,\ldots,r_1\}$, $\mathbf{r_2} = \{1,\ldots,r_2\}$, *and*

$$\mathbf{r}_\mathbf{1}^*(x,y) \stackrel{\triangle}{=} \{k \in \mathbf{r_1} | f^k(x,y) - \|f(x,y)_+\|_\infty = 0\}, \tag{2.12a}$$

$$\mathbf{r}_\mathbf{2}^*(y) \stackrel{\triangle}{=} \{k \in \mathbf{r_2} | g^k(y) = 0\}. \tag{2.12b}$$ ∎

Next, we will define a test function, which plays a crucial role in determining the value of the penalty π that is sufficiently large to ensure the local equivalence between

$\mathbf{P}$ and $\mathbf{P}_\pi$ near a point $\hat{x} \in \mathbb{R}^n$. We need the following building blocks: Let

$$A(x,y) \triangleq \begin{pmatrix} f_y(x,y) \\ g_y(y) \end{pmatrix} \tag{2.13a}$$

be an $(r_1 + r_2) \times m$ matrix with

$$f_y(x,y) \triangleq (\nabla_y f^1(x,y), \ldots, \nabla_y f^{r_1}(x,y))^T, \tag{2.13b}$$

$$g_y(y) \triangleq (\nabla g^1(y), \ldots, \nabla g^{r_2}(y))^T, \tag{2.13c}$$

and let

$$B(x,y) \triangleq \operatorname{diag}(B_1(x,y), B_2(y)) \tag{2.13d}$$

be an $(r_1+r_2)\times(r_1+r_2)$ diagonal matrix defined in terms of the two diagonal matrices

$$B_1(x,y) \triangleq \operatorname{diag}([f^1(x,y) - \|f(x,y)_+\|_\infty]^2, \ldots, [f^{r_1}(x,y) - \|f(x,y)_+\|_\infty]^2), \tag{2.13e}$$

$$B_2(y) \triangleq \operatorname{diag}([g^1(y)]^2, \ldots, [g^{r_2}(y)]^2). \tag{2.13f}$$

Furthermore, let $z : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^{r_1+r_2}$ be defined by

$$z(x,y) \triangleq (\eta(x,y), \xi(x,y))^T \triangleq [A(x,y)A(x,y)^T + B(x,y)]^+ A(x,y)\nabla_y \phi(x,y), \tag{2.13g}$$

where $\eta(x,y) \in \mathbb{R}^{r_1}$, $\xi(x,y) \in \mathbb{R}^{r_2}$, and M^+ denotes the pseudoinverse[2] of the matrix M.

Using a similar construction as in [6], we define for any $\pi > 0$ the test function $t_\pi : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ by

$$t_\pi(x,y) \triangleq -\pi + \sigma \sum_{k=1}^{r_1} |\eta^k(x,y)|, \tag{2.13h}$$

where $\sigma > 1$.

The function $\eta(\cdot,\cdot)$ has the following properties, which will ensure that the test function in (2.13h) is well-defined for all $x \in \mathbb{R}^n$ and $y \in \mathbb{R}^m$, and it is continuous whenever $\eta(\cdot,\cdot)$ is continuous. Note that $\eta(x,y)$ is under certain assumptions related to the multipliers of the "inner problem" in (1.2); see the proof of Lemma 2.7.

LEMMA 2.7. *Suppose Assumption* 2.1 *holds and* $\sigma > 1$ *in* (2.13h).

(i) *Then* $\eta(\cdot,\cdot)$ *is well-defined for all* $x \in \mathbb{R}^n$ *and* $y \in \mathbb{R}^m$.

(ii) *If* $x \in \mathbb{R}^n$ *and* $y \in \mathbb{R}^m$ *are such that* $\nabla_y f^k(x,y), k \in \mathbf{r_1^*}(x,y)$ *(see* (2.12a)*), together with* $\nabla g^k(y), k \in \mathbf{r_2^*}(y)$ *(see* (2.12b)*), are linearly independent, then* $A(x,y)A(x,y)^T + B(x,y)$ *(see* (2.13a), (2.13d)*) is positive definite, and* $\eta(\cdot,\cdot)$ *is continuous at* $x \in \mathbb{R}^n$ *and* $y \in \mathbb{R}^m$.

(iii) *If* $x \in \mathbb{R}^n$ *and* $\pi > 0$ *is such that* $t_\pi(x,y_x) \le 0$ *for some* $y_x \in \hat{Y}_\pi(x)$*, and* $\nabla_y f^k(x,y_x)$*,* $k \in \mathbf{r_1^*}(x,y_x)$ *(see* (2.12a)*), together with* $\nabla g^k(y_x), k \in \mathbf{r_2^*}(y_x)$ *(see* (2.12b)*), are linearly independent, then* $y_x \in \hat{Y}(x)$ *and* $\psi(x) = \psi_\pi(x)$.

[2]The pseudoinverse of a real matrix M is obtained by first taking a singular-value decomposition $M = PDQ$, with P and Q unitary matrices, and D diagonal, and then setting $M^+ = Q^T D^+ P^T$. The pseudoinverse of a diagonal matrix is obtained by replacing the ith diagonal term d_{ii} with $1/d_{ii}$ whenever $d_{ii} \neq 0$, otherwise with 0; see [11].

Proof. (i) By Theorem 4 in section 5.4 in [11], the pseudoinverse is unique, and hence $\eta(\cdot,\cdot)$ is uniquely defined for all $x \in \mathbb{R}^n$ and $y \in \mathbb{R}^m$.

(ii) Let $x \in \mathbb{R}^n$ and $y \in \mathbb{R}^m$ be such that $\nabla_y f^k(x,y), k \in \mathbf{r}_1^*(x,y)$ (see (2.12a)), together with $\nabla g^k(y), k \in \mathbf{r}_2^*(y)$ (see (2.12b)), are linearly independent. By the definition in (2.13g), $z(x,y)$ satisfies the equation

$$[A(x,y)A(x,y)^T + B(x,y)]z(x,y) - A(x,y)\nabla_y \phi(x,y) = 0, \tag{2.14a}$$

which is also the first-order necessary optimality condition for the unconstrained convex quadratic optimization problem

$$\min_{z \in \mathbb{R}^{r_1+r_2}} \{\|-\nabla_y \phi(x,y) + A(x,y)^T z\|^2 + \langle z, B(x,y)z\rangle\}. \tag{2.14b}$$

We will first show that $z(x,y)$ is the unique solution of (2.14b). Since (2.14b) is a quadratic problem, we need only to show that the quadratic function being minimized is positive definite. Clearly, this function is positive semidefinite. Let $z = (\eta, \xi)$. Then the quadratic part of the cost function in (2.14b) can be written as follows:

$$\begin{aligned} &\langle z, [A(x,y)A(x,y)^T + B(x,y)]z\rangle \\ &= \|f_y(x,y)^T\eta + g_y(y)^T\xi\|^2 + \langle \eta, B_1(x,y)\eta\rangle + \langle \xi, B_2(x,y)\xi\rangle. \end{aligned} \tag{2.15a}$$

Hence, the quadratic function in (2.14b) is positive definite if and only if

$$\|f_y(x,y)^T\eta + g_y(y)^T\xi\|^2 + \langle \eta, B_1(x,y)\eta\rangle + \langle \xi, B_2(x,y)\xi\rangle = 0 \tag{2.15b}$$

implies that $\eta = 0$ and $\xi = 0$. Now, when (2.15b) holds, we must have that $\eta^k = 0$ for all $k \notin \mathbf{r}_1^*(x,y)$, and $\xi^k = 0$ for all $k \notin \mathbf{r}_2^*(y)$. Hence, (2.15b) implies that

$$\sum_{k \in \mathbf{r}_1^*(x,y)} \eta^k \nabla_y f^k(x,y) + \sum_{k \in \mathbf{r}_2^*(y)} \xi^k \nabla g^k(y) = 0. \tag{2.15c}$$

It now follows from the linear independence hypothesis that (2.15c), and hence also (2.15b), hold if and only if $\eta = 0$ and $\xi = 0$. This shows that $[A(x,y)A(x,y)^T + B(x,y)]$ is positive definite, and hence $z(x,y)$ is the unique solution of (2.14b).

Next, since there is an unique solution to (2.14b), it follows that $[A(x,y)A(x,y)^T + B(x,y)]$ is invertible, and the inverse is identical to the pseudoinverse. Hence,

$$z(x,y) \stackrel{\triangle}{=} [A(x,y)A(x,y)^T + B(x,y)]^{-1} A(x,y)\nabla_y \phi(x,y). \tag{2.15d}$$

Since $[A(x,y)A(x,y)^T + B(x,y)]$ is positive definite, there exists $\epsilon > 0$ such that $[A(x',y')A(x',y')^T + B(x',y')]$ is positive definite for all $(x',y') \in \mathbb{B}((x,y),\epsilon)$. Hence, (2.15d) holds, with $x = x'$ and $y = y'$, for all $(x',y') \in \mathbb{B}((x,y),\epsilon)$, which implies that $z(\cdot,\cdot)$ is continuous at (x,y).

(iii) Let $x \in \mathbb{R}^n$, $y_x \in \mathbb{R}^m$, and $\pi > 0$ be such that $t_\pi(x,y_x) \leq 0$, $y_x \in \hat{Y}_\pi(x)$, and $\nabla_y f^k(x,y_x)$, $k \in \mathbf{r}_1^*(x,y_x)$, together with $\nabla g^k(y_x), k \in \mathbf{r}_2^*(y_x)$, are linearly independent at (x,y_x). Then y_x is a minimizer for the problem (see (2.3a) and (2.4b))

$$\min_{y \in Y} \max_{k \in \mathbf{r}} \{-\phi_\pi^k(x,y)\}, \tag{2.15e}$$

and it follows from first-order optimality conditions (see [17]) that there exist multipliers $\nu \in \mathbb{R}^r$, with $\nu^k \geq 0, k \in \mathbf{r}, \sum_{k=1}^r \nu^k = 1$, and $\mu \in \mathbb{R}^{r_2+1}$, with $\mu^k \geq 0$,

$k \in \{0, 1, \ldots, r_2\}$, $\sum_{k=0}^{r_2} \mu^k = 1$, such that

$$\mu^0 \left[\sum_{k=1}^{r} -\nu^k \nabla_y \phi_\pi^k(x, y_x) \right] + \sum_{k=1}^{r_2} \mu^k \nabla g^k(y_x) = 0, \tag{2.15f}$$

$$\mu^0 \left[\sum_{k=1}^{r} \nu^k (-\phi_\pi^k(x, y_x) + \omega_\pi(x, y_x)) \right] + \sum_{k=1}^{r_2} \mu^k g^k(y_x) = 0. \tag{2.15g}$$

By the linear independence hypothesis, $\mu^0 > 0$. Using (2.3b) and (2.3c), (2.15f) can be rewritten as

$$-\nabla_y \phi(x, y_x) + \sum_{k=1}^{r_1} \nu^k \pi \nabla_y f^k(x, y_x) + \sum_{k=1}^{r_2} \frac{\mu^k}{\mu^0} \nabla g^k(y_x) = 0, \tag{2.15h}$$

and, also using the fact that each term in (2.15g) must be nonnegative, (2.15g) can be rewritten as

$$\nu^k \pi (f^k(x, y_x) - \|f(x, y_x)_+\|_\infty) = 0, \quad k \in \mathbf{r_1}, \tag{2.15i}$$

$$\nu^r \pi \|f(x, y_x)_+\|_\infty = 0, \tag{2.15j}$$

$$\frac{\mu^k}{\mu^0} g^k(y_x) = 0, \quad k \in \mathbf{r_2}. \tag{2.15k}$$

Then we see from (2.14b), and the definitions (2.13a) and (2.13d), that

$$\min_{\eta \in \mathbb{R}^{r_1}, \xi \in \mathbb{R}^{r_2}} \left\{ \left\| -\nabla_y \phi(x, y_x) + \sum_{k=1}^{r_1} \eta^k \nabla_y f^k(x, y_x) + \sum_{k=1}^{r_2} \xi^k \nabla g^k(y_x) \right\|^2 + \sum_{k=1}^{r_1} [\eta^k (f^k(x, y_x) - \|f(x, y_x)_+\|_\infty)]^2 + \sum_{k=1}^{r_2} [\xi^k g^k(y_x)]^2 \right\} \geq 0. \tag{2.15l}$$

Since the cost function in (2.15l) is nonnegative for all vectors $\eta \in \mathbb{R}^{r_1}$ and $\xi \in \mathbb{R}^{r_2}$, it follows, by taking $\eta = (\nu^1 \pi, \ldots, \nu^{r_1} \pi)$ and $\xi = (\mu^1/\mu^0, \ldots, \mu^{r_2}/\mu^0)$ and (2.15h), (2.15i), (2.15k), that

$$\begin{aligned} &\min_{\eta \in \mathbb{R}^{r_1}, \xi \in \mathbb{R}^{r_2}} \left\{ \left\| -\nabla_y \phi(x, y_x) + \sum_{k=1}^{r_1} \eta^k \nabla_y f^k(x, y_x) + \sum_{k=1}^{r_2} \xi^k \nabla g^k(y_x) \right\|^2 \right. \\ &\qquad \left. + \sum_{k=1}^{r_1} [\eta^k (f^k(x, y_x) - \|f(x, y_x)_+\|_\infty)]^2 + \sum_{k=1}^{r_2} [\xi^k g^k(y_x)]^2 \right\} \\ \leq \quad &\left\| -\nabla_y \phi(x, y_x) + \sum_{k=1}^{r_1} \nu^k \pi \nabla_y f^k(x, y_x) + \sum_{k=1}^{r_2} \frac{\mu^k}{\mu^0} \nabla g^k(y_x) \right\|^2 \\ &+ \sum_{k=1}^{r_1} [\nu^k \pi (f^k(x, y_x) - \|f(x, y_x)_+\|_\infty)]^2 + \sum_{k=1}^{r_2} \left[\frac{\mu^k}{\mu^0} g^k(y_x) \right]^2 = 0. \end{aligned} \tag{2.15m}$$

Since the linear independence property holds at (x, y_x), it follows from the proof of part (ii) that (2.14b) has a unique solution. Hence, in view of (2.15l)–(2.15m), $\eta(x, y_x) = (\nu^1 \pi, \ldots, \nu^{r_1} \pi)$.

Suppose that $\|f(x,y_x)_+\|_\infty > 0$. Then by (2.15j), $\nu^r = 0$, and hence $\sum_{k=1}^{r_1} \nu^k = 1$. Now, since $t_\pi(x,y_x) \leq 0$ by assumption (see (2.13h)),

$$\pi \geq \sigma \sum_{k=1}^{r_1} |\eta^k(x,y_x)| = \sigma \sum_{k=1}^{r_1} \nu^k \pi = \sigma\pi. \tag{2.15n}$$

However, this is a contradiction because $\sigma > 1$. Hence $f(x,y_x) \leq 0$. Since $y_x \in \hat{Y}_\pi(x)$, we have that for every $y' \in Y$ such that $f(x,y') \leq 0$,

$$\begin{aligned} \phi(x,y') &= \phi(x,y') - \pi\|f(x,y')_+\|_\infty \\ &\leq \phi(x,y_x) - \pi\|f(x,y_x)_+\|_\infty \\ &= \phi(x,y_x). \end{aligned} \tag{2.15o}$$

Hence, $y_x \in \hat{Y}(x)$, and

$$\begin{aligned} \psi_\pi(x) &= \phi(x,y_x) - \pi\|f(x,y_x)_+\|_\infty \\ &= \psi(x). \end{aligned} \tag{2.15p}$$

This completes the proof. ∎

In the following, for any $S \subset \mathbb{R}^m$ and $\rho > 0$, let $S + \mathbb{B}_\rho \triangleq \{y \in \mathbb{R}^m | \|y - y'\| \leq \rho, y' \in S\}$. Furthermore, we denote the convergence of an infinite (sub)sequence $\{x_i\}_{i\in K}$, $K \in \mathbb{N}$, to a point x, by $x_i \to^K x$.

LEMMA 2.8. *Suppose Assumptions* 2.1 *and* 2.6 *hold. Then, for every $\hat{x} \in \mathbb{R}^n$, there exist a compact set $\Omega(\hat{x}) \subset \mathbb{R}^m$ and a scalar $\rho_{\hat{x}} > 0$ such that*

(i) *for every $x \in \mathbb{B}(\hat{x}, \rho_{\hat{x}})$ and $y \in \Omega(\hat{x})$, $A(x,y)A(x,y)^T + B(x,y)$ (see* (2.13a), (2.13d)*) is positive definite, and hence $\nabla_y f^k(x,y), k \in \mathbf{r}_1^*(x,y)$ (see* (2.12a)*), together with $\nabla g^k(y), k \in \mathbf{r}_2^*(y)$ (see* (2.12b)*), are linearly independent, and*

(ii) $\hat{Y}(\hat{x}) + \mathbb{B}_{\rho_{\hat{x}}} \subset \Omega(\hat{x})$.

Proof. Let $\hat{x} \in \mathbb{R}^n$ be arbitrary. By Assumption 2.6, $\nabla_y f^k(\hat{x},\hat{y}), k \in \mathbf{r}_1^*(\hat{x},\hat{y})$, together with $\nabla g^k(\hat{y}), k \in \mathbf{r}_2^*(\hat{y})$, are linearly independent for any $\hat{y} \in \hat{Y}(\hat{x})$. It follows from Lemma 2.7(ii) that $A(\hat{x},\hat{y})A(\hat{x},\hat{y})^T + B(\hat{x},\hat{y})$ is positive definite. Thus, by continuity of $A(\cdot,\cdot)$ and $B(\cdot,\cdot)$, there exist a compact set $\Omega(\hat{x}) \subset \mathbb{R}^m$ and a $\rho_{\hat{x}} > 0$ such that $A(x,y)A(x,y)^T + B(x,y)$ is positive definite for all $x \in \mathbb{B}(\hat{x}, \rho_{\hat{x}})$ and $y \in \Omega(\hat{x})$, and $\hat{Y}(\hat{x}) + \mathbb{B}_{\rho_{\hat{x}}} \subset \Omega(\hat{x})$.

By positive definiteness, both sides of (2.15a) are strictly positive for all $x \in \mathbb{B}(\hat{x}, \rho_{\hat{x}})$, $y \in \Omega(\hat{x})$, and $z = (\eta, \xi) \neq 0$, with $\eta \in \mathbb{R}^{r_1}$ and $\xi \in \mathbb{R}^{r_2}$. Hence, (2.15b) must imply that (η, ξ) in (2.15b) is zero for all $x \in \mathbb{B}(\hat{x}, \rho_{\hat{x}})$ and $y \in \Omega(\hat{x})$. However, then $\nabla_y f^k(x,y), k \in \mathbf{r}_1^*(x,y)$, together with $\nabla g^k(y), k \in \mathbf{r}_2^*(y)$, must be linearly independent for all $x \in \mathbb{B}(\hat{x}, \rho_{\hat{x}})$ and $y \in \Omega(\hat{x})$. Because if that was not true, we may have (2.15b) satisfied for $(\eta, \xi) \neq 0$. This completes the proof. ∎

LEMMA 2.9. *Suppose that Assumption* 2.1 *holds. Then, for every $\hat{x} \in \mathbb{R}^n$, $\pi > 0$, $\rho > 0$, and $\epsilon > 0$, there exist $\hat{\pi} \in [\pi, \infty)$ and $\hat{\rho} \in (0, \rho]$ such that $\hat{Y}_{\hat{\pi}}(x) \subset \hat{Y}(\hat{x}) + \mathbb{B}_\epsilon$ for all $x \in \mathbb{B}(\hat{x}, \hat{\rho})$.*

Proof. Let $\hat{x} \in \mathbb{R}^n$, $\pi > 0$, $\rho > 0$, and $\epsilon > 0$ be arbitrary. To prove the desired result, we will show that (i) there exists a $\hat{\pi} \in [\pi, \infty)$ such that $\hat{Y}_{\hat{\pi}}(\hat{x}) \subset \hat{Y}(\hat{x}) + \mathbb{B}_{\epsilon/2}$, and (ii) there exists a $\hat{\rho} \in (0, \rho]$ such that $\hat{Y}_{\hat{\pi}}(x) \subset \hat{Y}_{\hat{\pi}}(\hat{x}) + \mathbb{B}_{\epsilon/2}$ for all $x \in \mathbb{B}(\hat{x}, \hat{\rho})$.

(i) Let the set-valued function $\Gamma : [0, \infty) \to 2^{\mathbb{R}^m}$ be defined by

$$\Gamma(s) \triangleq \arg\max_{y \in Y} \phi'(s,y), \tag{2.16a}$$

where $\phi' : [0,\infty) \times \mathbb{R}^m \to \mathbb{R} \bigcup \{-\infty\}$ is defined by

$$\phi'(s,y) \stackrel{\triangle}{=} \begin{cases} \phi(\hat{x},y) - \dfrac{1}{s}\|f(\hat{x},y)_+\|_\infty, & s > 0, \\ \phi(\hat{x},y), & s = 0, \|f(\hat{x},y)_+\|_\infty = 0, \\ -\infty, & s = 0, \|f(\hat{x},y)_+\|_\infty > 0. \end{cases} \tag{2.16b}$$

First, we show that $\Gamma(\cdot)$ is outer semicontinuous at $s = 0$ in the sense of Kuratowski–Painlevé; see [21, 17]. By Theorem 5.3.7 in [17], we need only to show that the outer limit of $\{\Gamma(s_i)\}_{i=0}^\infty$ is contained in $\Gamma(0)$ for any sequence $\{s_i\}_{i=0}^\infty \subset [0,\infty)$ such that $s_i \to 0$. Let $\{s_i\}_{i=0}^\infty \subset [0,\infty)$ be such that $s_i \to 0$, and let $\hat{y} \in \mathbb{R}^m$ be a point in the outer limit of $\{\Gamma(s_i)\}_{i=0}^\infty$. Then there exists a sequence $\{y_i\}_{i=0}^\infty$ such that $y_i \in \Gamma(s_i)$ for all $i \in \mathbb{N}$, and $y_i \to \hat{y}$, as $i \to \infty$.

Now, consider the hypographs (see [21, 17]) of the problems $\max_{y\in Y} \phi'(s_i,y)$ given by

$$E_i \stackrel{\triangle}{=} \{(y^0,y) \in \mathbb{R}^{m+1} \mid y \in Y,\ y^0 \le \phi'(s_i,y)\} \tag{2.16c}$$

and of the problem $\max_{y\in Y} \phi'(0,y)$ given by

$$E \stackrel{\triangle}{=} \{(y^0,y) \in \mathbb{R}^{m+1} \mid y \in Y,\ y^0 \le \phi'(0,y)\}. \tag{2.16d}$$

By Theorem 3.3.2 in [17], the sequence of sets $\{E_i\}_{i=0}^\infty$ converges to E in the Kuratowski–Painlevé sense (see [21, 17]) if and only if (a) for any $y' \in Y$, $\liminf_{i\to\infty} \phi'(s_i,y') \ge \phi'(0,y')$, and (b) for any infinite sequence $\{y_i'\}_{i\in K} \subset Y$, $K \subset \mathbb{N}$, such that $y_i' \to^K y'$, as $i \to \infty$, $\limsup_{i\to^K\infty} \phi'(s_i,y') \le \phi'(0,y')$.

First, we consider (a). Suppose $y' \in Y$. Then we have directly from (2.16b) that $\lim_{i\to\infty} \phi'(s_i,y') = \phi'(0,y')$.

Second, we consider (b). Let $\{y_i'\}_{i\in K} \subset Y$ be an infinite sequence, $K \subset \mathbb{N}$, such that $y_i' \to^K y'$, as $i \to \infty$. Without loss of generality, we assume that $y_i' \to y'$, as $i \to \infty$. Now, we have two cases.

Case I. Suppose $\|f(\hat{x},y')_+\|_\infty > \delta$ for some $\delta > 0$. Then by continuity of $f(\cdot,\cdot)$ there exists an $i_0 \in \mathbb{N}$ such that $\|f(\hat{x},y_i')_+\|_\infty \ge \delta/2$ for all $i > i_0$. Hence, for all $i > i_0$, such that $s_i > 0$, $\phi'(s_i,y_i') = \phi(\hat{x},y_i') - \|f(\hat{x},y_i')_+\|_\infty/s_i \le \phi(\hat{x},y_i') - \delta/(2s_i)$, and for all $i > i_0$, such that $s_i = 0$, $\phi'(s_i,y_i') = -\infty$. Since $s_i \to 0$, we have that $\lim_{i\to\infty} \phi'(s_i,y_i') = \phi'(0,y') = -\infty$.

Case II. Suppose $f(\hat{x},y') \le 0$. Then by (2.16b), $\limsup_{i\to\infty} \phi'(s_i,y_i') \le \phi(\hat{x},y') = \phi'(0,y')$.

Hence, by Theorem 3.3.2 in [17], $\{E_i\}_{i=0}^\infty$ converges to E. As a consequence of the convergence of $\{E_i\}_{i=0}^\infty$ to E, Theorem 3.3.3 in [17] states that any accumulation point of a sequence of global maximizers of $\max_{y\in Y} \phi'(s_i,y)$ is a global maximizer of $\max_{y\in Y} \phi'(0,y)$. Hence, $\hat{y} \in \Gamma(0)$, which is a contradiction. Hence, we have that $\Gamma(\cdot)$ is outer semicontinuous at $s = 0$.

Next, let $y^* \in \Gamma(0)$. It follows from (2.16b) and Assumption 2.1(ii) that $f(\hat{x},y^*) \le 0$ and $y^* \in \hat{Y}(\hat{x})$. Hence, $\Gamma(0) \subset \hat{Y}(\hat{x})$.

Finally, by outer semicontinuity of $\Gamma(\cdot)$ at $s = 0$, there exists a $\hat{\pi} \in [\pi,\infty)$ such that $\hat{Y}_{\hat{\pi}}(\hat{x}) \subset \hat{Y}(\hat{x}) + \mathbb{B}_{\epsilon/2}$, and (i) holds.

(ii) Let $\hat{\pi}$ be as in (i). First, we show that $\hat{Y}_{\hat{\pi}}(\cdot)$ is outer semicontinuous at $\hat{x}$. By Theorem 5.3.7 in [17], we need only to show that the outer limit of $\{\hat{Y}_{\hat{\pi}}(x_i)\}_{i=0}^\infty$ is

contained in $\hat{Y}_{\hat{\pi}}(\hat{x})$ for any sequence $\{x_i\}_{i=0}^{\infty} \subset \mathbb{R}^n$ such that $x_i \to \hat{x}$. Let $\{y_i\}_{i=0}^{\infty}$ be an arbitrary sequence such that $y_i \in \hat{Y}_{\hat{\pi}}(x_i)$ and $y_i \to \hat{y}$. Then

$$\phi(x_i, y_i) - \hat{\pi}\|f(x_i, y_i)_+\|_\infty \geq \phi(x_i, y) - \hat{\pi}\|f(x_i, y)_+\|_\infty \tag{2.16e}$$

for all $i \in \mathbb{N}$ and $y \in Y$. Hence, by adding $\phi(\hat{x}, \hat{y}) - \hat{\pi}\|f(\hat{x}, \hat{y})_+\|_\infty$ to both sides of (2.16e), and rearranging terms, we obtain that

$$\begin{aligned}\phi(\hat{x}, \hat{y}) - \hat{\pi}\|f(\hat{x}, \hat{y})_+\|_\infty \geq \max_{y \in Y}\{&\phi(\hat{x}, \hat{y}) - \hat{\pi}\|f(\hat{x}, \hat{y})_+\|_\infty - \phi(x_i, y_i) \\ &+ \hat{\pi}\|f(x_i, y_i)_+\|_\infty + \phi(x_i, y) - \hat{\pi}\|f(x_i, y)_+\|_\infty\}.\end{aligned} \tag{2.16f}$$

It now follows by the continuity of the right-hand side of (2.16f) that $\phi(\hat{x}, \hat{y}) - \hat{\pi}\|f(\hat{x}, \hat{y})_+\|_\infty \geq \psi_{\hat{\pi}}(\hat{x})$, and hence that $\hat{y} \in \hat{Y}_{\hat{\pi}}(\hat{x})$. Hence, $\hat{Y}_{\hat{\pi}}(\cdot)$ is outer semicontinuous at $\hat{x}$, which implies that (ii) holds. Now, the conclusion follows directly from (i) and (ii). ∎

THEOREM 2.10. *Suppose Assumptions* 2.1 *and* 2.6 *hold. Then, for any* $\hat{x} \in \mathbb{R}^n$, $\mathbf{IP}(\hat{x}, 0)$ *is locally calm at* $\hat{x}$.

Proof. Let $\hat{x} \in \mathbb{R}^n$. By Lemmas 2.7(ii) and 2.8, there exist a compact set $\Omega(\hat{x}) \subset \mathbb{R}^m$ and $\rho_{\hat{x}} > 0$ such that $\eta(\cdot, \cdot)$ is continuous on $\mathbb{B}(\hat{x}, \rho_{\hat{x}}) \times \Omega(\hat{x})$ and $\hat{Y}(\hat{x}) + \mathbb{B}_{\rho_{\hat{x}}} \subset \Omega(\hat{x})$. Hence,

$$\pi^* \triangleq \max_{x \in \mathbb{B}(\hat{x}, \rho_{\hat{x}})} \max_{y \in \Omega(\hat{x})} \sigma \sum_{k=1}^{r_1} |\eta^k(x, y)|, \tag{2.17a}$$

with $\sigma > 1$, is well-defined. By Lemma 2.9, there exist $\hat{\rho} \in (0, \rho_{\hat{x}}]$ and $\hat{\pi} \geq \pi^*$ such that $\hat{Y}_{\hat{\pi}}(x) \subset \hat{Y}(\hat{x}) + \mathbb{B}_{\rho_{\hat{x}}}$ for all $x \in \mathbb{B}(\hat{x}, \hat{\rho})$. Let $x \in \mathbb{B}(\hat{x}, \hat{\rho})$ be arbitrary. Then $\hat{Y}_{\hat{\pi}}(x) \subset \Omega(\hat{x})$, and hence, for any $y_x \in \hat{Y}_{\hat{\pi}}(x)$, we have by (2.17a) and (2.13h) that $t_{\hat{\pi}}(x, y_x) \leq 0$, and by Lemma 2.8 that $\nabla_y f^k(x, y_x), k \in \mathbf{r}_1^*(x, y_x)$, together with $\nabla g^k(y_x), k \in \mathbf{r}_2^*(y_x)$, are linearly independent. Hence, by Lemma 2.7(iii), $y_x \in \hat{Y}(x)$. Next, let $y \in Y$ and $u \in \mathbb{R}^{r_1}$ be such that $f(x, y) \leq u$. Then

$$\begin{aligned}\phi(x, y_x) &= \phi(x, y_x) - \hat{\pi}\|f(x, y_x)_+\|_\infty \\ &\geq \phi(x, y) - \hat{\pi}\|f(x, y)_+\|_\infty \\ &\geq \phi(x, y) - \hat{\pi}\|(f(x, y) - u)_+\|_\infty - \hat{\pi}\|u\|_\infty \\ &= \phi(x, y) - \hat{\pi}\|u\|_\infty.\end{aligned} \tag{2.17b}$$

By (2.6), $v(x, 0) = \phi(x, y_x)$. For every $u \in \mathbb{R}^{r_1}$ such that $v(x, u) > -\infty$, there exists $y'_u \in Y$ such that $f(x, y'_u) \leq u$ and $\phi(x, y'_u) = v(x, u)$. Hence, by (2.17b), for every $u \in \mathbb{R}^{r_1}$ such that $v(x, u) > -\infty$, we have that

$$v(x, u) - v(x, 0) \leq \hat{\pi}\|u\|_\infty. \tag{2.17c}$$

Since (2.17c) also holds for $u \in \mathbb{R}^{r_1}$ such that $v(x, u) = -\infty$, we have that (2.17c) holds for every $u \in \mathbb{R}^{r_1}$. Finally, because $x \in \mathbb{B}(\hat{x}, \hat{\rho})$ was assumed arbitrary, the conclusion follows with $\hat{\alpha} = \hat{\pi}$. ∎

3. Approximations to $\mathbf{P}_\pi$. In view of Theorem 2.3, $\mathbf{P}$ can be solved by solving $\mathbf{P}_\pi$ for a sufficiently large $\pi > 0$. To facilitate the solution of $\mathbf{P}_\pi$, we introduce two approximations. First, for any set $Y_N \subset Y$, $N \in \mathbb{N} \triangleq \{1, 2, \dots\}$, of finite cardinality

and $\pi > 0$, we define the approximation $\psi_{\pi,N} : \mathbb{R}^n \to \mathbb{R}$ to the function $\psi_\pi(\cdot)$ (see (2.3a)) by

$$\psi_{\pi,N}(x) \stackrel{\triangle}{=} \max_{y \in Y_N} \omega_\pi(x, y), \tag{3.1a}$$

with $\omega_\pi(\cdot,\cdot)$ as in (2.3d).

Second, we introduce a smoothing technique that can be found in [1, 14, 18]. For any $\pi > 0$ and $p > 0$, let $\omega_{\pi,p} : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ be the smooth approximation to $\omega_\pi(\cdot,\cdot)$ defined by

$$\omega_{\pi,p}(x, y) \stackrel{\triangle}{=} -\frac{1}{p} \ln \left(\sum_{k=1}^{r} e^{-p\phi_\pi^k(x,y)} \right). \tag{3.1b}$$

Hence, for any $\pi > 0$, $N \in \mathbb{N}$, and $p > 0$ we define a family of min-max approximations to $\mathbf{P}_\pi$ by

$$\mathbf{P}_{\pi,N,p} \qquad \min_{x \in \mathbb{R}^n} \psi_{\pi,N,p}(x), \tag{3.1c}$$

where $\psi_{\pi,N,p} : \mathbb{R}^n \to \mathbb{R}$ is defined by

$$\psi_{\pi,N,p}(x) \stackrel{\triangle}{=} \max_{y \in Y_N} \omega_{\pi,p}(x, y). \tag{3.2a}$$

Referring to section 2.1 in [17], we find that a continuous optimality function, $\theta_{\pi,N,p} : \mathbb{R}^n \to \mathbb{R}$, for the problem $\mathbf{P}_{\pi,N,p}$ is given by

$$\theta_{\pi,N,p}(x) \stackrel{\triangle}{=} -\min_{\bar{\xi} \in \bar{G}\psi_{\pi,N,p}(x)} \xi^0 + \frac{1}{2}\|\xi\|^2, \tag{3.2b}$$

where $\bar{\xi} = (\xi^0, \xi) \in \mathbb{R}^{n+1}$, with $\xi \in \mathbb{R}^n$, and

$$\bar{G}\psi_{\pi,N,p}(x) \stackrel{\triangle}{=} \operatorname*{conv}_{y \in Y_N} \left\{ \begin{pmatrix} \psi_{\pi,N,p}(x) - \omega_{\pi,p}(x, y) \\ \nabla_x \omega_{\pi,p}(x, y) \end{pmatrix} \right\}. \tag{3.2c}$$

We require that the error associated with the discretization of the set Y satisfies a certain relation as specified in Assumption 3.1(ii). Note also that Assumption 3.1(iv) is a Mangasarian–Fromowitz-type constraint qualification.

ASSUMPTION 3.1. *We assume that*

(i) $\phi(\cdot,\cdot), f^k(\cdot,\cdot), k \in \mathbf{r_1}$, *and* $g^k(\cdot), k \in \mathbf{r_2}$, *are twice continuously differentiable;*

(ii) *there exist a strictly decreasing function* $\Delta : \mathbb{N} \bigcup \{\infty\} \to [0, \infty)$, *with the property that* $\Delta(N) \to 0$, *as* $N \to \infty$, *and* $\Delta(\infty) \stackrel{\triangle}{=} 0$, *and constants* $N_0 \in \mathbb{N}$, $C < \infty$ *such that for every* $N \geq N_0$ *and* $y \in Y$ *there exists a* $y' \in Y_N$ *such that*

$$\|y - y'\| \leq C\Delta(N); \tag{3.3a}$$

(iii) *for every* $N \geq N_0$ *and* $x \in \mathbb{R}^n$

$$\{y \in Y_N \mid f(x, y) \leq 0\} \neq \emptyset; \tag{3.3b}$$

(iv) *for any* $x \in \mathbb{R}^n$ *and* $y \in Y$ *there exist an* $h \in \mathbb{R}^m$ *and* $\hat{u} > 0$ *such that for all* $k \in \mathbf{r_1}$ *satisfying* $f^k(x,y) = 0$, $\langle \nabla_y f^k(x,y), h\rangle < 0$, *and for all* $u \in (0,\hat{u}]$, $g(y+uh) \leq 0$. ∎

For example, if Y is the unit cube in $\mathbb{R}^m$, i.e., $Y = I^m$, with $I = [0,1]$, then we can define $Y_N = I_N^m$, where

$$I_N = \{0, 1/a(N), 2/a(N), \ldots, (a(N)-1)/a(N), 1\}, \tag{3.3c}$$

with $a(N) = 2^{N-N_0}N_0$. In this case, $\Delta(N) = 1/a(N)$ and $C = \frac{1}{2}m^{1/2}$.

We need the following notation: Let $\hat{Y}_N : \mathbb{R}^n \to 2^{\mathbb{R}^m}$ and $\hat{Y}_{\pi,N} : \mathbb{R}^n \to 2^{\mathbb{R}^m}$ be defined by

$$\hat{Y}_N(x) \stackrel{\triangle}{=} \arg\max_{y \in Y_N} \{\phi(x,y) | f(x,y) \leq 0\}, \tag{3.4a}$$

$$\hat{Y}_{\pi,N}(x) \stackrel{\triangle}{=} \{y \in Y_N \mid \omega_\pi(x,y) = \psi_{\pi,N}(x)\} = \arg\max_{y \in Y_N} \{\phi(x,y) - \pi\|f(x,y)_+\|_\infty\}. \tag{3.4b}$$

LEMMA 3.2. *Suppose Assumptions* 2.1 *and* 3.1 *hold and* $\pi > 0$. *If* $x_i \to \hat{x}$, $y_i \to \hat{y}$, $N_i \to \infty$, *as* $i \to \infty$, *with* $y_i \in \hat{Y}_{\pi,N_i}(x_i)$ *and* $N_i \in \mathbb{N}$, *for all* $i \in \mathbb{N}$, *then* $\hat{y} \in \hat{Y}_\pi(\hat{x})$.

Proof. Let $\pi > 0$, $\{x_i\}_{i=0}^\infty \subset \mathbb{R}^n$, $\{y_i\}_{i=0}^\infty \subset \mathbb{R}^m$, $\{N_i\}_{i=0}^\infty \subset \mathbb{N}$, $\hat{x} \in \mathbb{R}^n$, and $\hat{y} \in Y$ be such that $y_i \in \hat{Y}_{\pi,N_i}(x_i)$, $x_i \to \hat{x}$, $y_i \to \hat{y}$, and $N_i \to \infty$. Then

$$\phi(x_i,y_i) - \pi\|f(x_i,y_i)_+\|_\infty \geq \phi(x_i,y) - \pi\|f(x_i,y)_+\|_\infty \tag{3.5a}$$

for all $i \in \mathbb{N}$ and $y \in Y_{N_i}$. Hence, by adding $\phi(\hat{x},\hat{y}) - \pi\|f(\hat{x},\hat{y})_+\|$ to both sides of (3.5a), and rearranging terms, we obtain that

$$\begin{aligned}\phi(\hat{x},\hat{y}) - \pi\|f(\hat{x},\hat{y})_+\|_\infty \geq \max_{y \in Y_{N_i}} \{&\phi(\hat{x},\hat{y}) - \pi\|f(\hat{x},\hat{y})_+\|_\infty - \phi(x_i,y_i) \\ &+ \pi\|f(x_i,y_i)_+\|_\infty + \phi(x_i,y) - \pi\|f(x_i,y)_+\|_\infty\}.\end{aligned} \tag{3.5b}$$

It now follows from the continuity of the right-hand side of (3.5b) (see Corollary 5.4.2 in [17]) and Assumption 3.1(ii) that $\phi(\hat{x},\hat{y}) - \pi\|f(\hat{x},\hat{y})_+\|_\infty \geq \psi_\pi(\hat{x})$, and hence that $\hat{y} \in \hat{Y}_\pi(\hat{x})$. ∎

LEMMA 3.3. *Suppose Assumptions* 2.1 *and* 3.1 *hold. Then,*

(i) *for every* $\hat{x} \in \mathbb{R}^n$ *and* $\epsilon > 0$, *there exist* $\hat{\pi} < \infty$, $\hat{N} < \infty$, *and* $\hat{\rho} > 0$, *such that* $\hat{Y}_{\pi,N}(x) \subset \hat{Y}(\hat{x}) + \mathbb{B}_\epsilon$, *for all* $x \in \mathbb{B}(\hat{x},\hat{\rho})$, $\pi \geq \hat{\pi}$, *and* $N \geq \hat{N}$, $N \in \mathbb{N}$, *and*

(ii) *for every* $\hat{x} \in \mathbb{R}^n$, $N \in \mathbb{N}$, $N \geq N_0$, *with* N_0 *as in Assumption* 3.1(ii), *and* $\epsilon > 0$, *there exist* $\hat{\pi} < \infty$ *and* $\hat{\rho} > 0$, *such that* $\hat{Y}_{\pi,N}(x) \subset \hat{Y}_N(\hat{x}) + \mathbb{B}_\epsilon$, *for all* $x \in \mathbb{B}(\hat{x},\hat{\rho})$ *and* $\pi \geq \hat{\pi}$.

Proof. (i) For any $s \geq 0$, $t \in T \stackrel{\triangle}{=} \{t \in (0,1] \mid 1/t \in \mathbb{N}\} \bigcup \{0\}$, and $x \in \mathbb{R}^n$, let $w = (s,t,x) \in \mathbb{R}^{n+2}$. We define the set-valued function $W : [0,\infty) \times T \times \mathbb{R}^n \to 2^{\mathbb{R}^m}$ by

$$W(w) \stackrel{\triangle}{=} \arg\max_{y \in Y_t^*} \tilde{\phi}(s,x,y), \tag{3.6a}$$

where $Y_t^* \stackrel{\triangle}{=} Y_N$, with $N = 1/t$, for $t \in T$, $t > 0$, $Y_t^* \stackrel{\triangle}{=} Y$ for $t = 0$, and $\tilde{\phi} : [0,\infty) \times \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R} \bigcup \{-\infty\}$ is defined by

$$\tilde{\phi}(s,x,y) \stackrel{\triangle}{=} \begin{cases} \phi(x,y) - \dfrac{1}{s}\|f(x,y)_+\|_\infty, & s > 0, \\ \phi(x,y), & s = 0, \|f(x,y)_+\|_\infty = 0, \\ -\infty, & s = 0, \|f(x,y)_+\|_\infty > 0. \end{cases} \tag{3.6b}$$

Let $\hat{x} \in \mathbb{R}^n$ be arbitrary. We will first show that $W(\cdot)$ is outer semicontinuous at $\hat{w} = (0,0,\hat{x})$ in the sense of Kuratowski–Painlevé; see [21, 17]. By Theorem 5.3.7 in [17], we need only to show that the outer limit of $\{W(w_i)\}_{i=0}^\infty$ is contained in $W(\hat{w})$ for any sequence $\{w_i\}_{i=0}^\infty \subset [0,\infty) \times T \times \mathbb{R}^n$ such that $w_i \to \hat{w}$. Let $\{w_i\}_{i=0}^\infty$, with $w_i = (s_i, t_i, x_i) \in [0,\infty) \times T \times \mathbb{R}^n$, be such that $w_i \to \hat{w}$, and let $\hat{y} \in \mathbb{R}^m$ be a point in the outer limit of $\{W(w_i)\}_{i=0}^\infty$. Then there exists a sequence $\{y_i\}_{i=0}^\infty$ such that $y_i \in W(w_i)$ for all $i \in \mathbb{N}$, and $y_i \to \hat{y}$, as $i \to \infty$.

Now, consider the hypographs (see [21, 17]) of the problems $\max_{y \in Y_{t_i}^*} \tilde{\phi}(s_i, x_i, y)$ given by

$$E_i \stackrel{\triangle}{=} \{(y^0, y) \in \mathbb{R}^{m+1} \mid y \in Y_{t_i}^*,\ y^0 \le \tilde{\phi}(s_i, x_i, y)\} \tag{3.6c}$$

and of the problem $\max_{y \in Y} \tilde{\phi}(0, \hat{x}, y)$ given by

$$E \stackrel{\triangle}{=} \{(y^0, y) \in \mathbb{R}^{m+1} \mid y \in Y,\ y^0 \le \tilde{\phi}(0, \hat{x}, y)\}. \tag{3.6d}$$

By Theorem 3.3.2 in [17], the sequence of sets $\{E_i\}_{i=0}^\infty$ converges to E in the Kuratowski–Painlevé sense (see [21, 17]) if and only if (a) for any $\tilde{y} \in Y$, there exists a sequence $\{\tilde{y}_i\}_{i=0}^\infty$, with $\tilde{y}_i \in Y_{t_i}^*$, such that $\tilde{y}_i \to \tilde{y}$, as $i \to \infty$, and $\liminf_{i\to\infty} \tilde{\phi}(s_i, x_i, \tilde{y}_i) \ge \tilde{\phi}(0, \hat{x}, \tilde{y})$, and (b) for every infinite sequence $\{\tilde{y}_i\}_{i \in K}$, with $K \subset \mathbb{N}$, such that $\tilde{y}_i \in Y_{t_i}^*$ for all $i \in K$, and $\tilde{y}_i \to^K \tilde{y}$, as $i \to \infty$, $\limsup_{i \to^K \infty} \tilde{\phi}(s_i, x_i, \tilde{y}_i) \le \tilde{\phi}(0, \hat{x}, \tilde{y})$.

First, consider (a). Suppose that $\tilde{y} \in Y$. Now, we have two cases.

Case I. Suppose $\|f(\hat{x}, \tilde{y})_+\|_\infty > 0$. Then $\tilde{\phi}(0, \hat{x}, \tilde{y}) = -\infty$, and hence by Assumption 3.1(ii) there exists a sequence $\{\tilde{y}_i\}_{i=0}^\infty \subset Y$ such that $\tilde{y}_i \in Y_{t_i}^*$, for all $i \in \mathbb{N}$, $\tilde{y}_i \to \tilde{y}$, as $i \to \infty$, and $\liminf_{i\to\infty} \tilde{\phi}(s_i, x_i, \tilde{y}_i) \ge \tilde{\phi}(0, \hat{x}, \tilde{y})$.

Case II. Suppose $f(\hat{x}, \tilde{y}) \le 0$. We infer from Assumption 3.1(iv) that there exist $h \in \mathbb{R}^m$, $\delta > 0$, and $u^* > 0$ such that for all $u \in (0, u^*]$

$$f^k(\hat{x}, \tilde{y} + uh) \le -\delta u \quad \forall\ k \in \mathbf{r_1}, \tag{3.6e}$$

$$g(\tilde{y} + uh) \le 0. \tag{3.6f}$$

Let $L < \infty$ be a Lipschitz constant for $f^k(\cdot,\cdot), k \in \mathbf{r_1}$, on $\mathbb{B}(\hat{x}, 1) \times \mathbb{B}(\tilde{y}, u^*\|h\|)$. Hence, by (3.6e), for all $u \in (0, u^*/2]$, $x \in \mathbb{B}(\hat{x}, 1)$, and $y \in \mathbb{B}(\tilde{y} + uh, u^*\|h\|/2)$, we have that

$$f^k(x,y) \le -\delta u + L(\|x - \hat{x}\| + \|\tilde{y} + uh - y\|) \quad \forall\ k \in \mathbf{r_1}, \tag{3.6g}$$

Let $\alpha = \delta/(2L)$. Then there exists $u^{**} \in (0, u^*/2]$ such that for all $u \in (0, u^{**}]$, $\alpha u \le \min\{1,\ u^*\|h\|/2\}$. Let $u \in (0, u^{**}]$. Then, for all $x \in \mathbb{B}(\hat{x}, \alpha u)$ and $y \in \mathbb{B}(\tilde{y} + uh, \alpha u)$,

$$f^k(x,y) \le 0 \quad \forall\ k \in \mathbf{r_1}. \tag{3.6h}$$

Since $x_i \to \hat{x}$ and $t_i \to 0$, as $i \to \infty$, there exists $i_0 \in \mathbb{N}$ such that for all $i \ge i_0$, $\|x_i \to \hat{x}\| \le \alpha u^{**}$ and $C\Delta(1/t_i) \le \alpha u^{**}$ (see Assumption 3.1(ii)). For all $i \ge i_0$, we

define $u_i = \max\{\|x_i - \hat{x}\|,\ C\Delta(1/t_i)\}/\alpha$, and $y_i' = \tilde{y} + u_i h$. By (3.6f), $g(y_i') \leq 0$, and hence $y_i' \in Y$. Then by Assumption 3.1(ii), for every $i \geq i_0$ there exists $\tilde{y}_i \in Y_{t_i}^*$ such that $\|y_i' - \tilde{y}_i\| \leq C\Delta(1/t_i)$. It now follows by construction that $\{\tilde{y}_i\}_{i=i_0}^{\infty}$ is such that $\tilde{y}_i \to \tilde{y}$, as $i \to \infty$, and by (3.6h) that $\|f(x_i, \tilde{y}_i)_+\|_\infty = 0$ for all $i \geq i_0$. Hence, by continuity of $\phi(\cdot,\cdot)$, $\lim_{i\to\infty} \tilde{\phi}(s_i, x_i, \tilde{y}_i) = \tilde{\phi}(0, \hat{x}, \tilde{y})$.

Second, consider (b). Let $\{\tilde{y}_i\}_{i\in K}$ be an infinite sequence, $K \subset \mathbb{N}$, such that $\tilde{y}_i \in Y_{t_i}^*$, for all $i \in K$, $\tilde{y}_i \to^K \tilde{y}$, as $i \to \infty$. Without loss of generality, we assume that $\tilde{y}_i \to \tilde{y}$, as $i \to \infty$. Now, we have two cases.

Case I. Suppose $\|f(\hat{x}, \tilde{y})_+\|_\infty > \delta$ for some $\delta > 0$. Then by continuity of $f(\cdot,\cdot)$, there exists an $i_0 \in \mathbb{N}$ such that $\|f(x_i, \tilde{y}_i)_+\|_\infty \geq \delta/2$ for all $i > i_0$. Hence, for all $i > i_0$, such that $s_i > 0$, $\tilde{\phi}(s_i, x_i, \tilde{y}_i) = \phi(x_i, \tilde{y}_i) - \|f(x_i, \tilde{y}_i)_+\|_\infty/s_i \leq \phi(x_i, \tilde{y}_i) - \delta/(2s_i)$, and for all $i > i_0$, such that $s_i = 0$, $\tilde{\phi}(s_i, x_i, \tilde{y}_i) = -\infty$. Since $w_i \to \hat{w}$, we have that $s_i \to 0$, and hence $\lim_{i\to\infty} \tilde{\phi}(s_i, x_i, \tilde{y}_i) = \tilde{\phi}(0, \hat{x}, \tilde{y}) = -\infty$.

Case II. Suppose $f(\hat{x}, \tilde{y}) \leq 0$. Then it follows directly from (3.6b) that $\limsup_{i\to\infty} \tilde{\phi}(s_i, x_i, \tilde{y}_i) \leq \phi(\hat{x}, \tilde{y}) = \tilde{\phi}(0, \hat{x}, \tilde{y})$.

Hence, by Theorem 3.3.2 in [17], $\{E_i\}_{i=0}^{\infty}$ converges to E. As a consequence of the convergence of $\{E_i\}_{i=0}^{\infty}$ to E, Theorem 3.3.3 in [17] states that any accumulation point of a sequence of global maximizers of $\max_{y\in Y_{t_i}^*} \tilde{\phi}(s_i, x_i, y)$ is a global maximizer of $\max_{y\in Y} \tilde{\phi}(0, \hat{x}, y)$. Hence, $\hat{y} \in W(\hat{w})$. So we have that $W(\cdot)$ is outer semicontinuous at $\hat{w} = (0, 0, \hat{x})$.

Next, let $y^* \in W(\hat{w})$, with $\hat{w} = (0, 0, \hat{x})$. It follows from Assumption 2.1(ii) and (3.6b) that $f(\hat{x}, y^*) \leq 0$ and $y^* \in \hat{Y}(\hat{x})$. Hence, $W(\hat{w}) \subset \hat{Y}(\hat{x})$.

Next, let $\epsilon > 0$. Then, by outer semicontinuity of $W(\cdot)$ at $\hat{w} = (0, 0, \hat{x})$, there exists $\rho > 0$ such that $W(w) \subset W(\hat{w}) + \mathbb{B}_\epsilon$ for all $w \in [0, \infty) \times T \times \mathbb{R}^n$ with $\|w - \hat{w}\|_\infty \leq \rho$. Hence, for all $\pi \geq 1/\rho$, $N \geq 1/\rho$, $N \in \mathbb{N}$, and $x \in \mathbb{B}(\hat{x}, \rho)$, $\hat{Y}_{\pi,N}(x) \subset \hat{Y}(\hat{x}) + \mathbb{B}_\epsilon$.

(ii) Using the same arguments as in (i), we obtain (ii). This completes the proof. ∎

The approximating, smooth functions in (3.1b) have the property (see [1, 14, 18]) that

$$0 \leq \omega_\pi(x, y) - \omega_{\pi,p}(x, y) \leq \frac{1}{p} \ln r \tag{3.7}$$

for all $x \in \mathbb{R}^n$, $y \in Y$, and $\pi > 0$. Hence, for all $x \in \mathbb{R}^n$ and $\pi > 0$

$$\begin{aligned} \psi_\pi(x) &= \max_{y\in Y} \omega_\pi(x, y) \\ &\leq \max_{y\in Y} \omega_{\pi,p}(x, y) + \frac{1}{p} \ln r \\ &= \psi_{\pi,p}(x) + \frac{1}{p} \ln r, \end{aligned} \tag{3.8a}$$

with

$$\psi_{\pi,p}(x) \stackrel{\triangle}{=} \max_{y\in Y} \omega_{\pi,p}(x, y). \tag{3.8b}$$

Next, it also follows from (3.7) that for all $x \in \mathbb{R}^n$ and $\pi > 0$

$$\begin{aligned} \psi(x) &= \max_{y\in Y} \omega_\pi(x, y) \\ &\geq \max_{y\in Y} \omega_{\pi,p}(x, y) \\ &= \psi_{\pi,p}(x). \end{aligned} \tag{3.8c}$$

By the same arguments leading to (3.8a) and (3.8c), we have that

$$0 \le \psi_{\pi,N}(x) - \psi_{\pi,N,p}(x) \le \frac{1}{p} \ln r \tag{3.9}$$

for all $x \in \mathbb{R}^n$, $\pi > 0$, and $N \in \mathbb{N}$.

LEMMA 3.4. *Suppose Assumptions* 2.1 *and* 3.1(ii) *hold. Then, for every bounded set $S \subset \mathbb{R}^n$ and $\pi > 0$, there exists a constant $K < \infty$ such that for all $N \ge N_0$, with N_0 as in Assumption* 3.1(ii)*, $p > 0$, and $x \in S$,*

$$0 \le \psi_\pi(x) - \psi_{\pi,N}(x) \le K\Delta(N), \tag{3.10a}$$

$$0 \le \psi_{\pi,p}(x) - \psi_{\pi,N,p}(x) \le K\Delta(N). \tag{3.10b}$$

Proof. Since $\phi_\pi^k(\cdot,\cdot), k \in \mathbf{r}$, are continuously differentiable, they are Lipschitz continuous on bounded sets. Hence, $\omega_\pi(\cdot,\cdot)$ is also Lipschitz continuous on bounded sets. First, because $Y_N \subset Y$, we always have that $\psi_{\pi,N}(x) \le \psi_\pi(x)$. Second, let $S \subset \mathbb{R}^n$ be a bounded set, and let $L < \infty$ be a Lipschitz constant for $\omega_\pi(\cdot,\cdot)$ on S. For any $x \in S$, there must exists a $y_x \in Y$ such that $\psi_\pi(x) = \omega_\pi(x, y_x)$. By Assumption 3.1(ii), there exists $y'_x \in Y_N$ such that $\|y'_x - y_x\| \le C\Delta(N)$. Hence,

$$\psi_{\pi,N}(x) \ge \omega_\pi(x, y'_x) \ge \omega_\pi(x, y_x) - LC\Delta(N) = \psi_\pi(x) - LC\Delta(N). \tag{3.10c}$$

Hence, (3.10a) holds with $K = LC$.

Next, $\omega_{\pi,p}(\cdot,\cdot)$, defined in (3.1b), has gradient with respect to y

$$\nabla_y \omega_{\pi,p}(x,y) \triangleq \sum_{k=1}^r \mu_{\pi,p}^k(x,y) \nabla_y \phi_\pi^k(x,y), \tag{3.10d}$$

where, for any $k^* \in \mathbf{r}$,

$$\mu_{\pi,p}^{k^*}(x,y) \triangleq \frac{\exp[-p\phi_\pi^{k^*}(x,y)]}{\sum_{k\in\mathbf{r}} \exp[-p\phi_\pi^k(x,y)]}. \tag{3.10e}$$

Hence, by the mean value theorem and (3.10d)–(3.10e), we have that for all $p > 0$, $x \in \mathbb{R}^n$, and $y, y' \in Y$

$$|\omega_{\pi,p}(x,y') - \omega_{\pi,p}(x,y)| \le r \sum_{k=1}^r \|\nabla_y \phi_\pi^k(x, y + s(y'-y))\| \; \|y' - y\| \tag{3.10f}$$

for some $s \in [0,1]$. Hence, $\omega_{\pi,p}(\cdot,\cdot)$ is Lipschitz continuous on bounded sets with a Lipschitz constant independent of p. The result now follows by the same arguments as for (3.10a). ∎

LEMMA 3.5. *Suppose that Assumptions* 2.1 *and* 3.1(i) *hold. Then, for every bounded set $S \subset \mathbb{R}^n$ and $\pi > 0$, there exists an $L < \infty$ such that*

$$\left\langle v, \frac{\partial^2 \omega_{\pi,p}(x,y)}{\partial x^2} v \right\rangle \le pL\|v\|^2 \tag{3.11a}$$

for all $y \in Y$, $x \in S$, $v \in \mathbb{R}^n$, and $p \ge 1$. ∎

Proof. Let $\pi > 0$ be arbitrary. By Assumption 3.1(i), $\omega_{\pi,p}(\cdot, y)$, $y \in Y$, is twice differentiable with gradient

$$\nabla_x \omega_{\pi,p}(x,y) \triangleq \sum_{k=1}^r \mu_{\pi,p}^k(x,y) \nabla_x \phi_\pi^k(x,y), \tag{3.11b}$$

where $\mu^k_{\pi,p}(x,y)$ is given by (3.10e), and Hessian matrix

$$\frac{\partial^2 \omega_{\pi,p}(x,y)}{\partial x^2} \stackrel{\triangle}{=} \sum_{k=1}^{r} \left[\nabla_x \mu^k_{\pi,p}(x,y) \nabla_x \phi^k_\pi(x,y)^T + \mu^k_{\pi,p}(x,y) \frac{\partial^2 \phi^k_\pi(x,y)}{\partial x^2} \right], \tag{3.11c}$$

where, for any $k^* \in \mathbf{r}$,

$$\nabla_x \mu^{k^*}_{\pi,p}(x,y) \stackrel{\triangle}{=} p \mu^{k^*}_{\pi,p}(x,y) \sum_{k=1}^{r} \mu^k_{\pi,p}(x,y) (\nabla_x \phi^k_\pi(x,y) - \nabla \phi^{k^*}_\pi(x,y)). \tag{3.11d}$$

Let $S \subset \mathbb{R}^n$ be bounded. Then by continuity there exists a $K < \infty$ such that $\|\nabla_x \phi^k_\pi(x,y)\| \leq K$ and $\langle v, \partial^2 \phi^k_\pi(x,y)/\partial x^2 v \rangle \leq K\|v\|^2$ for all $x \in S$, $v \in \mathbb{R}^n$, $y \in Y$, and $k \in \mathbf{r}$. Then, for all $x \in S$,

$$\|\nabla_x \mu^{k^*}_{\pi,p}(x,y)\| \leq p \sum_{k=1}^{r} \|\nabla_x \phi^k_\pi(x,y) - \nabla_x \phi^{k^*}_\pi(x,y)\| \leq 2prK. \tag{3.11e}$$

Hence, there exists $K_1 < \infty$ such that

$$\left\langle v, \sum_{k \in \mathbf{r}} \nabla_x \mu^k_{\pi,p}(x,y) \nabla_x \phi^k_\pi(x,y)^T v \right\rangle \leq pK_1 \|v\|^2 \tag{3.11f}$$

for all $x \in S$, $v \in \mathbb{R}^n$, and $y \in Y$. By inspection, $0 < \mu^k_{\pi,p}(x,y) < 1$ for all $x \in \mathbb{R}^n$, $y \in Y$, $k \in \mathbf{r}$, and $p > 0$. Hence, for $p \geq 1$, $x \in S$, $y \in Y$, and $v \in \mathbb{R}^n$

$$\begin{aligned} \left\langle v, \frac{\partial^2 \omega_{\pi,p}(x,y)}{\partial x^2} v \right\rangle &\leq pK_1 \|v\|^2 + rK\|v\|^2 \\ &\leq p(K_1 + rK)\|v\|^2. \end{aligned} \tag{3.11g}$$

Hence, $L = K_1 + rK$. This completes the proof. ∎

LEMMA 3.6. *Suppose that Assumptions* 2.1 *and* 3.1(ii) *hold and that the sequences* $\{x_i\}_{i=0}^\infty \subset \mathbb{R}^n$, $\{N_i\}_{i=0}^\infty \subset \mathbb{N}$, *and* $\{p_i\}_{i=0}^\infty \subset (0,\infty)$ *are such that* $x_i \to \hat{x}$, $p_i \to \infty$, *and* $N_i \to \infty$, *as* $i \to \infty$. *Then, for any* $\pi > 0$, $\limsup \theta_{\pi,N_i,p_i}(x_i) \leq \theta_\pi(\hat{x})$. ∎

Proof. For every i, let

$$\bar{\xi}_i \stackrel{\triangle}{=} (\xi^0_i, \xi_i) \in \underset{\bar{\xi} \in \bar{G}\psi_{\pi,N_i,p_i}(x_i)}{\arg\min} \xi^0 + \frac{1}{2}\|\xi\|^2. \tag{3.12a}$$

Then there exist multipliers $\nu^j_i \geq 0$, a set $\mathbf{J}_i \subset \mathbb{N}$, and $y_{i,j} \in Y_{N_i}$ such that $\sum_{j \in \mathbf{J}_i} \nu^j_i = 1$,

$$\xi^0_i = \sum_{j \in \mathbf{J}_i} \nu^j_i [\psi_{\pi,N_i,p_i}(x_i) - \omega_{\pi,p_i}(x_i, y_{i,j})], \tag{3.12b}$$

and

$$\xi_i = \sum_{j \in \mathbf{J}_i} \nu^j_i \nabla_x \omega_{\pi,p_i}(x_i, y_{i,j}). \tag{3.12c}$$

In view of (3.10e), we have that all $\mu^k_{\pi,p_i}(x_i, y_{i,j}) \geq 0$, and $\sum_{k\in\mathbf{r}} \mu^k_{\pi,p_i}(x_i, y_{i,j}) = 1$. Hence, we obtain the following expression for $\bar{\xi}_i$:

$$\bar{\xi}_i = \sum_{j\in\mathbf{J}_i} \nu_i^j \sum_{k\in\mathbf{r}} \mu^k_{\pi,p_i}(x_i, y_{i,j}) \begin{pmatrix} \psi_{\pi,N_i,p_i}(x_i) - \omega_{\pi,p_i}(x_i, y_{i,j}) \\ \nabla_x \phi^k_\pi(x_i, y_{i,j}) \end{pmatrix}. \tag{3.12d}$$

Now, let

$$\bar{\zeta} \stackrel{\triangle}{=} (\zeta^{-1}, \zeta^0, \zeta) = \sum_{j\in\mathbf{J}_i} \nu_i^j \sum_{k\in\mathbf{r}} \mu^k_{\pi,p_i}(x_i, y_{i,j}) \begin{pmatrix} \phi^k_\pi(x_i, y_{i,j}) - \omega_\pi(x_i, y_{i,j}) \\ \psi_\pi(x_i) - \omega_\pi(x_i, y_{i,j}) \\ \nabla_x \phi^k_\pi(x_i, y_{i,j}) \end{pmatrix}. \tag{3.12e}$$

Then by inspection $\bar{\zeta} \in \bar{G}\psi_\pi(x_p)$, and hence

$$\begin{aligned} -\theta_\pi(x_i) &\leq \zeta^{-1} + \zeta^0 + \frac{1}{2}\|\zeta\|^2 \\ &= \sum_{j\in\mathbf{J}_i} \nu_i^j \{[\psi_\pi(x_i) - \psi_{\pi,N_i,p_i}(x_i)] - [\omega_\pi(x_i, y_{i,j}) - \omega_{\pi,p_i}(x_i, y_{i,j})]\} \\ &\quad + \zeta^{-1} - \theta_{\pi,N_i,p_i}(x_i). \end{aligned} \tag{3.12f}$$

Now, for any $j \in \mathbf{J}_i$ and $k* \in \mathbf{r}$,

(3.12g)

$$\begin{aligned} \mu^{k*}_{\pi,p_i}(x_i, y_{i,j})[\phi^{k*}_\pi(x_i, y_{i,j}) - \omega_\pi(x_i, y_{i,j})] &= \frac{\phi^{k*}_\pi(x_i, y_{i,j}) - \omega_\pi(x_i, y_{i,j})}{\sum_{k\in\mathbf{r}} \exp\{p_i[(\phi^{k*}_\pi(x_i, y_{i,j}) - \phi^k_\pi(x_i, y_{i,j})]\}} \\ &\leq \frac{\phi^{k*}_\pi(x_i, y_{i,j}) - \omega_\pi(x_i, y_{i,j})}{\exp\{p_i[\phi^{k*}_\pi(x_i, y_{i,j}) - \omega_\pi(x_i, y_{i,j})]\}} \\ &\leq \frac{1}{p_i[\exp(1)]}. \end{aligned}$$

It now follows from (3.7), (3.9), (3.10a), (3.12g), and (3.12f) that

$$-\theta_\pi(x_i) \leq -\theta_{\pi,N_i,p_i}(x_i) + \frac{r}{p_i[\exp(1)]} + \frac{1}{p_i}\ln r + K\Delta(N_i), \tag{3.12h}$$

with $K < \infty$ as in (3.10a). By continuity of $\theta_\pi(\cdot)$ and (3.12h), the result follows. ∎

4. Algorithm for P. In view of Theorem 2.5, **P** can be solved by solving $\mathbf{P}_\pi$ for a sufficiency large penalty $\pi > 0$. However, a priori the size of the penalty is unknown. Hence, in the following algorithm we use the test function defined in (2.13h) to control the penalty π.

As shown in section 3, $\psi_\pi(\cdot)$ can be approximated by $\psi_{\pi,N,p}(\cdot)$. Hence, $\mathbf{P}_\pi$ can be approximately solved by solving $\mathbf{P}_{\pi,N,p}$. The following algorithm adaptively increases the precision parameters N and p, based on a series of tests, such that for all $x \in \mathbb{R}^n$, $\psi_{\pi,N,p}(x)$ converges to $\psi_\pi(x)$. For given π, N, and p, the algorithm calls the Pironneau–Polak–Pshenichnyi min-max algorithm (see [17, 20]) as a subroutine to perform one iteration on $\mathbf{P}_{\pi,N,p}$.

In the algorithm below, let $\Delta : \mathbb{N} \to \mathbb{R}$, $N_0 \in \mathbb{N}$, be as in Assumption 3.1, and let $Y_N \subset Y, N \geq N_0$ be the finite-cardinality subsets of Y in the definition of $\mathbf{P}_{\pi,N,p}$.

ALGORITHM 4.1.

Parameters. $\alpha, \beta, \mu, \rho \in (0,1)$; $\tau_1 \geq \ln(r_1+1)$; $\sigma_{-1}, \tau_2 > 0$; $\sigma, \kappa > 1$; $\gamma \gg 1$; $\pi_{-1} > 0$; $p_0 \geq 1$; $\hat{p} \geq p_0$, $\hat{p} \gg 1$; $\zeta \in \mathbb{N}$, $\zeta \geq 2$; $N_0 \in \mathbb{N}$.

Data. $x_0 \in \mathbb{R}^n$.

Step 0. Set $i = 0$, $j = 0$, $k = 0$, and $\delta = 1$.

Step 1. Compute $y_i \in \hat{Y}_{\pi_{i-1},N_i}(x_i)$, and the smallest eigenvalue $\sigma_{\min}(x_i, y_i)$ of the matrix $[A(x_i,y_i)A(x_i,y_i)^T + B(x_i,y_i)]$ (see (2.13a), (2.13d)), and set $\pi = \pi_{i-1}$.

Step 2. **If** $\sigma_{\min}(x_i, y_i) \geq \sigma_{i-1}$, set $\sigma_i = \sigma_{i-1}$, and go to **Step 3**.
Else, set $\sigma_i = \mu\sigma_{i-1}$, and go to **Step 10**.

Step 3. **If** $t_\pi(x_i, y_i) \leq 0$ (see (2.13h)), set $\pi_i = \pi$, and go to **Step 4**.
Else, go to **Step 10**.

Step 4. Compute $\theta_{\pi_i,N_i,p_i}(x_i)$ and the augmented search direction (see (3.2b)),

$$(h^0_{\pi_i,N_i,p_i}(x_i), h_{\pi_i,N_i,p_i}(x_i)) = -\operatorname*{arg\,min}_{\bar{\xi} \in \bar{G}\psi_{\pi_i,N_i,p_i}(x_i)} \xi^0 + \frac{1}{2}\|\xi\|^2. \tag{4.1a}$$

Step 5. Compute $x_{i+1} = x_i + \lambda_{\pi_i,N_i,p_i}(x_i) h_{\pi_i,N_i,p_i}(x_i)$, where the Armijo step size

$$\lambda_{\pi_i,N_i,p_i}(x_i) = \max_{s \in \mathbb{N}} \{\beta^s \mid \psi_{\pi_i,N_i,p_i}(x_i + \beta^s h_{\pi_i,N_i,p_i}(x_i)) - \psi_{\pi_i,N_i,p_i}(x_i) \leq \alpha\beta^s \theta_{\pi_i,N_i,p_i}(x_i)\}. \tag{4.1b}$$

Step 6. **If**

$$\Delta\psi_i \stackrel{\triangle}{=} \psi_{\pi_i,N_i,p_i}(x_{i+1}) - \psi_{\pi_i,N_i,p_i}(x_i) \geq -\frac{\tau_1}{p_i} - \tau_2\Delta(N_i), \tag{4.1c}$$

go to **Step 7**.
Else, set $N_{i+1} = N_i$ and $p_{i+1} = p_i$, replace i by $i+1$, and go to **Step 1**.

Step 7. Set $N_{i+1} \in \mathbb{N}$ equal to the smallest integer satisfying

$$\Delta(N_{i+1}) \leq \min\left\{\max\left\{\frac{1-\rho}{\tau_2}|\Delta\psi_i|,\ \Delta(\zeta N_i)\right\},\ \frac{\gamma-1}{\gamma}\Delta(N_i)\right\}. \tag{4.1d}$$

Step 8. **If** (initial stage)

$$\max\left\{\frac{\tau_1}{\rho|\Delta\psi_i|},\ \frac{\gamma p_i}{\gamma-1}\right\} \leq \hat{p} \ \text{ and } \ \delta = 1, \tag{4.1e}$$

set $p_{i+1} = \max\{\tau_1/(\rho|\Delta\psi_i|)\ ,\ \gamma p_i/(\gamma-1)\}$, replace k by $k+1$, i by $i+1$, and go to **Step 1**.
Elseif (switch stage)

$$\max\left\{\frac{\tau_1}{\rho|\Delta\psi_i|},\ \frac{\gamma p_i}{\gamma-1}\right\} > \hat{p} \ \text{ and } \ \delta = 1, \tag{4.1f}$$

set $\delta = \max\{2\ ,\ \gamma\hat{p}/((\gamma-1)(k+1))\}$, $p_{i+1} = \delta(k+2)$, replace k by $k+1$, i by $i+1$, and go to **Step 1**.
Else (final stage) go to **Step 9**.

Step 9. **If** $\delta(i+2) < \gamma p_i/(\gamma-1)$, set $p_{i+1} = p_i$, replace i by $i+1$, and go to **Step 1**.
Else find the smallest $k^* \in \mathbb{N}$ such that $k \leq k^* \leq i$ and $\delta(k^*+2) \geq \gamma p_i/(\gamma-1)$, and set $p_{i+1} = \delta(k^*+2)$, replace k by k^*+1, i by $i+1$, and go to **Step 1**.

Step 10. Set $x^*_j = x_i$, $\pi = \kappa^{j+1}\pi_{-1}$, replace j by $j+1$, and go to **Step 3**. ∎

LEMMA 4.2. *Suppose that Algorithm* 4.1 *has generated a sequence* $\{p_i\}_{i=0}^{\infty}$. *Then the following hold:*

(i) *If the test in* (4.1c) *is satisfied an infinite number of times, then there exists an* $i^* \in \mathbb{N}$ *such that* p_{i^*+1} *is set in the "switch stage" of Step* 8.

(ii) *If there exists an* $i^* \in \mathbb{N}$ *such that* p_{i^*+1} *is set in the "switch stage" of Step* 8, *then for all* $i < i^*$ *such that* (4.1c) *is satisfied,* p_{i+1} *is set in the "initial stage," and for all* $i > i^*$ *such that* (4.1c) *is satisfied,* p_{i+1} *is set in the "final stage."*

Proof. (ii) Suppose that there exists an $i^* \in \mathbb{N}$ such that p_{i^*+1} is set in the "switch stage." Then Algorithm 4.1 sets $\delta = \max\{2, \gamma\hat{p}/((\gamma-1)(k+1))\} \geq 2 > 1$ in iteration i^*. Hence, (4.1e)–(4.1f) cannot hold for $i > i^*$, and p_{i+1} must be set in the "final stage" of Step 8 for all $i > i^*$ such that (4.1c) holds. Hence, p_{i+1} must be set in the "initial stage" of Step 8 for all $i < i^*$ such that (4.1c) holds.

(i) Suppose for the sake of a contradiction that for all $i \in \mathbb{N}$, p_{i+1} is not set in the "switch stage" of Step 8. Then $\delta = 1$ for all $i \in \mathbb{N}$ because $\delta = 1$ for $i = 0$, and δ is only changed in the "switch stage" of Step 8. Hence, because $\delta = 1$ for all $i \in \mathbb{N}$ and the hypothesis that p_{i+1} is not set in the "switch stage," p_{i+1} is set by the "initial stage" of Step 8 for all $i \in \mathbb{N}$ such that (4.1c) is satisfied. Hence, $p_{i+1} \geq \gamma p_i/(\gamma-1)$ for all $i \in \mathbb{N}$ such that (4.1c) is satisfied. Since $p_{i+1} \geq \gamma p_i/(\gamma-1)$ an infinite number of times, there must exists an $i^{**} \in \mathbb{N}$ such that $\max\{\tau_1/(\rho|\Delta\psi_{i^{**}}|),\ \gamma p_{i^{**}}/(\gamma-1)\} > \hat{p}$; see the "initial stage" of Step 8. Hence, (4.1f) is satisfied for $i = i^{**}$, but (4.1e) is not. This is a contradiction, which completes the proof. ∎

The mechanisms in Algorithm 4.1 can by described as follows. Step 2 ensures that the linear independence property of Assumption 2.6 is eventually satisfied at (x_i, y_i), and Step 3 ensures that the test function remains nonpositive. In view of Lemma 2.7(iii), we see that Steps 2 and 3 increase the penalty π to a sufficiently large value that ensures local equivalence between $\mathbf{P}$ and $\mathbf{P}_\pi$.

Suppose that π^* is sufficiently large; i.e., there exists an $i^* \in \mathbb{N}$ such that $\pi_i = \pi^*$ for all $i > i^*$. Then Algorithm 4.1 solves the sequence of approximating problem $\{\mathbf{P}_{\pi^*,N_i,p_i}\}_{i=i*}^{\infty}$, associated with a sequence of monotonically increasing precision parameters N_i, p_i that diverge to infinity. At a given precision level, N', p', say, Algorithm 4.1 computes iterates that approach a stationary point of the approximating problem $\mathbf{P}_{\pi^*,N',p'}$. When the current iterate is sufficiently close to a stationary point for $\mathbf{P}_{\pi^*,N',p'}$, as determined by the test in (4.1c), the precision level is increased from N', p' to N'', p'', say. Algorithm 4.1 then continues by computing iterates that are approaching a stationary point of $\mathbf{P}_{\pi^*,N'',p''}$ until the test in (4.1c) again determines that the precision level has to be increased. The last iteration of the previous precision level is used as a "warm start" for calculations on the next precision level.

It becomes gradually harder and harder to satisfy (4.1c) as $N_i, p_i \to \infty$. Hence, as the precision level is increased, the iterates generated by Algorithm 4.1 gradually get closer and closer to a stationary point of the current approximating problem before the precision level is increased. Thus, Algorithm 4.1 computes approximate solutions to a sequence of approximating problems $\{\mathbf{P}_{\pi^*,N_i,p_i}\}_{i=i^*}^{\infty}$ with higher and higher precision as the number of iterations increases.

The sequences of precision parameters $\{N_i\}_{i=0}^{\infty}$ and $\{p_i\}_{i=0}^{\infty}$ are not determined a priori but are constructed by Algorithm 4.1. When (4.1c) is satisfied, the precision level is increased by an amount determined by Steps 7, 8, and 9.

In the "early" iterations, i.e., before the test in (4.1e) fails, the smoothing precision parameter is increased by an amount related to the value of the cost-decrease $\Delta\psi_i$. When $|\Delta\psi_i|$ is large, p_{i+1} tends to be only marginally larger than p_i, with a minimum

increase of $p_i/(\gamma - 1)$. On the other hand, when $|\Delta\psi_i|$ is small, p_{i+1} tends to be augmented by a considerable amount.

When the test in (4.1e) fails, δ is set to be larger than 1 in the "switch stage" of Step 8. Hence, for all subsequent iterations, the increase of the precision parameter will be determined by the "final stage," i.e., Step 8, of Algorithm 4.1. In the "final stage," the precision parameter is augmented by a multiple of δ whenever $p_{i+1} > p_i$.

The increase of the precision parameters N_i, p_i are motivated by the following considerations: (i) Suppose that the algorithm parameter $\tau_1 = \ln(r_1+1)$, where $r_1 \in \mathbb{N}$ is as in (2.3b), $\tau_2 = K$, where $K < \infty$ is as in Lemma 3.4, $\psi(x_{i+1}) = \psi_{\pi^*}(x_{i+1})$, and $\psi(x_i) = \psi_{\pi^*}(x_i)$; then we have by (3.8a), (3.8c), (3.9), and Lemma 3.4 that

$$\psi(x_{i+1}) - \psi(x_i) \leq \Delta\psi_i + \frac{\tau_1}{p_i} + \tau_2\Delta(N_i). \tag{4.2}$$

Hence, $\psi(x_{i+1}) - \psi(x_i) < 0$ whenever the test in (4.1c) fails; i.e., the precision is not increased as long as the new iterate guarantees a decrease in the cost function $\psi(\cdot)$. The constant K in Lemma 3.4 may seldom be known. In absence of any information about K, we recommend setting $\tau_2 = 1$. Note that larger values for τ_2 will drive N to infinity faster. (ii) When (4.1c) is satisfied, we can no longer guarantee that $\psi(x_{i+1}) - \psi(x_i) < 0$, and we set N_{i+1} and p_{i+1} to be larger than N_i and p_i, which, hopefully, will ensure that $\psi(x_{i+2}) - \psi(x_{i+1}) < 0$ will hold. (iii) If the current iterate is very close to a stationary point of the approximating problem, $|\Delta\psi_i|$ tends to become extremely small. Hence, the factor ζ (see (4.1d)) and the fixed increase of p_i in the "final stage" is introduced to prevent N_i, p_i from becoming very large prematurely. (iv) Lemma 4.3 below must hold.

Let $t > 0$ be the desired tolerance on the solution. Then every $p_i \gg \ln(r_1+1)/t$ is associated with an error (see (3.7)) $\ln(r_1+1)/p_i \ll t$. Hence, we recommend that the algorithm parameter $\hat{p}$, used to decide when to switch from the "initial stage" to the "final stage," be set equal to $\ln(r_1+1)/t$. Furthermore, we recommend to set γ equal to a large number, e.g., 10^5, to avoid any practical influence on the determination of p_{i+1}.

The parameter $\rho \in (0, 1)$ controls how the error associated with the discretization of Y compares with the error associated with the smoothing of $\omega_\pi(\cdot, \cdot)$. When ρ is close to unity, the error associated with the discretization tends to be "small" and the error associated with smoothing tends to be "large." When ρ is close to zero, the situation is reversed. Since a fine discretization implies a high computational cost, it can be efficient to bias the approximation error towards the smoothing error by selecting ρ close to 0.

Algorithm 4.1 is quite insensitive to the selection of the parameters $\sigma_{-1} > 0$ and $\mu \in (0, 1)$ used in Step 2. However, note that larger values of σ_{-1} and μ will cause the penalty π to increase faster. We recommend setting $\sigma_{-1} = 10^{-5}$ and $\mu = 0.5$.

LEMMA 4.3. *Suppose that Assumption* 2.1 *holds and that the sequences* $\{x_i\}_{i=0}^\infty$, $\{N_i\}_{i=0}^\infty$, *and* $\{p_i\}_{i=0}^\infty$ *are generated by Algorithm* 4.1. *Then the following hold:*

(i) *The sequences* $\{N_i\}_{i=0}^\infty$ *and* $\{p_i\}_{i=0}^\infty$ *are monotonically increasing, and, if* $p_{i+1} > p_i$, *then* $p_{i+1} \geq \gamma p_i/(\gamma - 1)$, *and, if* $N_{i+1} > N_i$, *then* $\Delta(N_{i+1}) \leq (\gamma - 1)\Delta(N_i)/\gamma$, *with* γ *as in Algorithm* 4.1.

(ii) *If* $\{x_i\}_{i=0}^\infty$ *has an accumulation point, then* $N_i \to \infty$, $p_i \to \infty$, *and* $\sum_{i=0}^\infty 1/p_i = \infty$.

Proof. (i) If the test in (4.1c) fails, then $N_{i+1} = N_i$ and $p_{i+1} = p_i$. If the test in (4.1c) is satisfied, then, according to Step 7 of Algorithm 4.1 (see (4.1d)), $\Delta(N_{i+1}) \leq (\gamma - 1)\Delta(N_i)/\gamma$. Next, consider the construction of $\{p_i\}_{i=0}^\infty$. If the test

in (4.1c) is satisfied, then we have three cases corresponding to the "initial," "switch," and "final" stages of Step 8.

Case I. Suppose that p_{i+1} is defined as in the "initial stage" of Step 8 in Algorithm 4.1. Then $p_{i+1} \geq \gamma p_i/(\gamma - 1)$.

Case II. Suppose that p_{i+1} is defined as in the "switch stage" of Step 8 in Algorithm 4.1. Then

$$\begin{aligned} p_{i+1} &= \max\left\{2, \frac{\gamma}{\gamma-1}\frac{\hat{p}}{k+1}\right\}(k+2) \\ &\geq \frac{\gamma}{\gamma-1}\hat{p}. \end{aligned} \tag{4.3a}$$

If $i > 0$, then, by Lemma 4.2(ii), p_i was constructed according to the "initial stage" of Step 8. Hence, it follows from the definition of p_i and (4.1e) that $p_i \leq \hat{p}$. Hence, by (4.3a) we have that $p_{i+1} \geq \gamma p_i/(\gamma - 1)$. If $i = 0$, then $p_{i+1} \geq \gamma p_i/(\gamma - 1)$ because $\hat{p} \geq p_0$.

Case III. Suppose that p_{i+1} is defined as in the "final stage" of Step 8; see Step 9. Then $p_{i+1} \geq \gamma p_i/(\gamma - 1)$ whenever $p_{i+1} > p_i$. Hence, (i) holds.

(ii) Suppose that Algorithm 4.1 has generated the sequence $\{x_i\}_{i=0}^{\infty}$ with accumulation point $\hat{x}$ and that at least one of the sequences $\{N_i\}_{i=0}^{\infty}$, $\{p_i\}_{i=0}^{\infty}$ are bounded from above. Now, we have three cases.

Case I. Suppose that both $\{N_i\}_{i=0}^{\infty}$ and $\{p_i\}_{i=0}^{\infty}$ are bounded. Then the test in (4.1c) can only be satisfied a finite number of times, because otherwise (4.1d) would have caused $\{N_i\}_{i=0}^{\infty}$ to diverge to infinity. Hence, there must exist an $i^* \in \mathbb{N}$, an $N^* < \infty$, and a $p^* < \infty$ such that for all $i > i^*$, $N_i = N^*$, $p_i = p^*$, and

$$\psi_{\pi_i,N^*,p^*}(x_{i+1}) - \psi_{\pi_i,N^*,p^*}(x_i) < -\frac{\tau_1}{p^*} - \tau_2\Delta(N^*). \tag{4.3b}$$

By inspection, $\psi_{\pi'',N}(x) - \psi_{\pi',N}(x) \leq 0$ for all $\pi'' \geq \pi'$, $N \in \mathbb{N}$, and $x \in \mathbb{R}^n$. Hence, by (3.9) and (4.3b), we have that for all $i > i^*$

$$\begin{aligned} &\psi_{\pi_{i+1},N^*}(x_{i+1}) - \psi_{\pi_i,N^*}(x_i) \\ &= \psi_{\pi_{i+1},N^*}(x_{i+1}) - \psi_{\pi_i,N^*}(x_{i+1}) + \psi_{\pi_i,N^*}(x_{i+1}) - \psi_{\pi_i,N^*}(x_i) \\ &\leq 0 + \psi_{\pi_i,N^*,p^*}(x_{i+1}) - \psi_{\pi_i,N^*,p^*}(x_i) + \frac{1}{p^*}\ln(r_1+1) \\ &< -\frac{\tau_1}{p^*} - \tau_2\Delta(N^*) + \frac{1}{p^*}\ln(r_1+1) \\ &\leq -\tau_2\Delta(N^*). \end{aligned} \tag{4.3c}$$

Thus, $\psi_{\pi_i,N^*}(x_i) \to -\infty$, as $i \to \infty$. However, there exists an infinite subset $K \subset \mathbb{N}$ such that $x_i \to^K \hat{x}$, as $i \to \infty$. If $\{\pi_i\}_{i=0}^{\infty}$ is bounded, then there exists an $i^{**} \geq i^*$ such that $\pi_i = \pi^*$ for all $i > i^{**}$, and hence by continuity, $\psi_{\pi^*,N^*}(x_i) \to^K \psi_{\pi^*,N^*}(x^*)$, as $i \to \infty$. If $\pi_i \to \infty$, then we can infer from Lemma 3.3(ii) that $\psi_{\pi_i,N^*}(x_i) \to^K \psi_{N^*}(x^*)$, as $i \to \infty$. This is a contradiction.

Case II. Suppose that $\{N_i\}_{i=0}^{\infty}$ is bounded, but $\{p_i\}_{i=0}^{\infty}$ diverges to infinity. Then the test in (4.1c) can only be satisfied a finite number of times, because otherwise (4.1d) would have caused $\{N_i\}_{i=0}^{\infty}$ to diverge to infinity. Since $p_{i+1} = p_i$ whenever the test in (4.1c) fails, it follows that $p_{i+1} > p_i$ only a finite number of times. Hence, $\{p_i\}_{i=0}^{\infty}$ has to be bounded, which is a contradiction.

Case III. Suppose that $\{N_i\}_{i=0}^{\infty}$ diverges to infinity, but $\{p_i\}_{i=0}^{\infty}$ is bounded from above. Then the test in (4.1c) must be satisfied an infinite number of times, because otherwise $\{N_i\}_{i=0}^{\infty}$ would not have diverged to infinity. Hence, Algorithm 4.1 enters

Step 8 an infinite number of times. By Lemma 4.2, there exists an $i^* \in \mathbb{N}$ such that p_{i+1} is set by the "final stage" for all $i > i^*$ such that (4.1c) is satisfied. Since $\{p_i\}_{i=0}^{\infty}$ is bounded from above and Step 8 is entered a infinite number of times, we must have that (see Step 9)

$$\delta(i+2) < \frac{\gamma p_i}{\gamma - 1} \tag{4.3d}$$

for an infinite number of iterations. However, since there exists an $p^* < \infty$ such that $p_i \leq p^*$ for all $i \in \mathbb{N}$, (4.3d) cannot be satisfied for an infinite number of iterations, which is a contradiction.

Hence, $N_i \to \infty$ and $p_i \to \infty$, as $i \to \infty$. Next, we prove that $\sum_{i=0}^{\infty} 1/p_i = \infty$. Since $p_{i+1} > p_i$ only if (4.1c) is satisfied, and $p_i \to \infty$, as $i \to \infty$, the test in (4.1c) must be satisfied an infinite number of times. Hence, by Lemma 4.2, there exists an $i^* \in \mathbb{N}$ such that for all $i > i^*$, p_{i+1} is set by the "final stage" of Step 8 whenever (4.1c) is satisfied. Hence, for all $i > i^*$, $p_{i+1} = p_i$ or $p_{i+1} \leq \delta(i+2)$; see Step 9. The final result now follows from the fact that $\sum_{i=0}^{\infty} 1/i = +\infty$. ∎

LEMMA 4.4. *Suppose that Assumptions* 2.1 *and* 3.1(i) *hold. For every bounded set* $S \subset \mathbb{R}^n$, $\pi > 0$, *and* $\alpha, \beta \in (0,1)$, *there exists a* $K_S < \infty$ *such that for all* $p \geq 1$, $N \in \mathbb{N}$, *and* $x \in S$

$$\psi_{\pi,N,p}(x + \lambda_{\pi,N,p}(x) h_{\pi,N,p}(x)) - \psi_{\pi,N,p}(x) \leq \alpha \frac{K_S}{p} \theta_{\pi,N,p}(x), \tag{4.4a}$$

where $\lambda_{\pi,N,p}(x)$ *and* $h_{\pi,N,p}(x)$ *are the step size and search direction of Algorithm* 4.1*; see* (4.1a,b).

Proof. In section 2.1 in [17], we find the following equivalent form of $\theta_{\pi,N,p}(\cdot)$, see (3.2b):

$$\theta_{\pi,N,p}(x) = \min_{h \in \mathbb{R}^n} \max_{y \in Y_N} \ \omega_{\pi,p}(x,y) - \psi_{\pi,N,p}(x) + \langle \nabla_x \omega_{\pi,p}(x,y), h \rangle + \frac{1}{2}\|h\|^2. \tag{4.4b}$$

Let $S \subset \mathbb{R}^n$ be bounded. It follows from Assumption 2.1 and (3.2b)–(3.2c) that there exists a constant $M < \infty$ such that $\|h_{\pi,N,p}(x)\| \leq M$ for all $x \in S$, $N \in \mathbb{N}$, and $p > 0$. Next, let $S_B \subset \mathbb{R}^n \triangleq \{x \in \mathbb{R}^n \mid \|x - x'\| \leq M, x' \in S\}$, and let $L \in [1, \infty)$ be the constant corresponding to S_B such that (3.11a) holds for all $x \in S_B$, $y \in Y$, $v \in \mathbb{R}^n$, and $p \geq 1$. Then, for all $\lambda \in (0,1]$, $x \in S$, $N \in \mathbb{N}$, and $p \geq 1$, we have by expansion, Lemma 3.5, and (4.4b) that for some $s \in [0,1]$

(4.4c)

$$\begin{aligned}
&\psi_{\pi,N,p}(x + \lambda h_{\pi,N,p}(x)) - \psi_{\pi,N,p}(x) \\
&= \max_{y \in Y_N} \{\omega_{\pi,p}(x + \lambda h_{\pi,N,p}(x), y) - \psi_{\pi,N,p}(x)\} \\
&= \max_{y \in Y_N} \Bigg\{ \omega_{\pi,p}(x,y) - \psi_{\pi,N,p}(x) \\
&\qquad + \lambda \langle \nabla_x \omega_{\pi,p}(x,y), h_{\pi,N,p}(x) \rangle \\
&\qquad + \frac{\lambda^2}{2} \left\langle h_{\pi,N,p}(x), \frac{\partial^2 \omega_{\pi,p}(x + s\lambda h_{\pi,N,p}(x), y)}{\partial x^2} h_{\pi,N,p}(x) \right\rangle \Bigg\} \\
&\leq \lambda \max_{y \in Y_N} \left\{ \omega_{\pi,p}(x,y) - \psi_{\pi,N,p}(x) + \langle \nabla_x \omega_{\pi,p}(x,y), h_{\pi,N,p}(x) \rangle + \frac{\lambda}{2} pL \|h_{\pi,N,p}(x)\|^2 \right\} \\
&= \lambda \left(\theta_{\pi,N,p}(x) + \frac{1}{2}(\lambda pL - 1)\|h_{\pi,N,p}(x)\|^2 \right).
\end{aligned}$$

Hence, for all $\lambda \in (0, 1/(pL)]$

$$\begin{aligned} &\psi_{\pi,N,p}(x + \lambda h_{\pi,N,p}(x)) - \psi_{\pi,N,p}(x) - \alpha\lambda\theta_{\pi,N,p}(x) \\ &\leq \lambda(1-\alpha)\theta_{\pi,N,p}(x) \leq 0. \end{aligned} \tag{4.4d}$$

Now, it follows from (4.4d) and the step-size rule in (4.1b) that

$$\lambda_{\pi,N,p}(x) \geq \frac{\beta}{pL} \tag{4.4e}$$

for all $x \in S$, $N \in \mathbb{N}$, and $p \geq 1$. Hence, the conclusion follows with $K_S = \beta/L$. This completes the proof. ∎

THEOREM 4.5. *Suppose that Assumptions* 2.1 *and* 3.1 *hold and that Algorithm* 4.1 *has generated a bounded sequence* $\{x_i\}_{i=0}^{\infty}$ *and a finite sequence* $\{x_j^*\}_{j=0}^{j^*}$. *Then there exist an infinite subset* $K \subset \mathbb{N}$ *and an* $\hat{x} \in \mathbb{R}^n$ *such that* $x_i \to^K \hat{x}$ *and* $\theta_{\pi^*}(\hat{x}) = 0$, *where* $\pi^* = \kappa^{j^*+1}\pi_{-1}$, *with* κ, π_{-1} *as in Algorithm* 4.1.

Proof. Since $\{x_j^*\}_{j=0}^{j^*}$ is a finite sequence, there exists an $i^* \in \mathbb{N}$ such that $\pi_i = \pi^* \triangleq \kappa^{j^*+1}\pi_{-1}$ for all $i > i^*$. For the sake of a contradiction, suppose that there exists an $\epsilon > 0$ such that

$$\limsup_{i\to\infty} \theta_{\pi^*,N_i,p_i}(x_i) \leq -\epsilon. \tag{4.5a}$$

Since $\{x_i\}_{i=0}^{\infty}$ is a bounded sequence, it has at least one accumulation point. Hence, by Lemma 4.3(ii), $p_i, N_i \to \infty$, as $i \to \infty$. Next, by Lemma 4.4 there exists an $M < \infty$ such that

$$\psi_{\pi^*,N_i,p_i}(x_{i+1}) - \psi_{\pi^*,N_i,p_i}(x_i) \leq \alpha\frac{M}{p_i}\theta_{\pi^*,N_i,p_i}(x_i) \tag{4.5b}$$

for all $i > i^*$. Now, for all $N \in \mathbb{N}$ and $p > 0$, let

$$\tilde{\psi}_{\pi^*,N,p}(x) \triangleq \psi_{\pi^*,N,p}(x) + \frac{\gamma}{p}\ln r + \gamma K\Delta(N), \tag{4.5c}$$

where $\gamma > 1$ is as in Algorithm 4.1 and $K < \infty$ as in Lemma 3.4. Now, we have three cases corresponding to whether p and N were increased or not in Steps 7 and 8 of Algorithm 4.1.

Case I. Suppose that $p_i < p_{i+1}$ and $N_i < N_{i+1}$. Then by Lemma 4.3(i)

$$p_{i+1} \geq \frac{\gamma}{\gamma - 1}p_i, \tag{4.5d}$$

$$\Delta(N_{i+1}) \leq \frac{\gamma - 1}{\gamma}\Delta(N_i), \tag{4.5e}$$

and we have that for all $i > i^*$

$$\begin{aligned}
&\tilde{\psi}_{\pi^*,N_{i+1},p_{i+1}}(x_{i+1}) - \tilde{\psi}_{\pi^*,N_i,p_i}(x_i) \\
&= \psi_{\pi^*,N_{i+1},p_{i+1}}(x_{i+1}) - \psi_{\pi^*,N_i,p_i}(x_i) \\
&\quad + \left(\frac{\gamma}{p_{i+1}} - \frac{\gamma}{p_i}\right)\ln r + \gamma K(\Delta(N_{i+1}) - \Delta(N_i)) \\
&= \psi_{\pi^*,N_{i+1},p_{i+1}}(x_{i+1}) - \psi_{\pi^*,N_i,p_i}(x_{i+1}) + \psi_{\pi^*,N_i,p_i}(x_{i+1}) - \psi_{\pi^*,N_i,p_i}(x_i) \\
&\quad + \left(\frac{\gamma}{p_{i+1}} - \frac{\gamma}{p_i}\right)\ln r + \gamma K(\Delta(N_{i+1}) - \Delta(N_i)) \\
&\leq \psi_{\pi^*,N_{i+1}}(x_{i+1}) - \psi_{\pi^*,N_i}(x_{i+1}) + \frac{1}{p_i}\ln r + \alpha\frac{M}{p_i}\theta_{\pi^*,N_i,p_i}(x_i) \\
&\quad + \left(\frac{\gamma}{p_{i+1}} - \frac{\gamma}{p_i}\right)\ln r + \gamma K(\Delta(N_{i+1}) - \Delta(N_i)) \\
&\leq \psi_{\pi^*}(x_{i+1}) - \psi_{\pi^*}(x_{i+1}) + \frac{1}{p_i}\ln r + K\Delta(N_i) + \alpha\frac{M}{p_i}\theta_{\pi^*,N_i,p_i}(x_i) \\
&\quad + \left(\frac{\gamma}{p_{i+1}} - \frac{\gamma}{p_i}\right)\ln r + \gamma K(\Delta(N_{i+1}) - \Delta(N_i)) \\
&\leq \alpha\frac{M}{p_i}\theta_{\pi^*,N_i,p_i}(x_i).
\end{aligned} \tag{4.5f}$$

Case II. Suppose that $p_i = p_{i+1}$ and $N_i < N_{i+1}$. Then (4.5e) holds, and we have that for all $i > i^*$

$$\begin{aligned}
&\tilde{\psi}_{\pi^*,N_{i+1},p_{i+1}}(x_{i+1}) - \tilde{\psi}_{\pi^*,N_i,p_i}(x_i) \\
&= \psi_{\pi^*,N_{i+1},p_i}(x_{i+1}) - \psi_{\pi^*,N_i,p_i}(x_i) + \gamma K(\Delta(N_{i+1}) - \Delta(N_i)) \\
&= \psi_{\pi^*,N_{i+1},p_i}(x_{i+1}) - \psi_{\pi^*,N_i,p_i}(x_{i+1}) + \psi_{\pi^*,N_i,p_i}(x_{i+1}) - \psi_{\pi^*,N_i,p_i}(x_i) \\
&\quad + \gamma K(\Delta(N_{i+1}) - \Delta(N_i)) \\
&\leq \psi_{\pi^*,p_i}(x_{i+1}) - \psi_{\pi^*,p_i}(x_{i+1}) + K\Delta(N_i) + \alpha\frac{M}{p_i}\theta_{\pi^*,N_i,p_i}(x_i) \\
&\quad + \gamma K(\Delta(N_{i+1}) - \Delta(N_i)) \\
&\leq \alpha\frac{M}{p_i}\theta_{\pi^*,N_i,p_i}(x_i).
\end{aligned} \tag{4.5g}$$

Case III. Suppose that $p_i = p_{i+1}$ and $N_i = N_{i+1}$. Then we have that for all $i > i^*$

$$\begin{aligned}
&\tilde{\psi}_{\pi^*,N_{i+1},p_{i+1}}(x_{i+1}) - \tilde{\psi}_{\pi^*,N_i,p_i}(x_i) \\
&= \psi_{\pi^*,N_i,p_i}(x_{i+1}) - \psi_{\pi^*,N_i,p_i}(x_i) \\
&\leq \alpha\frac{M}{p_i}\theta_{\pi^*,N_i,p_i}(x_i).
\end{aligned} \tag{4.5h}$$

By Lemma 4.3(ii), $\sum_{i=0}^{\infty} 1/p_i = +\infty$. Hence, by (4.5a) and (4.5f)–(4.5h), $\tilde{\psi}_{\pi^*,N_i,p_i}(x_i) \to -\infty$, as $i \to \infty$. Then we also must have $\psi_{\pi^*,N_i,p_i}(x_i) \to -\infty$, as $i \to \infty$. Let x^* be an accumulation point of $\{x_i\}_{i=0}^{\infty}$. Then there exists an infinite subset $K^* \subset \mathbb{N}$ such that $x_i \to^{K^*} x^*$, and by (3.9) and (3.10a) $|\psi_{\pi^*,N_i,p_i}(x_i) - \psi_{\pi^*}(x^*)| \leq |\psi_{\pi^*,N_i,p_i}(x_i) - \psi_{\pi^*,N_i}(x_i)| + |\psi_{\pi^*,N_i}(x_i) - \psi_{\pi^*}(x_i)| + |\psi_{\pi^*}(x_i) - \psi_{\pi^*}(x^*)| \to^{K^*} 0$, as $i \to \infty$, which is a contradiction. Thus,

$$\limsup_{i\to\infty} \theta_{\pi^*,N_i,p_i}(x_i) = 0. \tag{4.5i}$$

Hence, by Lemma 3.6 and (4.5i), there have to exist an infinite subset $K \subset \mathbb{N}$ and an $\hat{x} \in \mathbb{R}^n$ such that $x_i \to^K \hat{x}$ and $\theta_{\pi^*}(\hat{x}) = 0$. This completes the proof. ∎

LEMMA 4.6. *Suppose that Assumptions* 2.1 *and* 2.6 *hold. Then the smallest eigenvalue* $\sigma_{\min}(\cdot,\cdot)$ *of the matrix-valued function* $[A(\cdot,\cdot)A(\cdot,\cdot)^T + B(\cdot,\cdot)]$ *(see* (2.13a), (2.13d)*) is continuous, and for every compact set* $S \subset \mathbb{R}^n$,

$$\min_{x\in S} \min_{y\in\hat{Y}(x)} \sigma_{\min}(x,y) > 0. \tag{4.6}$$

Proof. For any $x \in \mathbb{R}^n$ and $y \in \mathbb{R}^m$, let $C(x,y) = A(x,y)A(x,y)^T + B(x,y)$, with the smallest eigenvalue $\sigma_{\min}(x,y)$. Since $\sigma_{\min}(x,y) = \min_{\|v\|=1}\langle v, C(x,y)v\rangle$ and $C(\cdot,\cdot)$ is continuous, it follows from Corollary 5.4.2 in [17] that $\sigma_{\min}(\cdot,\cdot)$ is continuous.

Next, let $S \subset \mathbb{R}^n$ be a compact set. By Theorem 5.4.3 in [17], $\hat{Y}(\cdot)$ (see (2.4a)) is outer semicontinuous and compact-valued. Hence, by Theorem 5.4.1 in [17], $\min_{y\in\hat{Y}(\cdot)} \sigma_{\min}(\cdot,y)$ is lower semicontinuous. Since the infimum of a lower semicontinuous function over a compact set is attained, (4.6) follows from Lemma 2.7(ii) and Assumption 2.6. ∎

THEOREM 4.7. *Suppose that Assumptions* 2.1, 2.6, *and* 3.1 *hold and that* $\{x_i\}_{i=0}^{\infty}$ *is a bounded sequence generated by Algorithm* 4.1. *Then there exist an* $\hat{x} \in \mathbb{R}^n$ *and an infinite subset* $K \subset \mathbb{N}$ *such that* $x_i \to^K \hat{x}$, *as* $i \to \infty$, *and* $\hat{x}$ *is a stationary point for* $\mathbf{P}$.

Proof. Let $\{x_j^*\}$ be the sequence generated by Algorithm 4.1 in Step 10. We will show that $\{x_j^*\}$ must be a finite sequence. For the sake of a contradiction, suppose that $\{x_j^*\}_{j=0}^{\infty}$ is an infinite sequence. Since $\{x_i\}_{i=0}^{\infty}$ is a bounded sequence, $\{x_j^*\}_{j=0}^{\infty}$ is bounded, and hence there must exist an infinite subset $K^* \subset \mathbb{N}$ and $x^{**} \in \mathbb{R}^n$ such that $x_j^* \to^{K^*} x^{**}$, as $i \to \infty$.

By Lemmas 2.7(ii) and 2.8, there exist a compact set $\Omega(x^{**})$ and a $\rho_{x^{**}}$ such that $\eta(\cdot,\cdot)$ is continuous on $\mathbb{B}(x^{**},\rho_{x^{**}}) \times \Omega(x^{**})$, and

$$\hat{Y}(x^{**}) + \mathbb{B}_{\rho_{x^{**}}} \subset \Omega(x^{**}). \tag{4.7a}$$

Hence,

$$\pi^{**} \stackrel{\triangle}{=} \max_{x\in\mathbb{B}(x^{**},\rho_{x^{**}})} \max_{y\in\Omega(x^{**})} \sigma \sum_{k=1}^{r_1} |\eta^k(x,y)| \tag{4.7b}$$

is well-defined, and therefore $t_\pi(x,y) \le 0$ for all $\pi \ge \pi^{**}$, $x \in \mathbb{B}(x^{**},\rho_{x^{**}})$, and $y \in \Omega(x^{**})$. Since $\{x_j^*\}$ is an infinite sequence, $\pi_i \to \infty$, as $i \to \infty$. Hence, there exists $i_0 \in \mathbb{N}$ such that $\pi_i > \pi^{**}$ for all $i \ge i_0$. Hence, $t_{\pi_{i-1}}(x_i,y) \le 0$ for all $i > i_0$, $x_i \in \mathbb{B}(x^{**},\rho_{x^{**}})$, and $y \in \Omega(x^{**})$. By Lemma 4.3(ii), $N_i \to \infty$, as $i \to \infty$. Let $\{y_i\}_{i=0}^{\infty}$ be the sequence generated by Algorithm 4.1 in Step 1. Then by Lemma 3.3(i) and (4.7a) there exist $i_1 \ge i_0$ and $\rho_1 \in (0,\rho_{x^{**}}]$ such that for all $i > i_1$ with $x_i \in \mathbb{B}(x^{**},\rho_1)$,

$$y_i \in \hat{Y}_{\pi_{i-1},N_i}(x_i) \subset \hat{Y}(x^{**}) + \mathbb{B}_{\rho_{x^{**}}} \subset \Omega(x^{**}). \tag{4.7c}$$

Therefore, for all $i > i_1$ and $x_i \in \mathbb{B}(x^{**},\rho_1)$, $t_{\pi_{i-1}}(x_i,y_i) \le 0$.

Next, by Lemma 4.6 there exists $\epsilon > 0$ such that

$$2\epsilon = \min_{x\in\mathbb{B}(x^{**},\rho_1)} \min_{y\in\hat{Y}(x)} \sigma_{\min}(x,y). \tag{4.7d}$$

Moreover, $\sigma_{\min}(\cdot,\cdot)$ is continuous, and hence uniformly continuous on $\mathbb{B}(x^{**},\rho_1) \times \Omega(x^{**})$. Hence, there exists $\rho_2 \in (0,\rho_1]$ such that

$$|\sigma_{\min}(x',y') - \sigma_{\min}(x'',y'')| \leq \epsilon \tag{4.7e}$$

for all $x', x'' \in \mathbb{B}(x^{**},\rho_1)$, and $y', y'' \in \Omega(x^{**})$, with $\|x'-x''\| \leq \rho_2$ and $\|y'-y''\| \leq \rho_2$. By Lemma 3.3(i), there exist $\rho_3 \in (0,\rho_2]$ and $i_2 > i_1$ such that for all $i > i_2$ with $x_i \in \mathbb{B}(x^{**},\rho_3)$,

$$y_i \in \hat{Y}_{\pi_{i-1},N_i}(x_i) \subset \hat{Y}(x^{**}) + \mathbb{B}_{\rho_2} \subset \Omega(x^{**}). \tag{4.7f}$$

Consequently, for all $i > i_2$ with $x_i \in \mathbb{B}(x^{**},\rho_3)$, there exists $y_i' \in \hat{Y}(x^{**})$ such that $\|y_i' - y_i\| \leq \rho_2$. Hence, by (4.7d) and (4.7e), we have that for all $i > i_2$ with $x_i \in \mathbb{B}(x^{**},\rho_3)$,

$$\sigma_{\min}(x_i,y_i) = \sigma_{\min}(x_i,y_i) - \sigma_{\min}(x^{**},y_i') + \sigma_{\min}(x^{**},y_i') \geq -\epsilon + 2\epsilon = \epsilon. \tag{4.7g}$$

Since for all $i > i_1$ and $x_i \in \mathbb{B}(x^{**},\rho_1)$, $t_{\pi_{i-1}}(x_i,y_i) \leq 0$; i.e., the test in Step 3 is satisfied for all $\pi \geq \pi_{i-1}$. Since $x_j^* \to^{K^*} x^{**}$, there must exist an infinite set $K^{**} \subset \mathbb{N}$, with elements diverging to infinity, such that $\sigma_{\min}(x_i,y_i) < \sigma_{i-1}$ for all $i \in K^{**}$; i.e., the test in Step 2 fails an infinite number of times. Hence, $\sigma_i \to 0$, as $i \to \infty$. Therefore, there exists an $i_3 > i_2$ such that for all $i > i_3$, $\sigma_{i-1} < \epsilon$. Now, we have that for all $i > i_3$ with $x_i \in \mathbb{B}(x^{**},\rho_3)$, $\sigma_{\min}(x_i,y_i) \geq \sigma_{i-1}$ and $t_\pi(x_i,y_i) \leq 0$ for all $\pi \geq \pi_{i-1}$. Consequently, no $x_i \in \mathbb{B}(x^{**},\rho_3)$ is converted into x_j^* after i_3, which is a contradiction. Hence, $\{x_j^*\}_{j=0}^{j^*}$ is a finite sequence with $j^* < \infty$.

It follows from Theorem 4.5 there exist an infinite subset $K \subset \mathbb{N}$ and an $\hat{x} \in \mathbb{R}^n$ such that $x_i \to^K \hat{x}$ and $\theta_{\pi^*}(\hat{x}) = 0$, with $\pi^* \triangleq \kappa^{j^*+1}\pi_{-1}$. Furthermore, there exists $i^* \in \mathbb{N}$ such that for all $i \geq i^*$ the test in Step 2 is satisfied and the test in Step 3 is satisfied with $\pi = \pi_{i-1}$; i.e., there exists $\sigma^* > 0$ such that for all $i \geq i^*$, $\pi_i = \pi^*$, $\sigma_{\min}(x_i,y_i) \geq \sigma^*$, and

$$t_{\pi^*}(x_i,y_i) \leq 0. \tag{4.7h}$$

Now, $\{y_i\}_{i=0}^{\infty} \subset Y$, which is compact. Hence, there exist $L \subset K$ and $\hat{y} \in Y$ such that $y_i \to^L \hat{y}$, as $i \to \infty$. By Lemma 4.3(ii), $N_i \to \infty$. Hence, Lemma 3.2 gives that $\hat{y} \in \hat{Y}_{\pi^*}(\hat{x})$. By Lemma 4.6, $\sigma_{\min}(\cdot,\cdot)$ is continuous, and hence by continuity, $\sigma_{\min}(\hat{x},\hat{y}) \geq \sigma^*$. This implies that $A(\hat{x},\hat{y})A(\hat{x},\hat{y})^T + B(\hat{x},\hat{y})$ is positive definite, and hence by Lemma 2.8(i), $\nabla_y f^k(\hat{x},\hat{y}), k \in \mathbf{r}_1^*(\hat{x},\hat{y})$, together with $\nabla g^k(\hat{y}), k \in \mathbf{r}_2^*(\hat{y})$, are linearly independent. By Lemma 2.7(ii), $t_{\pi^*}(\cdot,\cdot)$ is continuous at $(\hat{x},\hat{y})$. It follows from (4.7h) that $t_{\pi^*}(\hat{x},\hat{y}) \leq 0$. Hence, by Lemma 2.7(iii),

$$\psi(\hat{x}) = \psi_{\pi^*}(\hat{x}). \tag{4.7i}$$

It now follows from Theorems 2.5 and 2.10 that $\hat{x}$ is stationary for $\mathbf{P}$. This completes the proof. ∎

5. Numerical example. We illustrate Algorithm 4.1 by a numerical example computed on a 500 MHz PC running Matlab [15]. Let $x = (x^1,x^2,x^3) \in \mathbb{R}^3$, $y \in \mathbb{R}$, and

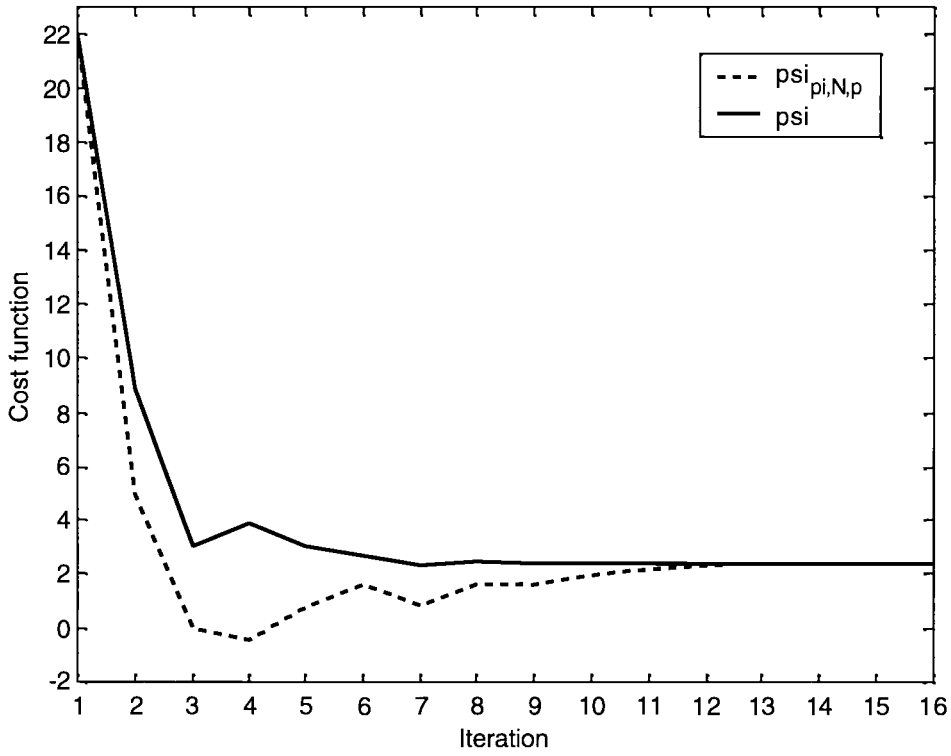


FIG. 1. *Decrease in the cost functions* $\psi_{\pi_i,N_i,p_i}(x_i)$ *and* $\psi(x_i)$.

$$\phi(x,y) = 3(x^1 - y)^2 + (2-y)(x^2)^2 + 5(x^3 + y)^2 + 2x^1 + 3x^2 - x^3 + e^{4y^2}, \tag{5.1a}$$

$$f(x,y) = \frac{1}{4}\sin(x^1 x^2) + y - \frac{1}{2}, \tag{5.1b}$$

$$g^1(y) = -y, \tag{5.1c}$$

$$g^2(y) = y - 1, \tag{5.1d}$$

i.e., $r_1 = 1$, $r_2 = 2$, and $Y = [0,1] \subset \mathbb{R}$.

Based on the reasoning in the paragraphs following Lemma 4.2, we take $\tau_1 = \ln 2$, $\tau_2 = 1$, $\sigma = \kappa = 2$, $\rho = 0.001$, $\mu = 0.5$, $\sigma_{-1} = 10^{-5}$, $\hat{p} = 5 \cdot 10^4$, $\gamma = 10^5$, and $\zeta = 2$. Furthermore, we set the step-size parameters to be $\alpha = 0.5$ and $\beta = 0.8$. The discretization scheme is such that Y_N contains $N+1$ equally spaced numbers in $[0,1]$, i.e., $Y_1 = \{0,1\}, Y_2 = \{0, 0.5, 1\}, Y_3 = \{0, 0.333, 0.667, 1\}$, etc., and $\Delta(N) = 1/N$. The approximation parameters are set to be $p_0 = 1$, $N_0 = 1$, and $\pi_{-1} = 1$, which give a coarse approximation.

Using the starting point $x_0 = (2,1,0)$, we obtain the local minimizer $\hat{x} = (-0.0033, -1.0002, -0.3928)$, with $\psi(\hat{x}) = 2.4100$. Figures 1–3 show how the approximating cost function $\psi_{\pi_i,N_i,p_i}(x_i)$ and the exact cost function $\psi(x_i)$ converge, and how the precision parameters N and p are gradually increased. The penalty π was increased

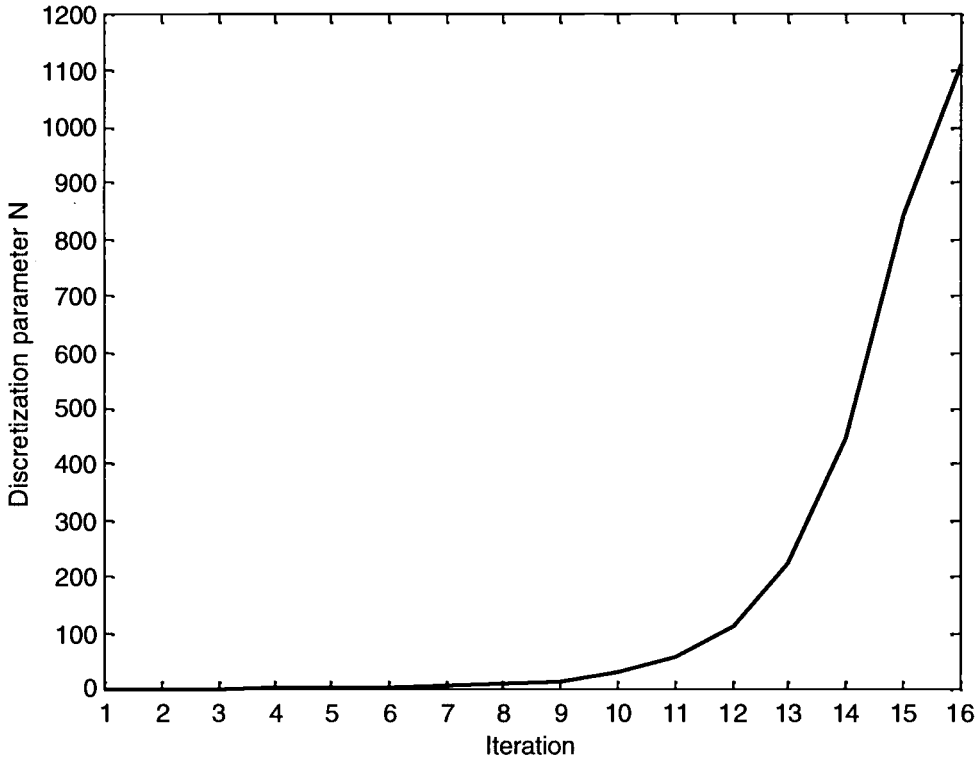


FIG. 2. *Increase in the discretization parameter* N_i.

to 1024 in the first iteration and remained constant after that throughout the rest of the computation. Obviously, the cost function $\psi(x_i)$ cannot be evaluated exactly in finite time; the values in Figure 1 were obtained by a fine discretization of Y. It can be seen from Figures 2 and 3 that the precision parameters N and p stay low until the iterate is close to a local minimizer. The initially coarse approximations reduce ill-conditioning potentially caused by a high smoothing precision parameter (see [18] for an examination of such effects) and computational cost caused by high discretization.

6. Conclusions. We have developed an implementable algorithm for a class of generalized semi-infinite min-max problems based on a sequential solution of gradually better-approximating finite min-max problems. The approximating problems are obtained by exact penalization, discretization, and smoothing. The penalty, discretization, and smoothing parameters are automatically adjusted by using a series of tests. Under mild assumptions, we have shown that if the algorithm generates a bounded sequence, then the penalty parameter remains bounded and there exists an accumulation point which satisfies a first-order optimality condition.

Clearly, discretization is a computationally expensive technique in high-dimensional spaces, and hence the proposed algorithm will be computationally inefficient for problems with a high-dimensional semi-infinite part, i.e., large m. In spite of this, we used a discretization technique because of the need for global maximizers of the inner problem of the min-max-min problem. Obviously, other global optimization techniques could have been used, but we have not evaluated the relative

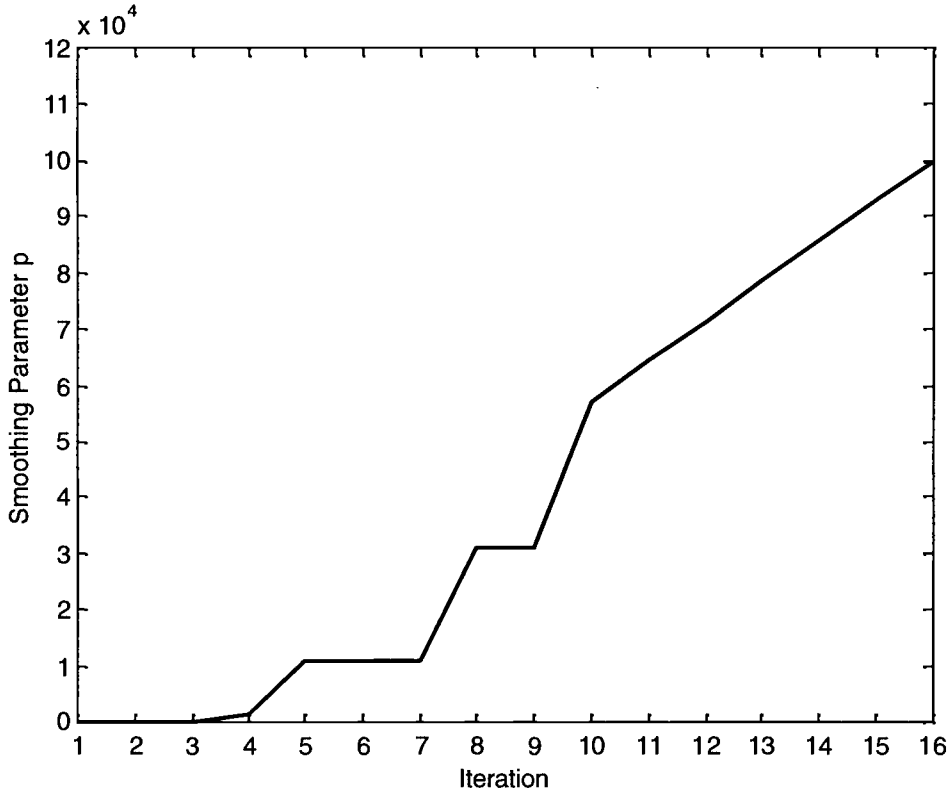


FIG. 3. *Increase in the smoothing parameter* p_i.

merits of alternative techniques.

Appendix A. The optimality condition for **P** derived in Theorem 2.5 (see also Theorem 2.10) can be related to the following optimality condition deduced from Theorem 3.3 in [24].

THEOREM A.1. *Suppose that* $\hat{x}$ *is a local minimizer for* **P**, *that Assumption* 2.1 *holds, and that the vectors* $\nabla_y f^k(\hat{x},y), k \in \mathbf{r_1^*}(\hat{x},y)$, *together with the vectors* $\nabla g^k(y)$, $k \in \mathbf{r_2^*}(y)$, *are linearly independent for all* $y \in \hat{Y}(\hat{x})$. *Then*

$$0 \in \operatorname*{conv}_{y\in\hat{Y}(\hat{x})} \{\nabla_x \phi(\hat{x},y) - f_x(\hat{x},y)^T \alpha(\hat{x},y)\}, \tag{A.1a}$$

where $\alpha(\hat{x},y) \in \mathbb{R}^{r_1}$, *together with* $\beta(\hat{x},y) \in \mathbb{R}^{r_2}$ *(not used here), are the unique Karush–Kuhn–Tucker multipliers for the "inner problem"* (1.2) *at the point* $y \in \hat{Y}(\hat{x})$; *i.e.,* $(\alpha(\hat{x},y), \beta(\hat{x},y))$ *satisfy*

$$\nabla_y \phi(\hat{x},y) - f_y(\hat{x},y)^T \alpha(\hat{x},y) - g_y(y)^T \beta(\hat{x},y) = 0, \tag{A.1b}$$

$$\alpha(\hat{x},y)^T f(\hat{x},y) + \beta(\hat{x},y)^T g(y) = 0, \tag{A.1c}$$

$$\alpha(\hat{x},y) \geq 0,\ \beta(\hat{x},y) \geq 0, \tag{A.1d}$$

$$f(\hat{x},y) \leq 0,\ g(y) \leq 0. \qquad \square \tag{A.1e}$$

THEOREM A.2. *Suppose that Assumption* 2.1 *holds, that* $\hat{x}$ *satisfies* (A.1a)*, and that the vectors* $\nabla_y f^k(\hat{x},y), k \in \mathbf{r}_1^*(\hat{x},y)$*, together with the vectors* $\nabla g^k(y), k \in \mathbf{r}_2^*(y)$*, are linearly independent for all* $y \in \hat{Y}(\hat{x})$*. If* $\pi > 0$ *is such that* $\psi(\hat{x}) = \psi_\pi(\hat{x})$*, and for all* $y \in \hat{Y}(\hat{x})$

$$\pi \geq \sum_{k=1}^{r_1} |\eta^k(\hat{x},y)|, \tag{A.2a}$$

with $\eta(\cdot,\cdot)$ *as in* (2.13g)*, then* $0 \in \bar{G}\psi_\pi(\hat{x})$.

Proof. By Carathéodory's theorem (see, e.g., Theorem 5.2.5 in [17]), (A.1a) holds if and only if there exist $\hat{y}_i \in \hat{Y}(\hat{x})$, $i \in \{1,\ldots,n+1\}$, and a multiplier vector $\hat{\mu} \in \Sigma_{n+1} \stackrel{\triangle}{=} \{\mu \in \mathbb{R}^{n+1} |\ \mu^i \geq 0, i \in \{1,\ldots,n+1\}, \sum_{i=1}^{n+1} \mu^i = 1\}$ such that

$$0 = \sum_{i=1}^{n+1} \hat{\mu}^i \nabla_x \phi(\hat{x},\hat{y}_i) - \sum_{i=1}^{n+1} \sum_{k=1}^{r_1} \hat{\mu}^i \alpha^k(\hat{x},\hat{y}_i) \nabla_x f^k(\hat{x},\hat{y}_i). \tag{A.2b}$$

We will now construct multipliers such that $0 \in \bar{G}\psi_\pi(\hat{x})$. Let $\pi > 0$ satisfy (A.2a) for all $y \in \hat{Y}(\hat{x})$,

$$\zeta_i^k \stackrel{\triangle}{=} \frac{1}{\pi} \alpha^k(\hat{x},\hat{y}_i),\ k \in \mathbf{r_1}, \tag{A.2c}$$

$$\zeta_i^0 \stackrel{\triangle}{=} 1 - \sum_{k=1}^{r_1} \zeta_i^k, \tag{A.2d}$$

$$\mu^i \stackrel{\triangle}{=} \hat{\mu}^i,\ i \in \{1,\ldots,n+1\}, \tag{A.2e}$$

and $y_i \stackrel{\triangle}{=} \hat{y}_i, i \in \{1,\ldots,n+1\}$. Trivially, $\mu \in \Sigma_{n+1}$, $y_i \in Y$, and $f(\hat{x},y_i) \leq 0$ for all $i \in \{1,\ldots,n+1\}$. Furthermore, for all $k \in \mathbf{r_1}$ and $i \in \{1,\ldots,n+1\}$

$$\begin{aligned} \mu^i \zeta_i^k [\phi_\pi^k(\hat{x},y_i) - \omega_\pi(\hat{x},y_i)] &= -\pi \mu^i \zeta_i^k f^k(\hat{x},y_i) \\ &= -\hat{\mu}^i \alpha^k(\hat{x},\hat{y}_i) f^k(\hat{x},\hat{y}_i) = 0, \end{aligned} \tag{A.2f}$$

because from (A.1c)–(A.1e), $\alpha^k(\hat{x},\hat{y}_i) f^k(\hat{x},\hat{y}_i) = 0$ for all $k \in \mathbf{r_1}$ and $i \in \{1,\ldots,n+1\}$. Also, $\phi_\pi^r(\hat{x},y_i) - \omega_\pi(\hat{x},y_i) = 0$ for all $i \in \{1,\ldots,n+1\}$. Next, by (A.2b)

(A.2g)

$$\begin{aligned} \sum_{i=1}^{n+1} \sum_{k=1}^{r} \mu^i \zeta_i^k \nabla_x \phi_\pi^k(\hat{x},y_i) &= \sum_{i=1}^{n+1} \mu^i \nabla_x \phi(\hat{x},y_i) - \sum_{i=1}^{n+1} \sum_{k=1}^{r_1} \mu^i \zeta_i^k \pi \nabla_x f^k(\hat{x},y_i) \\ &= \sum_{i=1}^{n+1} \hat{\mu}^i \nabla_x \phi(\hat{x},\hat{y}_i) - \sum_{i=1}^{n+1} \sum_{k=1}^{r_1} \hat{\mu}^i \alpha^k(\hat{x},\hat{y}_i) \nabla_x f^k(\hat{x},\hat{y}_i) = 0. \end{aligned}$$

It now remains to show that $\zeta_i^0 \geq 0$ for all $i \in \{1,\ldots,n+1\}$. It follows by inspection that the unique multipliers $\alpha(\hat{x},\hat{y}_i)$ and $\beta(\hat{x},\hat{y}_i)$ (see (A.1b)–(A.1e)) solve the minimization problem in (2.14b) with x and y replaced by $\hat{x}$ and $\hat{y}_i$, respectively. Hence, $\alpha(\hat{x},\hat{y}_i)$ and $\beta(\hat{x},\hat{y}_i)$ also satisfy the necessary optimality conditions for (2.14b) given in (2.14a). Since the solution of (2.14a) is unique under the linear independence assumption (see the proof of Lemma 2.7(ii)), we have by definition of $\eta(\cdot,\cdot)$ (see (2.13g))

that $\eta(\hat{x}, y_i) = \alpha(\hat{x}, \hat{y}_i)$ for all $i \in \{1, \ldots, n+1\}$. Hence,

$$\begin{aligned} \zeta_i^0 &= 1 - \sum_{k=1}^{r_1} \zeta_i^k = 1 - \frac{1}{\pi} \sum_{k=1}^{r_1} \alpha^k(\hat{x}, \hat{y}_i) \\ &= 1 - \frac{1}{\pi} \sum_{k=1}^{r_1} \eta^k(\hat{x}, y_i) \geq 1 - \frac{1}{\pi}\pi = 0. \end{aligned} \tag{A.2h}$$

This completes the proof. □

Acknowledgment. The authors are grateful to Professor R. Tyrrell Rockafellar for suggesting the use of exact penalization.

REFERENCES

[1] D.P. Bertsekas, *Constrained Optimization and Lagrange Multiplier Methods*, Academic Press, New York, 1982.

[2] J.F. Bonnans and A. Shapiro, *Perturbation Analysis of Optimization Problems*, Springer-Verlag, New York, 2000.

[3] J.V. Burke, *Calmness and exact penalization*, SIAM J. Control Optim., 29 (1991), pp. 493–497.

[4] F. Clarke, *Optimization and Nonsmooth Analysis*, Wiley, New York, 1983.

[5] O. Ditlevsen and H.O. Madsen, *Structural Reliability Methods*, Wiley, New York, 1996.

[6] T. Glad and E. Polak, *A multiplier method with automatic limitation of penalty growth*, Math. Programming, 17 (1979), pp. 140–155.

[7] T.J. Graettinger and B.H. Krogh, *The acceleration radius: A global performance measure for robotic manipulators*, IEEE J. Robotics and Automation, 4 (1988), pp. 60–69,

[8] R. Hettich and G. Still, *Semi-infinite programming models in robotics*, in Parametric Optimization and Related Topics II, J. Goddat et al., eds., Akademie-Verlag, Berlin, 1991, pp. 112–118.

[9] H. Th. Jongen, J.-J. Ruckmann, and O. Stein, *Generalized semi-infinite optimization: A first order optimality condition and examples*, Math. Programming, 83 (1998), pp. 145–158.

[10] A. Kaplan and R. Tichatschke, *On the numerical treatment of a class of semi-infinite terminal problems*, Optimization, 41 (1997), pp. 1–36.

[11] D. Kincaid and W. Cheney, *Numerical Analysis*, Brooks/Cole, New York, 1996.

[12] E. Levitin, *Reduction of Generalized Semi-infinite Programming Problems to Semi-infinite or Piece-wise Smooth Programming Problems*, Preprint 8-2001, University of Trier, Trier, Germany, 2001.

[13] E. Levitin and R. Tichatschke, *A branch-and-bound approach for solving a class of generalized semi-infinite programming problems*, J. Global Optim., 13 (1998), pp. 299–315.

[14] X. Li, *An entropy-based aggregate method for minimax optimization*, Engineering Optimization, 18 (1997), pp. 277–285.

[15] *Matlab Reference Manual, Version* 5.3*, (R*11*)*, MathWorks, Inc., Natick, MA, 1999.

[16] G. Di Pillo, *Exact penalty methods*, in Algorithms for Continuous Optimization: The State of the Art, E. Spedicato, ed., Kluwer Academic Publishers, Dordrecht, The Netherlands, 1994.

[17] E. Polak, *Optimization. Algorithms and Consistent Approximations*, Springer-Verlag, New York, 1997.

[18] E. Polak and J.O. Royset, *Algorithms with adaptive smoothing for finite min-max problems*, J. Optim. Theory Appl., 119 (2003).

[19] E. Polak and J.O. Royset, *Algorithms for finite and semi-infinite min-max-min problems using adaptive smoothing techniques*, J. Optim. Theory Appl., 119 (2003).

[20] B.N. Pshenichnyi and Yu. M. Danilin, *Chislennye Metody v Ekstremal'nykh Zadachakh*, Nauka, Moscow, 1975.

[21] R.T. Rockafellar and R.J.-B. Wets, *Variational Analysis*, Springer-Verlag, New York, 1997.

[22] J.O. Royset, A. Der Kiureghian, and E. Polak, *Reliability-based optimal structural design by the decoupling approach*, Reliability Engineering and System Safety, 73 (2001), pp. 213–221.

[23] J.-J. Ruckmann and A. Shapiro, *On first-order optimality conditions in generalized semi-infinite programming*, J. Optim. Theory Appl., 101 (1999), pp. 677–691.

[24] O. Stein, *First order optimality conditions for degenerate index sets in generalized semi-infinite programming*, Math. Oper. Res., 26 (2001), pp. 565–582.
[25] O. Stein and G. Still, *On optimality conditions for generalized semi-infinite programming problems*, J. Optim. Theory Appl., 104 (2000), pp. 443–458.
[26] G. Still, *Generalized semi-infinite programming: Theory and methods*, European J. Oper. Res., 119 (1999), pp. 301–313.
[27] G. Still, *Generalized semi-infinite programming: Numerical aspects*, Optimization, 49 (2001), pp. 223–242.
[28] G.-W. Weber, *Generalized semi-infinite optimization: On some foundations*, Vychisl. Tekhnol., 4 (1999), pp. 41–61.

SIAM J. OPTIM.
Vol. 14, No. 1, pp. 35–52

AN INFEASIBLE ACTIVE SET METHOD FOR QUADRATIC PROBLEMS WITH SIMPLE BOUNDS*

K. KUNISCH† AND F. RENDL‡

Abstract. A primal-dual active set method for quadratic problems with bound constraints is presented. Based on a guess on the active set, a primal-dual pair (x, s) is computed that satisfies the first order optimality condition and the complementarity condition. If (x, s) is not feasible, a new active set is determined, and the process is iterated. Sufficient conditions for the iterations to stop in a finite number of steps with an optimal solution are provided. Computational experience indicates that this approach often requires only a few (less than 10) iterations to find the optimal solution.

Key words. primal-dual active set method, convex programming

AMS subject classifications. 90C06, 90C20, 90C99

PII. S1052623400376135

1. Introduction.

We consider the convex programming problem

$$(P) \qquad \min J(x) \text{ subject to } x - b \leq 0, \tag{1.1}$$

where

$$J(x) := \frac{1}{2} x^T Q x + d^T x,$$

Q is a positive definite $n \times n$ matrix, and $b, d \in \mathbb{R}^n$. This problem has received considerable interest in the literature. We recall some of the more recent contributions.

Solution methods based on active sets and gradient projection are among the most popular approaches to solve (P) and can be traced back to the 1960s. More recent contributions from Moré and Toraldo [14, 15] indicate that this approach is applicable also for large-scale problems. The key steps here consist of using the conjugate gradient method to investigate a given face of the feasible region and the gradient projection method to move to a different face. Kočvara and Zowe [12] replace the gradient projection by successive overrelaxation with projection, thereby gaining efficiency as compared to [15].

Another solution strategy consists of treating the inequalities by the interior-point idea: a sequence of parameterized barrier functions is (approximately) minimized using Newton's method. The main computational effort consists of solving the Newton system to get the search direction. From the vast literature on this topic, we refer to the book by Wright [16]. More recently, Heinkenschloss, Ulbrich, and Ulbrich [10] developed an affine-scaling interior-point approach to general nonlinear bound-constrained problems, which does not assume strict complementarity to hold at local solutions. D'Appuzo et al. [7] present a parallel implementation of an interior-point method for box-constrained quadratic programming.

*Received by the editors August 3, 2000; accepted for publication (in revised form) July 9, 2002; published electronically May 15, 2003.

http://www.siam.org/journals/siopt/14-1/37613.html

†Institut für Mathematik, Universität Graz, A-8010 Graz, Austria (karl.kunisch@uni-graz.at). This author was supported in part by the Fonds zur Förderung der wissenschaftlichen Forschung (FWF), Austria, under SFB 03 "Optimierung und Kontrolle."

‡Institut für Mathematik, Universität Klagenfurt, A-9020 Klagenfurt, Austria (franz.rendl@uni-klu.ac.at).

Finally, trust–region-type methods have also been investigated to deal with bound-constrained problems. We refer to Coleman and Lin [4], Coleman and Liu [5], and Lin and Moré [13] for further details.

Our contribution to solve (P) consists of an infeasible active-set approach that was already successfully applied to constrained optimal control problems; see [1, 2]. The approach from [1, 2] was tailored to deal with specially structured elliptic partial differential equations. Here we investigate this approach for general convex quadratic problems of the form (P). Our approach is iterative. In each step we maintain the first order optimality condition and the complementarity constraint associated with (P); see (2.1) and (2.2) below. The iterations are carried out until primal and dual feasibility hold; see (2.3) and (2.4) below. A primary feature of the algorithm is its simplicity. Moreover, it does not rely on tuning parameters except in the degenerate case, where a parameter controlling the stopping criterion is used. In the nondegenerate case the algorithm terminates after finitely many steps. In fact, computational experience shows that typically only few (often only between 5 and 10) iterations are required to reach the optimal solution starting from an arbitrarily chosen initial active set. We succeed in proving finite step convergence of the algorithm from arbitrary initial data independently of assumptions on strict complementarity, provided appropriate sufficient conditions related to strict diagonal dominance of Q are satisfied.

The paper is organized as follows. At the end of this section we summarize notation used throughout this paper. In section 2 we describe the details of our algorithm. The main theoretical contributions are contained in section 3, which presents sufficient conditions for the method to converge in a finite number of iterations. The practical performance is described in section 4. We consider randomly generated problems, problems arising in mathematical physics, and problems from [14, 15] for the sake of comparison.

Notation. The following notation will be used throughout. For a subset $A \subseteq N := \{1, \dots, n\}$ and $x \in \mathbb{R}^n$ we write x_A for the components of x indexed by A, i.e., $x_A := (x_i)_{i \in A}$. The complement of A will be denoted by $\bar{A}$. If Q is a matrix and A and B are subsets of N, then $Q_{A,B}$ is the submatrix of Q, given by $Q_{A,B} = (q_{ij})_{i \in A, j \in B}$. If $A = B$ we write Q_A for $Q_{A,A}$. The vector of ones will be denoted by e. For $a, b \in \mathbb{R}^n$ we write $a \circ b$ to denote the vector of elementwise products, $a \circ b := (a_i b_i)_{i \in N} \in \mathbb{R}^n$.

2. The algorithm. To describe the algorithm that will be investigated analytically and computationally let $b, d \in \mathbb{R}^n$ and $Q = Q^T$ be given, with Q a positive definite $n \times n$ matrix. We consider the *convex quadratic minimization problem* with simple bound constraints (1.1).

Even though we could assume without loss of generality that $b = 0$, we prefer to maintain a general upper bound on x. The Karush–Kuhn–Tucker (KKT) system for (P) is given by

$$Qx + d + s = 0, \tag{2.1}$$

$$s \circ (x - b) = 0, \tag{2.2}$$

$$(KKT)$$

$$x - b \leq 0, \tag{2.3}$$

$$s \geq 0. \tag{2.4}$$

It is well known that a vector x together with a vector $s \in \mathbb{R}^n$ of Lagrange multipliers for the inequality constraints furnishes a (global) minimum of (P) if and only if (x, s) satisfies the KKT system. A solution pair (x, s) of KKT is called strictly complementary if there exists no index i such that $s_i = 0$ and $x_i = b_i$.

TABLE 2.1
Description of the algorithm.

Prototype Algorithm
Input: Q symmetric, positive definite $n \times n$ matrix, $b, d \in \mathbb{R}^n$. $A \subseteq N$, e.g., $A = N$.
Output: (x, s) optimal solution
repeat until (x,s) is optimal
Solve KKT(A), i.e., set $x_A = b_A,\ s_I = 0$ and compute x_I from (2.7) and s_A from (2.8);
$A := \{i : x_i > b_i \text{ or } s_i > 0\}$;

We now describe in some detail the approach sketched in the introduction. The crucial step in solving (P) is to identify those inequalities which are active, i.e., the set $A \subseteq N$, where the solution to (P) satisfies $x_A = b_A$. Then, with $I := N \backslash A$, we must have $s_I = 0$.

To compute the remaining elements x_I and s_A of x and s, we use (2.1) and partition the equations and variables according to A and I:

$$\begin{pmatrix} Q_A & Q_{A,I} \\ Q_{I,A} & Q_I \end{pmatrix} \begin{pmatrix} x_A \\ x_I \end{pmatrix} + \begin{pmatrix} d_A \\ d_I \end{pmatrix} + \begin{pmatrix} s_A \\ s_I \end{pmatrix} = 0. \tag{2.5}$$

The second set of equations can be solved for x_I, because Q_I is by assumption positive definite:

$$x_I = -Q_I^{-1}(d_I + Q_{I,A}\ b_A).$$

Substituting this into the first equation implies $s_A = -d_A - Q_{A,N}\ x$. If our guess for A would have been correct, then $x_I \leq b_I$ and $s_A \geq 0$ would have to hold. Suppose this is not the case. Then we need to make a new "guess" for A, which we denote by A^+. Let us first look at s_A. If $s_i > 0$, this confirms our previous guess $i \in A$, so we include i also in A^+. Consider now x_I. If $x_i > b_i$ we set $x_i = b_i$ in the next iteration. Hence we include i in A^+ also in this case. Formally we arrive at

$$A^+ := \{i : x_i > b_i \text{ or } s_i > 0\}. \tag{2.6}$$

This completes an intuitive description of one iteration of the algorithm. It is summarized in Table 2.1. To simplify notation we introduce for given A the following set $KKT(A)$ of equations:

$$KKT(A) \qquad Qx + d + s = 0, \quad x_A = b_A, \quad s_I = 0.$$

The solution of $KKT(A)$ is given by $x_A = b_A,\ s_I = 0$, and

$$x_I = -Q_I^{-1}(d_I + Q_{I,A}b_A), \tag{2.7}$$

$$s_A = -d_A - Q_A\ b_A - Q_{A,I}x_I. \tag{2.8}$$

We observe that if (x, s) satisfies $KKT(A)$, then (2.1) and (2.2) of KKT hold. The iterates of the algorithm are well-defined, because $KKT(A)$ has a unique solution for every $A \subseteq N$, due to $Q \succ 0$.

In Lemma 2.1 it is guaranteed that the set A changes in each iteration unless the algorithm stops. Hence the algorithm does not get trapped by generating the same (x, s) in two consecutive iterations.

LEMMA 2.1. *Let A and A^+ be the active sets in two consecutive iterations of the algorithm. Then either the current primal and dual variables satisfy* (2.1)–(2.4) *or* $A \neq A^+$.

Proof. Let A and A^+ be the active sets in two consecutive iterations. Suppose that $A = A^+$, and let $i \in A^+$. Then either $s_i > 0$ or $x_i > b_i$. But $x_i = b_i$ for $i \in A$, and $A = A^+$, so we cannot have $x_i > b_i$, and hence $x \leq b$. Therefore $i \in A^+$ implies that $s_i > 0$, i.e., $s_{A^+} > 0$. But we also have $s_I = s_{I^+} = 0$ because $A = A^+$, so $s \geq 0$. Therefore the solution of $KKT(A)$ is optimal, and the algorithm would have stopped before generating A^+. ∎

Remark. To guess the set A at the start, several obvious strategies could be employed. Using $A = N$, we get $x = b$ and $s = -(Qb + d)$. Setting $A = \emptyset$ gives $s = 0$ and $x = -Q^{-1}d$. In the latter case a linear system of order n has to be solved, which may be expensive for large-scale sparse problems. Alternatively, A may be selected at random.

3. Convergence analysis.

3.1. Index partition. To investigate the convergence behavior of the algorithm, we look at two consecutive iterations. Suppose that some iteration, say $k \geq 1$, is carried out with the set $A^k \subseteq N$, yielding $(x^{(k)}, s^{(k)})$ as the solution of $KKT(A^{(k)})$. According to (2.6), the new active set is

$$A^{(k+1)} = \{i : x_i^{(k)} > b_i \text{ or } s_i^{(k)} > 0\}.$$

Let $(x^{(k+1)}, s^{(k+1)})$ denote the solution of $KKT(A^{(k+1)})$. To avoid too many superscripts, we write

$$(A, x, s) \text{ for } (A^{(k)}, x^{(k)}, s^{(k)}) \text{ and } (B, y, t) \text{ for } (A^{(k+1)}, x^{(k+1)}, s^{(k+1)}).$$

Given A, we find that x, s, B, y, t are determined by

$$x_A = b_A,\ s_{\bar{A}} = 0,\ Qx + d + s = 0, \tag{3.1}$$

$$B = \{i : x_i > b_i \text{ or } s_i > 0\}, \tag{3.2}$$

$$y_B = b_B,\ t_{\bar{B}} = 0,\ Qy + d + t = 0. \tag{3.3}$$

The following partition of N into mutually disjoint subsets will be useful in our convergence analysis. We first partition A into

$$S := \{i \in A : s_i \leq 0\}$$

and $A \backslash S$. The set I is partitioned into

$$T := \{i \in I : x_i > b_i\}$$

and $I \backslash T$. In Table 3.1 we summarize the relevant information about x, s, y, t for this partition. The column for x indicates that $x_T > b_T, x_S = b_S$ and so on. A nonspecified entry, for instance y_S, indicates that y_S is in no specific relation to b_S. Finally, let $K_1 := \{i \in S : y_i > b_i\}$, let $K_2 := \{i \in I \backslash T : y_i > b_i\}$, and let $K := K_1 \cup K_2$. The set K contains the indices of primal infeasibility of y.

TABLE 3.1
Partition of index set N.

	s	t	x	y
T	$=0$		$> b$	$= b$
S	≤ 0	$=0$	$= b$	
$I \backslash T$	$=0$	$=0$	$\leq b$	
$A \backslash S$	> 0		$= b$	$= b$

3.2. Merit function and finite step convergence. We recall the well-known [3] augmented Lagrangian merit function for (P) given by

$$\mathcal{L}_c(x,s) = J(x) + s^T \max\left(x-b, -\frac{s}{c}\right) + \frac{c}{2}\left\|\max\left(x-b,-\frac{s}{c}\right)\right\|^2,$$

where $c > 0$ and the max-operation is acting componentwise. Here we shall employ a variation of $\mathcal{L}_c(x,s)$ given by

$$L_c(x,s) = \mathcal{L}_c(x, \max(0,s)).$$

Let us note that

$$L_c(x,s) = J(x) + \frac{c}{2}\|\max(x-b,0)\|^2,$$

provided that (x,s) satisfies the complementarity condition $s \circ (x-b) = 0$, which is the case for the iterates of our algorithm. In the remainder of this section we shall establish sufficient conditions for the decay of L_c along the iterates of the algorithm; i.e.,

$$L_c(y,t) - L_c(x,s) < 0$$

holds for any two consecutive pairs (x,s) and (y,t). In particular this implies convergence of the algorithm in finitely many steps. We start with some preliminary results. First, we investigate how the objective function changes during consecutive iterations. One cannot expect a monotone decrease of $J(x)$, as the iterates may be infeasible.

LEMMA 3.1. *Let (x,s), (y,t) and T be given as above. Then we have*

$$J(y) - J(x) = \frac{1}{2}(y-x)^T \begin{pmatrix} Q_T & 0 \\ 0 & -Q_{\bar{T}} \end{pmatrix} (y-x).$$

Proof. We use the Q-inner product, $\langle a,b\rangle_Q := a^T Q b$, with associated norm $\|a\|_Q^2 := \langle a,a\rangle_Q$ and get

$$J(y) - J(x) = \frac{1}{2}\|y\|_Q^2 - \frac{1}{2}\|x\|_Q^2 + z^T d,$$

where $z = y - x$. Using the identity

$$\|y\|_Q^2 - \|x\|_Q^2 = 2\langle y-x, y\rangle_Q - \|y-x\|_Q^2,$$

the right-hand side can now be rewritten to obtain

$$J(y) - J(x) = -\frac{1}{2}z^T Q z + z^T(Qy+d) = -\frac{1}{2}z^T Q z - z^T t.$$

The last equation follows from $Qy + d = -t$. Now $t_i = 0$ for $i \in S \cup (I \backslash T)$ and $z_i = 0$ on $A \backslash S$. Therefore $z^T t = \sum_{i \in T} z_i t_i$. Furthermore $t - s = -Qz$ and $t_i - s_i = t_i$ for $i \in T$, and hence

$$-\sum_{i \in T} z_i t_i = \sum_{i \in T} z_i (Qz)_i = z^T \begin{pmatrix} Q_{T,N} \\ 0 \end{pmatrix} z.$$

Summarizing, we see that

$$J(y) - J(x) = -\frac{1}{2} z^T \begin{pmatrix} Q_T & Q_{T,\bar{T}} \\ Q_{\bar{T},T} & Q_{\bar{T}} \end{pmatrix} z + z^T \begin{pmatrix} Q_T & \frac{1}{2} Q_{T,\bar{T}} \\ \frac{1}{2} Q_{\bar{T},T} & 0 \end{pmatrix} z. \qquad \square$$

LEMMA 3.2. *Let (x, s), (y, t) as well as T and K be given as above. Then we have*

$$\| \max(y - b, 0) \|^2 - \| \max(x - b, 0) \|^2 = \sum_{i \in K} |y_i - b_i|^2 - \sum_{i \in T} |x_i - b_i|^2.$$

Proof. The claim follows from the fact that x is infeasible precisely on T; see Table 3.1. Moreover, by the definition of the sets K_1 and K_2, the variable y is infeasible on K. $\square$

In summary we have proved the following result.

PROPOSITION 3.3. *For every two consecutive pairs (x, s) and (y, t) we have*

$$L_c(y, t) - L_c(x, s) = \frac{1}{2} z^T \begin{pmatrix} Q_T & 0 \\ 0 & -Q_{\bar{T}} \end{pmatrix} z + \frac{c}{2} \sum_{i \in K} |y_i - b_i|^2 - \frac{c}{2} \sum_{i \in T} |x_i - b_i|^2.$$

We can now state the following sufficient conditions (C1) and (C2) for decrease of the merit function. Some more notation is required. Let $\mu := \lambda_{\min}(Q) > 0$ denote the smallest eigenvalue of Q. Further, let

$$\nu := \max\{\|Q_{A,\bar{A}}\| : A \subset N, A \neq \emptyset, A \neq N\},$$

where $\| \cdot \|$ denotes the spectral norm. We also use the diagonal matrix $D := \operatorname{diag}(q_{11}, \ldots, q_{nn})$, consisting of the main diagonal elements of Q, and define

$$q := \min\{q_{ii} : i \in N\}.$$

Finally, let $r := \|Q - D\|$ denote the norm of Q with the elements from the main diagonal removed.

$$\text{condition (C1)} \qquad \operatorname{cond}(Q) < \left(\frac{\mu}{\nu}\right)^2 - 1, \tag{3.4}$$

$$\text{condition (C2)} \qquad \operatorname{cond}(Q) < \left(\frac{q}{r}\right)^2 - 1, \tag{3.5}$$

where $\operatorname{cond}(Q) = \frac{\lambda_{\max}(Q)}{\lambda_{\min}(Q)}$. We will show below that either (C1) and (C2) ensures strict decrease of the merit function L_c.

Conditions (3.4) and (3.5) both require some diagonal dominance of Q. In fact, turning to (3.4), we observe that ν is independent of the main diagonal of Q and can be bounded as

$$\nu \leq \|Q - D\| = r.$$

Consequently, (3.4) is implied by

$$\mathrm{cond}(Q) < \left(\frac{\mu}{r}\right)^2 - 1.$$

Clearly, this condition will hold if $r > 0$ is sufficiently small, relative to μ and $\mathrm{cond}(Q)$.

Turning to (3.5), assume that $\frac{r}{q} < 1$. This implies that

$$\|QD^{-1} - I\| \leq \|Q - D\|\|D^{-1}\| = \frac{r}{q} < 1,$$

which allows us to interpret (3.5) as a condition on the diagonal dominance of Q relative to $\mathrm{cond}(Q)$.

To see the effect of these two conditions in more detail, we consider the matrix

$$Q(\alpha) := \alpha I - (E - I) = \begin{pmatrix} \alpha & -1 & \dots & -1 \\ -1 & \ddots & \ddots & \vdots \\ \vdots & \ddots & & -1 \\ -1 & \dots & -1 & \alpha \end{pmatrix}.$$

Here $E = ee^T$ is the matrix of all ones. The eigenvalues of $Q(\alpha)$ are $a + 1 - n$ and $\alpha + 1$. We assume that $\alpha > n - 1$ to ensure $Q \succ 0$. In this case

$$\mathrm{cond}(Q) = \frac{\alpha+1}{\alpha+1-n},\ r = \|Q-D\| = \|E-I\| = n-1,\ q = \alpha,\ \text{and}\ \nu = \frac{n}{2}\ \text{for even}\ n.$$

Therefore (3.4) is equivalent to

$$\frac{\alpha+1}{\alpha+1-n} < 4\left(\frac{\alpha+1-n}{n}\right)^2 - 1.$$

This condition holds for $\alpha + 1 \geq 1.9n$. Similarly, it can be established that condition (3.5) is satisfied for $\alpha+1 \geq 1.8n$. In summary, $\alpha+1 > n$ ensures diagonal dominance of $Q(\alpha)$, but our conditions require a stronger form of diagonal dominance.

THEOREM 3.4. *Let $(x, s), (y, t)$ be two consecutive primal-dual iterates of the algorithm, and set $c = \|Q\| + \mu$. If* (C1) *holds, then we have*

$$2(L_c(y, t) - L_c(x, s)) \leq c_1\|y - x\|^2 < 0,$$

with $c_1 := (\|Q\| + \mu)(\frac{\nu}{\mu})^2 - \mu < 0$. Similarly, (C2) *implies that*

$$2(L_c(y, t) - L_c(x, s)) \leq c_2\|y - x\|^2 < 0,$$

with $c_2 := (\|Q\| + \mu)(\frac{r}{q})^2 - \mu < 0$. In both cases the algorithm stops after a finite number of iterations for every b and d in $\mathbb{R}^n$.

Proof. As before, we set $z = y - x$. We first note that for $i \in K$ we have $x_i \leq b_i$, hence $0 < y_i - b_i \leq y_i - x_i$, and therefore

$$\sum_{i\in K}(y_i - b_i)^2 \leq \|z_K\|^2.$$

Similarly, $y_T = b_T$, and hence

$$\sum_{i\in T}(x_i - b_i)^2 = \|z_T\|^2.$$

Using Proposition 3.3 we get

$$2(L_c(y,t) - L_c(x,s)) \leq \|Q\|\|z_T\|^2 - \mu\|z_{\bar{T}}\|^2 + c\|z_K\|^2 - c\|z_T\|^2.$$

The next goal is to bound $\|z_K\|$ in terms of $\|z\|$. On K_1 we have $s_{K_1} \leq 0, t_{K_1} = 0$, and therefore $(Qz)_{K_1} = s_{K_1} \leq 0$. Similarly, $s_{K_2} = t_{k_2} = 0$ on K_2, and thus $(Qz)_{K_2} = 0$. It follows that $(Qz)_K = Q_K z_K + Q_{K,\bar{K}} z_{\bar{K}} = \binom{s_{K_1}}{0}$. Taking the inner product with z_K and noting that $z_{K_1} \geq 0$ we find

$$z_K^T (Qz)_K = z_K^T Q_K z_K + z_K^T Q_{K,\bar{K}} z_{\bar{K}} = s_{K_1}^T z_{K_1} \leq 0. \tag{3.6}$$

We can now bound $\|z_K\|$ in the following two ways.

Variant 1 *to bound* $\|z_K\|$. We have

$$\mu\|z_K\|^2 \leq \|z_K^T Q_K z_K\| \leq \|z_K^T Q_{K,\bar{K}} z_{\bar{K}}\| \leq \nu\|z_K\|\|z_{\bar{K}}\|.$$

The first inequality follows from the definition of μ, the second uses (3.6), and the last follows from the definition of ν. So we get

$$\|z_K\|^2 \leq \frac{\nu^2}{\mu^2}\|z_{\bar{K}}\|^2 \leq \frac{\nu^2}{\mu^2}(\|z_T\|^2 + \|z_{\bar{T}}\|^2).$$

Variant 2 *to bound* $\|z_K\|$. Using again (3.6) and the definition of D we can alternatively write

$$0 \geq z_K^T (Qz)_K = z_K^T D_K z_K + z_K^T (Q_{K,N} - D_{K,N}) z.$$

From this we get $q\|z_K\|^2 \leq z_K^T D_K z_K \leq r\|z_K\|\|z\|$, and therefore

$$\|z_K\|^2 \leq \frac{r^2}{q^2}\|z\|^2 = \frac{r^2}{q^2}(\|z_T\|^2 + \|z_{\bar{T}}\|^2).$$

Variant 1 yields

$$2(L_c(y,t) - L_c(x,s)) \leq \left(\|Q\| - c + c\frac{\nu^2}{\mu^2}\right)\|z_T\|^2 + \left(c\frac{\nu^2}{\mu^2} - \mu\right)\|z_{\bar{T}}\|^2, \tag{3.7}$$

while variant 2 gives

$$2(L_c(y,t) - L_c(x,s)) \leq \left(\|Q\| - c + c\frac{r^2}{q^2}\right)\|z_T\|^2 + \left(c\frac{r^2}{q^2} - \mu\right)\|z_{\bar{T}}\|^2. \tag{3.8}$$

Note that setting $c := \|Q\| + \mu$ makes the coefficients of $\|z_T\|^2$ and $\|z_{\bar{T}}\|^2$ in (3.7) as well as (3.8) equal to one another.

Up to this point we have not yet used either of the conditions (C1) or (C2). Suppose now that (C1) holds. Then the coefficients in (3.7) are both equal to $(\|Q\| + \mu)\frac{\nu^2}{\mu^2} - \mu = c_1$. Moreover, condition (C1) implies that $\|Q\| + \mu < \mu(\frac{\mu}{\nu})^2$, and hence $c_1 < 0$.

Similarly, we find that in (3.8) the coefficients are both equal to c_2. If we assume that (C2) holds, then $c_2 < 0$.

It is clear that under the above conditions it is impossible that some active set A is reproduced twice. Since there is only a finite number, namely 2^n, of different sets, the algorithm stops with an optimal solution for every choice of b and d. ∎

We conclude this section with several remarks.

1. The main computational effort per iteration is the solution of the linear equation (2.7). The size of this system varies, but typically it is much smaller than n, since it needs only be solved for the inactive variables.
2. We emphasize that the iterates of the algorithm do not observe primal and dual feasibility. Thereby big changes to the active set are possible and occur in numerical practice. This is an extremely useful feature especially for large scale problems.
3. It is straightforward to extend the algorithm to both lower and upper bounds, i.e., to constraints of the form

$$l \le x \le u.$$

   We experimented also with this type of problem and did not find a significant difference in performance to the unilaterally constrained problem.
4. It is also straightforward to extend the present approach to strictly convex cost functions f. In this case the algorithm can be included in an outer SQP-type iteration.

3.3. Comparison to closely related methods. To compare the present algorithm with Bertsekas's projected Newton method [3, Chapter 1.5], we first recall a simple version of this algorithm.

(i) Given the current iterate x with $x \le b$, compute $A^B = \{i : x_i = b_i \text{ and } (\nabla J(x))_i < 0\}$ and set $I^B = N \backslash A^B$.
(ii) Compute a step δx by $(\delta x)_{I^B} = -Q_{I^B}^{-1}(Qx + d)_{I^B}$, $(\delta x)_{A^B} = -(Qx + d)_{A^B}$.
(iii) Compute a step size α based on an Armijo rule along the projected arc and set the new iterate $x_+ = P(x + \alpha \delta x)$.

Here P denotes the projection onto the feasible set, and the superscript B refers to Bertsekas. Furthermore $\nabla J(x) = Qx + d$ corresponds to $-s$ in the development of our algorithm. We note that $(x_+^B)_{A^B} = b$, due to steps (ii) and (iii) of the Bertsekas algorithm.

We point out some significant differences between Bertsekas's method and ours. Bertsekas's algorithm maintains primal feasibility throughout. To ensure convergence, a line search is performed in each iteration. In contrast our algorithm always satisfies the first order equation (2.1) and the complementarity condition (2.2). It neither utilizes a line search nor maintains primal feasibility (2.3). As a consequence the gradients determining the active sets are computed at different points for the two methods. If the primal variable of our algorithm is feasible, then the active sets of both methods coincide. This does not imply, however, that the following iterates are identical. The equality criterion $x_i = b_i$ in step (ii) of Bertsekas's algorithm causes difficulties in the convergence proof and can lead to jamming in numerical realizations. It is therefore suggested [3] to replace it by the criterion $0 \le b_i - x_i < \varepsilon_k$, where $\varepsilon_k \to 0^+$ as the iteration count $k \to \infty$. If Q is positive definite and strict complementarity holds at the solution (x, s) to KKT, then the Bertsekas algorithm converges in finitely many steps.

In our algorithm the new iterate is uniquely determined by the current active set A, and consequently it can generate only a finite number of different iterates. A convergence proof therefore amounts to verifying that primal and dual feasibilities (2.3) and (2.4) are reached.

The projected Newton method was refined in several papers; see, e.g., [14, 15]. The resulting implementations differ significantly from our algorithm not only in the

points already addressed above but also in the fact that they utilize tuning parameters. Our algorithm does not depend on sophisticated tuning parameters.

Next we note that (KKT) can be considered as a complementarity system. The numerical treatment of such a system has been studied extensively. We mention splitting schemes, damped Newton methods, and interior-point methods, all of which are described in [6, Chapter 5.8], for example. We now describe how the method of this paper relates to the damped Newton method as analyzed in [6]. For this purpose we define, for given $x \in \mathbb{R}^N, s \in \mathbb{R}^N$, the disjoint decomposition of N into subsets $\mathcal{A}, \mathcal{D}, \mathcal{I}$ as

$$\mathcal{A} = \{i : b_i - x_i < s_i\}, \quad \mathcal{D} = \{i : b_i - x_i = s_i\}, \quad \mathcal{I} = \{i : b_i - x_i > s_i\} \tag{3.9}$$

and introduce the mapping $H : \mathbb{R}^N \to \mathbb{R}^N$ by

$$H(x) = \min\{-Qx - d, b - x\}.$$

Note that the system

$$H(x) = 0, \ s = -(Qx + d),$$

characterizes a solution pair (x, s) to (2.1)–(2.4).

Let us consider a Newton step for solving $H(x) = 0$, and let x denote the current iterate associated with $s = -(Qx+d)$. We assume that x is nondegenerate, i.e., $\mathcal{D} = \emptyset$. Then H is differentiable at x, and a Newton step satisfying $H(x) + \nabla H(x)\delta x = 0$ has the form

$$\begin{pmatrix} -(Qx + d)_{\mathcal{I}} \\ (b - x)_{\mathcal{A}} \end{pmatrix} - \begin{pmatrix} Q_{\mathcal{I}} & Q_{\mathcal{I}\mathcal{A}} \\ 0 & I_{\mathcal{A}} \end{pmatrix} \begin{pmatrix} (\delta x)_{\mathcal{I}} \\ (\delta x)_{\mathcal{A}} \end{pmatrix} = 0, \tag{3.10}$$

where $I_{\mathcal{A}}$ is the identity matrix of appropriate dimension. Solving for δx results in

$$x^+_{\mathcal{A}} = x_{\mathcal{A}} + (\delta x)_{\mathcal{A}} = b_{\mathcal{A}} \tag{3.11}$$

and

$$x^+_{\mathcal{I}} = x_{\mathcal{I}} + (\delta x)_{\mathcal{I}} = -Q^{-1}_{\mathcal{I}}(d_{\mathcal{I}} + Q_{\mathcal{I}\mathcal{A}} b_{\mathcal{A}}). \tag{3.12}$$

Defining s^+ by means of $s^+ = -(Qx^+ + d)$ we have

$$s^+_{\mathcal{I}} = 0 \text{ and } x^+_{\mathcal{A}} = -d_{\mathcal{A}} - Q_{\mathcal{A}} b_{\mathcal{A}} - Q_{\mathcal{A}\mathcal{I}} x_{\mathcal{I}}. \tag{3.13}$$

From (3.11)–(3.13) we conclude that (x^+, s^+) from the above Newton step and (x, s) of $KKT(A)$ in our algorithm coincide if $A = \mathcal{A}$ and $I = \mathcal{I}$. If, as in our algorithm, one accepts the update (x^+, s^+) obtained from the full Newton step, then, since $s^+ \circ (x^+ - b) = 0$, the new active sets according to (3.9) and those determined from our prototype algorithm coincide.

The damped Newton algorithm, however, does not accept the full Newton step but rather utilizes a line search guaranteeing that the actual new iterate is nondegenerate and that $H^T H$ acts as a merit function. For the resulting algorithm it is shown in [6] that if $\tilde{x}$ is a nondegenerate accumulation point of the damped Newton method, then $(\tilde{x}, \tilde{s})$ with $\tilde{s} = -(Q\tilde{x}+d)$ is a solution to the complementarity problem. Nondegeneracy corresponds to strict complementarity of the solution to (KKT). The latter is not required for Theorem 3.4 to hold.

3.4. Equality constraints. We close this section with a discussion of why our approach may have difficulties if we allow equality constraints. For this purpose we consider the following problem:

$$\text{minimize } J(x) \text{ such that } Rx - r = 0, \quad x \le b.$$

Denoting by u the multipliers for the equality constraints, the KKT system for this problem is given by

$$Qx + d + s + R^T u = 0, \quad Rx = r, \quad s \circ (x - b) = 0, \quad x \le b, \quad s \ge 0.$$

Solving the system $Qx + d + s + R^T u = 0,\ Rx = r$ under the additional constraint that $x_A = b_A,\ s_I = 0$ leads to

$$\begin{pmatrix} Q_I & R_I^T \\ R_I & 0 \end{pmatrix} \begin{pmatrix} x_I \\ u \end{pmatrix} = \begin{pmatrix} -d_I - Q_{I,A} x_A \\ r - R_A x_A \end{pmatrix}.$$

This system need not be solvable even if R has full rank. R_I could, for instance, be the zero matrix for some $I \subseteq N$. On the other hand, this system is solvable for any I if there is only a single equation $\sum r_i x_i = r_0$ and $r_i \neq 0$ for all i. This occurs, for instance, if x represents a convex combination, $x \ge 0,\ \sum_i x_i = 1$. Numerical experiments for this case are given in section 4.5.

4. Computational experience. In section 3 we gave sufficient conditions for convergence of our algorithm. In this section we look at the practical behavior of the algorithm by considering a variety of test problems. We implemented the algorithm in MATLAB and provide the source code along with the routines for the examples of section 4.2 for download at

http://www-sci.uni-klu.ac.at/math-or/home/publications/active-ml.tar.

4.1. A randomly generated problem. To get a first impression, we study in some detail a randomly generated problem, where we vary Q by making it increasingly ill-conditioned. In order for the reader to be able to reproduce some of the following results, we provide the MATLAB commands that we used to generate the data Q, d and b.

```
>> n=500;
>> rand('seed',n);
>> p=sprandn(n,n,.1)+speye(n);
>> p=tril(triu(p,-100));
>> Q0=p*p';
>> d=rand(n,1)*20*n-10*n;
>> b=ones(n,1);
>> Astart=find(rand(n,1)>rand);
```

The problem of size $n = 500$ was generated so that $Q0$ is a sparse positive semidefinite matrix, which is singular. The matrices Q are obtained from $Q0$ by adding a small multiple of the identity matrix I:

$$Q = Q0 + \varepsilon I,$$

where $\varepsilon \in \{1, 10^{-1}, 10^{-4}, 10^{-7}, 10^{-10}\}$. The estimated condition numbers of these matrices are given in Table 4.1. We used the MATLAB command `condest` to compute them. The only nontrivial input to our algorithm is the initial guess A for the optimal active set. The algorithm is started by a randomly generated initial active set Astart; see the last command line above.

TABLE 4.1
Iteration counts for a random problem of size $n = 500$. $Q = Q0 + \varepsilon I$, where $Q0 \succeq 0$ is singular.

5	6	7	8	9	10	11	12	cond(Q)	ε
540	460	0	0	0	0	0	0	795	1
0	67	575	343	15	0	0	0	9.3e+03	1e-1
0	0	11	306	470	176	35	2	8.7e+06	1e-4
0	0	14	266	443	220	47	10	7.4e+09	1e-7
0	0	13	244	437	235	59	12	5.1e+12	1e-10

In Table 4.1 we summarize the iteration counts of our algorithm. Each line represents 1000 runs on the same problem with different starting sets A. The columns labeled with numbers *it* ranging from 5 to 12 indicate how often the algorithm stopped after *it* iterations. We note that the algorithm never took more than 12 iterations in all of the 5000 runs. For the well-conditioned problem with $\varepsilon = 1$, the algorithm in fact always stopped after no more than 6 iterations.

4.2. Biharmonic equation. A rich source of applications for our algorithm comes from applications in mathematical physics. We first consider a model problem describing small vertical deformations u of a horizontal, elastic thin plate occupying a domain $\Omega \subset \mathbb{R}^2$, which is clamped along its boundary Γ, under the influence of a vertical force of density $f \in L^2(\Omega)$, with the plate constrained to remain below an obstacle $\psi \in L^\infty(\Omega)$:

$$\left.\begin{array}{l} \min \ \frac{1}{2}\displaystyle\int_\Omega |\Delta u(x)|^2 dx - \int_\Omega f(x)u(x)dx \\ \text{over} \ \ u \in H^2(\Omega) \ \ \text{satisfying} \\ u = 0, \ \frac{\partial u}{\partial n} = 0 \ \ \text{on} \ \ \Gamma \\ u(x) \le \psi(x) \ \text{a.e. in} \ \Omega. \end{array}\right\} \tag{4.1}$$

We assume throughout that ψ is uniformly positive in a neighborhood of Γ. This implies, in particular, that the set of feasible solutions to (4.1) is nonempty. It is then well known that (4.1) admits a unique solution u^* in $H_0^2(\Omega) = \{v \in H^2(\Omega) : v|\Gamma = 0, \frac{\partial v}{\partial n}|\Gamma = 0\}$; see, e.g., [8, section 4.4]. The Lagrange multiplier λ associated with the inequality constraint is only a measure in $L^\infty(\Omega)^*$, the dual of $L^\infty(\Omega)$. The KKT system characterizing the solution to (4.1) is given by

$$\left.\begin{array}{l} (\Delta u^*, \Delta v)_{L^2} + \langle \lambda, v\rangle_{L^{\infty,*},L^\infty} = (f, v)_{L^2} \ \ \text{for all} \ \ v \in H_0^2(\Omega), \\ \langle \lambda, v\rangle_{L^{\infty,*},L^\infty} \ge 0 \ \ \text{for all} \ \ v \in H_0^2(\Omega), \ \text{with} \ v \ge 0 \\ \langle \lambda, u^* - \psi\rangle_{L^{\infty,*},L^\infty} = 0, \ u^* - \psi \le 0 \ \ \text{a.e. in} \ \Omega, \end{array}\right\} \tag{4.2}$$

where $(\cdot,\cdot)$ denotes the inner product in $L^2(\Omega)$, $\langle\cdot,\cdot\rangle_{L^{\infty,*},L^\infty}$ denotes the duality pairing between $L^\infty(\Omega)^*$ and $L^\infty(\Omega)$, and we recall that $H_0^2(\Omega) \subset L^\infty(\Omega)$.
Formally, (4.2) can be expressed as

$$\left.\begin{array}{l} \Delta^2 u^* + \lambda = f \ \ \text{in} \ \Omega, \ \ \text{with} \ u|\Gamma = \frac{\partial u}{\partial n}|\Gamma = 0, \\ \lambda \ge 0, u^* \le \psi, \lambda(x)(u^* - \psi)(x) = 0 \ \ \text{in} \ \Omega. \end{array}\right\} \tag{4.3}$$

To realize (4.1), discretization of the biharmonic Δ^2 with homogeneous Dirichlet and Neumann boundary conditions, as well as of u, λ, f, and ψ, is required. For our numerical tests we chose Ω as the unit square which was discretized by a uniform axiparallel grid of meshsize $h = \frac{1}{m+1}$, for fixed $m \in \mathbb{N}$, resulting in m^2 interior nodes. The biharmonic operator, including the boundary conditions, was discretized on the

basis of a 13-point finite difference stencil as described, e.g., in [9, page 105], resulting after proper ordering of the nodes in a positive definite $m^2 \times m^2$-matrix Q_h. It is worthwhile to point out that Q_h is not an M-matrix. The functions u, f, and ψ were approximated by their interior nodal values and are denoted by u_h, f_h, ψ_h. Approximating the integral by cuboids centered at the nodes of the grid, the discretization of (4.1) turns out to be

$$\left.\begin{array}{l} \min \ \frac{1}{2}\, u_h^T\, Q_h\, u_h + f_h^T u_h \\ \text{subject to} \ \ u_h \le \psi_h, \end{array}\right\} \tag{4.4}$$

which is of the form (P). For the discretized problem a Lagrange multiplier $\lambda_h \ge 0$ in $\mathbb{R}^{m^2}$ clearly exists. We solved (4.4) for different choices of f and ψ. Numerical results are presented for a typical case with

$$f(x,y) = -60(1-x^2)\ y\ e^{-7(x-.9)^2-4(y-.1)^2} + 100x(1-y)\ e^{-3(x-.2)^2-6(y-.8)^2}$$

(see Figure 4.1) and

$$\psi(x,y) = 4 \cdot 10^{-5}.$$

A straightforward application of our algorithm with $m = 128$ and an unstructured initialization chosen as $u_h = \psi_h$ requires 71 iterations and 172 seconds to reach the exact solution of (4.4). All computation times given are seconds on a DEC ALPHA workstation.

To reduce the number of iterations and to utilize the advantage of fast solution on coarse meshes a multilevel approach turned out to be very efficient. For this purpose we choose a sequence of grids characterized by meshsizes $h_i = 2^{-i}h$, $i = 0, 1, \ldots,$ where h is an initial coarse meshsize. The nested algorithm then consists of a coarse to fine sweep. The optimal active set for meshsize h_i is interpolated to the grid h_{i+1} and utilized as the initial active set for the problem with gridsize h_{i+1}. Utilizing this strategy for our test problem resulted in a significant reduction in the total number of iterations. In Table 4.2 we give the number of required iterations on each grid level, starting with $m = 8$. In our computational experiments we noted that the algorithm typically maintained dual feasibility throughout ($s \ge 0$), and only primal feasibility was violated. This phenomenon is known to hold if Q is an M-matrix; see [11].

We next compare our algorithm to the Moré–Toraldo refinement of Bertsekas's projected Newton algorithm for a problem that is widely studied in the literature. For that purpose we consider the obstacle problem for the harmonic equation, which arises from (4.1) by replacing the Laplacian with the Nabla operator and removing the homogeneous Neumann boundary conditions. All further specifications regarding the discretization of the operators, the choices for Ω, f, and ψ, are as in [15, section 7.1]. We solve this problem with our algorithm including grid refinement as described above. In Table 4.3 we provide the iteration counts and the number of inactive variables at the optimum. Comparing with Table 7.1 in [15] we observe an identical number of inactive variables for $m = 100$. Moreover, the iteration count is consistently less than half of that reported in [15] for a wide range of m.

Our numerical results for obstacle problems of numerical physics suggest that the prototype algorithm combined with grid refinement is highly competitive with the most efficient existing methods.

4.3. Lack of strict complementarity. Our convergence analysis of section 3 holds independent of any assumptions concerning strict complementarity; however, in

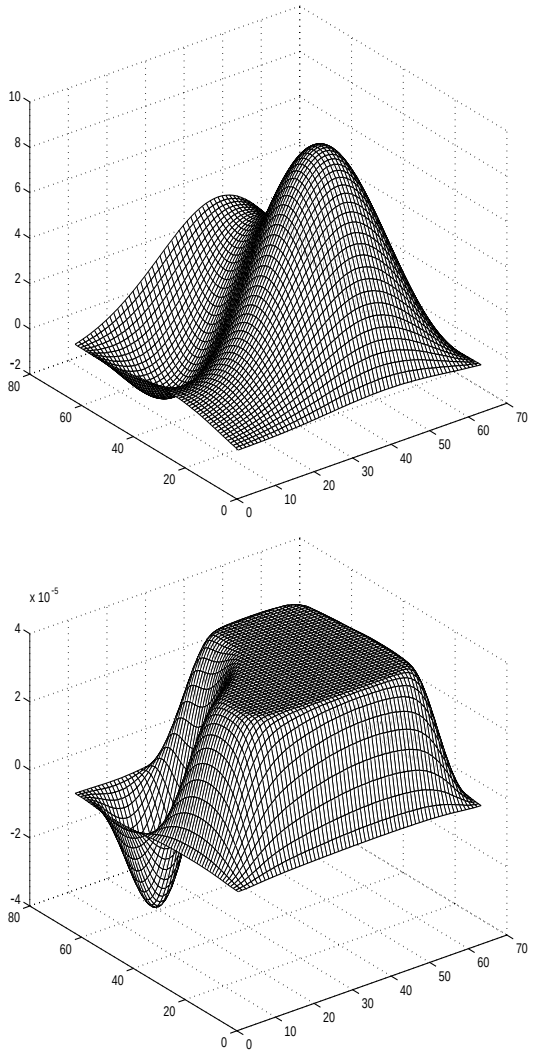


FIG. 4.1. *Force f acting on plate (top); deformation u of plate (bottom).*

TABLE 4.2
Number of iterations for the biharmonic obstacle problem.

m	8	16	32	64	128	256
iter	6	7	10	6	7	8
time (sec.)	0.01	0.08	0.6	2.7	26.7	364
largest system	26	139	647	2738	11251	45580

TABLE 4.3
Number of iterations for the obstacle problem.

m	25	50	100	200	400
iter	7	4	4	4	4
time (sec.)	0.05	0.27	1.5	12.2	127.3
final system	199	899	3843	15880	64636

TABLE 4.4
Iteration count for random problems from [14].

		3	4	5	6	7	8	9	10	11	12	13	14-16	fail
		$tol = 0$												
3	3	269	4812	4643	276									
3	12	4	157	1775	4080	2711	1062	189	22					
12	3			5	502	3449	4449	1395	199	1				
12	12				78	742	2219	2713	2291	930	423	209	86	309
		$tol = 10^{-12}$												
3	3	269	4812	4643	276									
3	12	4	157	1775	4080	2711	1062	189	22					
12	3			30	979	4386	3680	856	69					
12	12			62	568	2329	3339	2433	953	279	36	1		

finite precision computations the test $s_i > 0$ is sensitive to round-off errors. Therefore we replace it by $s_i > -tol$, where $tol > 0$ is a small tolerance. Similarly, we allow small violations of primal feasibility and set our stopping condition to

$$s_i \geq -tol, \quad x_i \leq b_i + tol \quad \text{for all } i.$$

With this modification, we tested our method on problems lacking strict complementarity, which we generated as follows.

First we partition N into three nonempty sets A, B, I. We select b and $Q \succ 0$ at random. Setting $x_A = b_A$, $x_B = b_B$, $x_I < b_I$ and $s_B = 0$, $s_I = 0$, $s_A > 0$ and defining $d := -Qx - s$ yields a solution pair (x, s) satisfying KKT, with $s_B = x_B - b_B = 0$. By construction the solution lacks strict complementarity on B.

Our computational tests for this type of problem with $tol = 10^{-6}$ did not indicate a significant performance difference compared to problems where strict complementarity holds.

The following class of examples, which are constructed in such a way that they are arbitrarily close to lacking strict complementarity, allows comparison with results presented in the literature. We follow the approach of [14] and generate random instances where both the condition number of Q and the degree of degeneracy can be set by the user. To allow comparisons with [14], we choose the dimension n of the problems to be $n = 100$. The parameter *ncond* sets the condition number of Q to e^{ncond}. We selected $ncond \in \{3, 12\}$. The number *ndeg* controls the degree of degeneracy of the problem. The example from [14] is constructed in such a way that for all i in the active set at the solution we set $s_i = 10^{-\mu_i \, ndeg}$, where μ_i is a uniformly distributed random number in $[0, 1]$. A large value of *ndeg* indicates that the optimal dual variable to the active constraint may be close to zero, depending on μ_i. Similar to [14] we choose $ndeg \in \{3, 12\}$. Finally, we selected the cardinality of active variables at the optimal solution to be $n/2$, with upper and lower bound constraints.

Table 4.4 contains computational results for random instances generated for the parameter combinations for *ndeg* and *ncond* which are specified in the first two columns. For each pair of parameter settings, 100 random instances are generated and solved with 100 different initial active sets. This gives 10000 runs for each line. The number of iterations needed for these 10000 runs is documented. We first ran all instances with $tol = 0$; i.e., we ignored potential difficulties arising from lack of strict complementarity.

For this choice of *tol* the algorithm solved all instances with no more than 10 iterations per run, except for the most difficult setting with $ndeg = 12, ncond = 12$.

For the latter, 309 out of 10000 instances failed due to cycling. Setting $tol = 10^{-12}$ for the stopping condition, as described above, led to convergence of all instances. The second part in the table summarizes the corresponding iteration counts.

The comparison with the results in [14] can be carried out on the basis of iteration counts since for both algorithms the most expensive step is the solution of the linear system on the inactive set. For both choices of *tol* our algorithm requires significantly fewer iterations, with a big majority of runs terminating with less than 10 iterations. Moreover, our algorithm is less sensitive to *ndeg* and *ncond*.

4.4. Regularization. The conditions (3.4) or (3.5) are sufficient for our method to converge. In practice we noticed, however, that the method also converges when these conditions are not satisfied, provided that Q is positive definite and not too ill-conditioned.

To make the method work independently of the conditioning of Q, we propose to use the following regularization term in the first few iterations of the algorithm. In iteration k of the algorithm we use

$$Q' := Q + \frac{t}{2^{k-1}} I \quad \text{for } k \le k_0$$

and set $Q' = Q$ after iteration k_0. Here the number t depends on the scaling of Q. In our tests we set $t = 1$ and $k_0 = 4$. With this modification we never ran into computational difficulties and managed to handle even problems with singular Q. A specific class of examples is given in the next subsection.

4.5. Projection onto polytope. Suppose we are given $n+1$ points $a_0, \ldots, a_n$ in $\mathbb{R}^m$. If P denotes the convex hull of $a_1, \ldots, a_n$, we consider the problem of finding the point $a \in P$ closest to a_0. Let $A = (a_1, \ldots, a_n)$. Then $a \in P$ if and only if $a = Ax,\ x \ge 0,\ \sum_i x_i = 1$. Thus we arrive at the following problem:

$$\text{minimize } \|Ax - a_0\|^2 \text{ such that } x \ge 0,\ \sum_i x_i = 1.$$

We generate our problems as follows. A is chosen to be a sparse random matrix of order $m \times n$, where $n > m$, with nonzero entries uniformly distributed in $(0, 1)$. We set $a_0 = 0$ and measure the distance of the polytope $P \subseteq \mathbb{R}^n_+$, contained in the positive orthant, from the origin. Under our assumptions, $Q = A^T A$ is positive semidefinite with dimension of the nullspace at least $n - m$, so the matrix is highly singular. To ensure that the iterates are well-defined, we have added a multiple of $(\frac{1}{2})^{k-1}$ times the identity for iterations $k = 1, \ldots, 4$. After iteration 4 we continued with the singular matrix Q, without encountering any computational difficulties. In Table 4.5 we present again accumulated results for 100 test runs with $n = 500$ and $n = 1000$. In both cases $m = n/2$. The meaning of the numbers in Table 4.5 is the same as that in Table 4.1. The number of iterations was never more than 10, despite the fact that we work with a singular matrix after iteration 4. Obviously, after the

TABLE 4.5
Distance of polytope from origin (iteration counts).

n	m	6	7	8	9	10	11
500	250	1	23	61	12	3	0
1000	500	0	6	56	33	5	0

TABLE 4.6
Cardinality of inactive set during iterations for $m = 128$. The third column provides the number of common elements to the subsequent inactive set.

Iteration	$n - \|A\|$	Common	Difference
1	11097	10928	169
2	11192	11092	100
3	11225	11191	34
4	11245	11225	20
5	11249	11245	4
6	11251	11249	2
7	11251	11251	0

first four iterations the approximation to the true active set was already close enough to the optimal solution so that no computational difficulties occurred.

4.6. Cholesky update/downdate. The algorithm which we use for the computational tests can be improved by utilizing the fact that as the method progresses, the current estimate of the active set gets closer to the active set at the optimum. Hence from a certain iteration on this set changes only in a few variables from one iteration to the next when compared to the total number of variables. To give some intuition, we provide in Table 4.6 the cardinalities of the inactive set (= size of the linear system to be solved) and the number of common elements to the subsequent inactive set for the biharmonic equation with $m = 128$, starting from the extrapolated optimal active set from $m = 64$; see Figure 4.1.

In the current version of the algorithm we do not use this feature and compute the Cholesky factor of the system from scratch in each iteration. It would certainly be worth testing to use the Cholesky factor from the previous iteration and to perform the required update and downdate steps to add or remove elements as necessary. Since version 5.3 of MATLAB does not support this feature for sparse matrices, we have no computational results yet for this variant of the algorithm.

5. Discussion. In this paper we presented an infeasible active set method. Under certain conditions we were able to prove global convergence of the method. Moreover, we demonstrated the efficiency of the method by considering a diverse set of test problems. As can be seen from these examples the algorithm converges in many cases for which the sufficient conditions of Theorem 3.4 are not satisfied. It could be the focus of further research efforts to narrow the gap between the classes of problems for which the algorithm converges efficiently in numerical practice and those for which convergence can be proved. Special features of the algorithm include its ability to find the exact numerical solution of the problem and the fact that at each iteration level the size of the linear system which must be solved is determined by the currently inactive set, which can be significantly smaller than the total set of variables. As a consequence the proposed algorithm differs significantly from interior-point methods, for example. From the numerical experiments we observe that the algorithm has the feature of "correcting" many active variables to inactive ones and vice versa during the early stages of the iterations. This is certainly one of its distinguishing practical features. The total number of iterations is frequently quite insensitive with respect to initialization. Lack of strict complementarity as well as singularity of the system matrix do not inhibit the efficiency of the algorithm. In the context of the obstacle problem for the plate equation we demonstrated that it can be useful to combine our algorithm with a multilevel approach.

REFERENCES

[1] M. Bergounioux, M. Haddou, M. Hintermüller, and K. Kunisch, *A comparison of a Moreau–Yosida-based active set strategy and interior point methods for constrained optimal control problems*, SIAM J. Optim., 11 (2000), pp. 495–521.

[2] M. Bergounioux, K. Ito, and K. Kunisch, *Primal-dual strategy for constrained optimal control problems*, SIAM J. Control Optim., 37 (1999), pp. 1176–1194.

[3] D.P. Bertsekas, *Constrained Optimization and Lagrange Multipliers*, Academic Press, New York, 1982.

[4] T.F. Coleman and Y. Lin, *An interior trust region approach for nonlinear minimization subject to bounds*, SIAM J. Optim., 6 (1996), pp. 418–445.

[5] T.F. Coleman and J. Liu, *An interior Newton method for quadratic programming*, Math. Program., 85 (1999), pp. 491–523.

[6] R.W. Cottle, J.-S. Pang, and R.E. Stone, *The Linear Complementarity Problem*, Academic Press, Boston, 1992.

[7] M. D'Apuzzo, M. Marino, P.M. Pardalos, and G. Toraldo, *A Parallel Implementation of a Potential Reduction Algorithm for Box-Constrained Quadratic Programming*, Tech. report, Center for Research on Parallel Computing and Supercomputers, Napoli, Italy, 2000.

[8] R. Glowinski, J.L. Lions, and R. Trémolières, *Numerical Analysis of Variational Inequalities*, North–Holland, Amsterdam, 1981.

[9] W. Hackbusch, *Elliptic Partial Differential Equations*, Springer, Berlin, 1992.

[10] H. Heinkenschloss, M. Ulbrich, and S. Ulbrich, *Superlinear and quadratic convergence of affine-scaling interior-point Newton methods for problems with simple bounds without strict complementarity assumption*, Math. Program., 86 (1999), pp. 615–635.

[11] R.H.W. Hoppe, *Multigrid algorithms for variational inequalities*, SIAM J. Numer. Anal., 24 (1987), pp. 1046–1065.

[12] M. Kočvara and J. Zowe, *An iterative two-step algorithm for linear complementarity problems*, Numer. Math., 68 (1994), pp. 95–106.

[13] C.-J. Lin and J.J. Moré, *Newton's method for large bound-constrained optimization problems*, SIAM J. Optim., 9 (1999), pp. 1100–1127.

[14] J.J. Moré and G. Toraldo, *Algorithms for bound constrained quadratic problems*, Numer. Math., 55 (1989), pp. 377–400.

[15] J.J. Moré and G. Toraldo, *On the solution of large quadratic programming problems with bound constraints*, SIAM J. Optim., 1 (1991), pp. 93–113.

[16] S.J. Wright, *Primal-Dual Interior Point Methods*, SIAM, Philadelphia, 1997.

SIAM J. OPTIM.
Vol. 14, No. 1, pp. 53–76

ON NUMERICAL SOLUTION OF THE MAXIMUM VOLUME ELLIPSOID PROBLEM*

YIN ZHANG† AND LIYAN GAO†

Abstract. In this paper we study practical solution methods for finding the maximum volume ellipsoid inscribing a given full-dimensional polytope in $\Re^n$ defined by a finite set of linear inequalities. Our goal is to design a general-purpose algorithmic framework that is reliable and efficient in practice. To evaluate the merit of a practical algorithm, we consider two key factors: the computational cost per iteration and the typical number of iterations required for convergence. In addition, numerical stability is an important factor. We investigate some new formulations upon which we build primal-dual type interior-point algorithms, and we provide theoretical justifications for the proposed formulations and algorithmic framework. Extensive numerical experiments have shown that one of the new algorithms is the method of choice among those tested.

Key words. maximum volume ellipsoid problem, primal-dual interior-point algorithms

AMS subject classifications. 90C25, 90C30

PII. S1052623401397230

1. Introduction.

The ellipsoidal approximation of polytopes is an important problem in its own right, while it is also a basic subroutine in a number of algorithms for different problems. One example is that Lenstra's algorithm for the integer programming feasibility problem [12, 13] uses the ellipsoidal approximation of polytopes as a subroutine.

Consider a full-dimensional polytope $\mathcal{P} \in \Re^n$ defined by m linear inequalities. For brevity, we will call the problem of finding the maximum volume ellipsoid inscribing $\mathcal{P}$ the MaxVE problem. The MaxVE problem has its root in the rounding of convex bodies in $\Re^n$. One of the earliest studies was done by John [7]. In particular, John's results imply that once the maximum volume inscribing ellipsoid $\mathcal{E}$ is found in $\mathcal{P}$, then $\mathcal{E} \subset \mathcal{P} \subset n\mathcal{E}$, where $n\mathcal{E}$ is the ellipsoid resulting from dilating $\mathcal{E}$ by a factor n about its center. Such a pair of ellipsoids is also called a *Löwner–John pair* for $\mathcal{P}$. That is, $\mathcal{E}$ provides an n-rounding for $\mathcal{P}$. Moreover, if $\mathcal{P}$ is centrally symmetric around the origin, then the rounding factor can be reduced to $\sqrt{n}$.

Ellipsoids have good geometric and computational properties that make them much easier to handle, both theoretically and computationally, than polytopes. For example, the global minimum of any quadratic in an ellipsoid can be located in polynomial time (see [25], for example), while finding such a global minimum in a polytope is generally an NP-hard problem. For many problems a fruitful and effective approach is to use ellipsoids to approximate polytopes in various theoretic and algorithmic settings. A celebrated example is Khachiyan's ellipsoid method [9]—the first polynomial-time algorithm for linear programming. Other applications include optimal design [20, 22], computational geometry (for example, [24]), and algorithm construction (for example, [4] and [21]).

*Received by the editors October 31, 2001; accepted for publication (in revised form) September 10, 2002; published electronically May 15, 2003. This research was supported in part by NSF grants DMS-9973339 and DMS-9872009, DOE grant DE-FG03-97ER25331, and DOE/LANL contract 03891-99-23.

http://www.siam.org/journals/siopt/14-1/39723.html

†Department of Computational and Applied Mathematics, Rice University, Houston, TX 77005 (yzhang@rice.edu, gaoly@rice.edu).

Recently, several randomized polynomial-time algorithms ([2, 8, 14], for example) have been proposed for approximating the volume of convex bodies. (Computing the volume itself is NP-hard.) In the case of a polytope, these algorithms require approximating the polytope by an ellipsoid.

It is known that the rounding of a polytope can be accomplished by the (shallow-cut) ellipsoid method in polynomial time (see, for example, [19, 4]). It is also known, however, that the ellipsoid method is not a practically efficient algorithm. A number of interior-point algorithms have been proposed in recent years for the MaxVE problems, for example, by Nesterov and Nemirovskii [17], Khachiyan and Todd [11] (also see [10] for a related problem), Nemirovskii [16], and Anstreicher [1].

Nesterov and Nemirovskii [17] constructed a three-stage barrier method for finding an ϵ-optimal ellipsoid $\mathcal{E}$ such that its volume $\mathrm{Vol}(\mathcal{E}) \geq \mathrm{Vol}(\mathcal{E}^*)e^{-\epsilon}$, where $\mathcal{E}^*$ is the maximum volume ellipsoid inscribing $\mathcal{P}$ and $\epsilon \in (0, 1)$. They obtained a complexity bound $O(m^{2.5}(n^2 + m)\ln(\frac{mR}{\epsilon}))$ for their algorithm, where m is the number of constraints and R is a priori known ratio of the radii of two concentric balls, the larger ball containing the given polytope $\mathcal{P}$ and the smaller one being contained in $\mathcal{P}$. The term n^2 comes from the requirement of solving linear systems involving an $n \times n$ matrix-valued variable.

Khachiyan and Todd [11] proposed an algorithm that attains the complexity estimate of $O(m^{3.5}\ln(\frac{mR}{\epsilon})\ln(\frac{n\ln R}{\epsilon}))$. The algorithm applies the basic barrier method to a small number of subproblems and requires only solving linear systems of $n + m$ equations to compute the involved Newton directions. In their formulation the matrix-valued variable is explicitly treated as dependent on another vector-valued variable during the solution of Newton linear systems.

Nemirovskii [16] showed that the MaxVE problem can be reformulated as a saddle-point problem in $m + n$ variables and solved by a path-following method for approximating saddle points of a sequence of self-concordant convex-concave functions as defined in [16]. Nemirovskii proved that the complexity of the algorithm is $O(m^{3.5}\ln(\frac{mR}{\epsilon}))$.

Most recently, Anstreicher [1] proposed an algorithm that uses key ideas of Khachiyan and Todd [11] but avoids solving the subproblems required in the Khachiyan and Todd algorithm. This way, Anstreicher's algorithm attains the complexity estimate of $O(m^{3.5}\ln(\frac{mR}{\epsilon}))$, which is the same as in [16]. Anstreicher also showed that first computing an approximate analytic center of the polytope can reduce the complexity to $O((mn^2 + m^{1.5}n)\ln(R) + m^{3.5}\ln(\frac{m}{\epsilon}))$.

In addition, Vandenberghe, Boyd, and Wu [23] proposed an algorithm for the class of MAXDET problems to which the MaxVE problem belongs. However, their algorithm does not take into account the special structure of the MaxVE problem.

All the aforementioned works are primarily concerned with the complexity issues, and the proposed algorithms are theoretical in nature. The objective of the present study is to identify or construct a numerically efficient and stable algorithm for solving general MaxVE problems. Our study is not aimed at solving very large-scale problems, so we will not consider aspects of exploiting sparsity and other special structures that may be present in the polytope-defining inequalities.

Since for many convex programs primal-dual interior-point algorithms have proven to be superior in practice than either primal or dual algorithms, we will mainly investigate primal-dual type algorithms, though we will also consider particular primal algorithms for the purpose of comparison.

Two features are common in all the known interior-point algorithms for solving the MaxVE problem. First, they are iterative in nature. Second, they require solving a linear system at each iteration to update the current iterate. Hence, in judging the practical efficiency of an algorithm, we must consider two key factors: (i) how many iterations the algorithm typically requires in practice for obtaining an approximate solution of a certain quality and (ii) how expensive it is to solve the relevant linear system at each iteration. Besides efficiency, another important consideration is the robustness of the algorithm. The robustness of an iterative algorithm is often determined by the numerical stability of the solution procedure for linear systems that has to be invoked at every iteration.

In most primal-dual algorithms for linear programming or semidefinite programming, at each iteration one solves a large linear system by reducing it to a smaller Schur complement system obtained by block elimination. Moreover, the coefficient matrix in the Schur complement system is often positive definite. This procedure has proven to be efficient and at the same time adequately stable. Likewise, in this paper we will try to identify primal-dual algorithms for which the corresponding linear systems can be reduced by block Gauss elimination to a well-behaved Schur complement system.

The paper is organized as follows. In section 2 we describe the formulation of the MaxVE problem. We introduce some primal-dual type interior-point algorithms in section 3 and give related theoretical results in section 4. We summarize the Khachiyan and Todd algorithm and our modification in section 5. Numerical comparative results on these four algorithms are presented in section 6. Finally, we offer some concluding remarks in section 7.

We now introduce some notation. For any given vector $v \in \Re^p$, we denote the $p \times p$ diagonal matrix with v on its diagonal either by $\mathrm{Diag}(v)$ or by its upper-case letter V whenever no confusion can occur. On the other hand, for a square matrix M, $\mathrm{diag}(M)$ is the vector formed by the diagonal of M. The Hadamard product is represented by the small circle "$\circ$." Unless otherwise specified, superscripts for vectors and subscripts for scalars that are not elements of a matrix are iteration counts. For a vector v, inequalities of the form $v > a$ are interpreted as componentwise, where a is a vector of the same size. For symmetric matrices, $A \succ B$, or equivalently $A - B \succ 0$, means that $A - B$ is positive definite. We use $\Re^m_+$ and $\Re^m_{++}$ to represent the nonnegative and positive orthants in $\Re^m$, respectively. The notation $\mathcal{S}^n_{++}$ represents the cone of all symmetric positive definite matrices in $\Re^{n\times n}$. For a set $\mathcal{W}$ in $\Re^m$, we denote its closure by $\mathrm{cl}(\mathcal{W})$. Finally, by default $\|\cdot\|$ represents the Euclidean norm unless otherwise specified.

2. The maximum volume ellipsoid problem. Consider a polytope $\mathcal{P}$ in $\Re^n$ given by

$$\mathcal{P} = \{v \in \Re^n : Av \le b\}, \tag{2.1}$$

where $A \in \Re^{m\times n}$, $m > n$, and $b \in \Re^m$. Recall that by definition a polytope is a bounded polyhedron. For convenience of discussion, we will make the following two assumptions throughout the paper:

A1. The matrix A has full rank n and contains no zero rows.

A2. There exists a strictly interior point $\bar{v} \in \mathcal{P}$ satisfying $A\bar{v} < b$.

In this paper, we will also make the assumption that m is a small multiple of n, that is, $n < m \ll n^2$.

Given a center $x \in \Re^n$ and a nonsingular scaling matrix $E \in \Re^{n\times n}$, an ellipsoid in $\Re^n$ centered at x can be defined as

$$\mathcal{E}(x, E) = \{v \in \Re^n : (v - x)^T (EE^T)^{-1}(v - x) \le 1\};$$

or, equivalently,

$$\mathcal{E}(x, E) = \{v \in \Re^n : v = x + Es \text{ and } \|s\| \le 1\}, \tag{2.2}$$

where $\|\cdot\|$ is the Euclidean norm in $\Re^n$. Clearly, an ellipsoid is uniquely determined by, and uniquely determines, the symmetric positive definite matrix EE^T, but E is not uniquely determined since the same ellipsoid can also be generated by EQ for any orthogonal matrix $Q \in \Re^{n\times n}$. Without loss of generality, we make the assumption that E itself is symmetric positive definite. With this restriction, every (nondegenerate) ellipsoid will have a unique representation $\mathcal{E}(x, E)$.

It is easy to see that the ellipsoid $\mathcal{E}(x, E)$ is contained in $\mathcal{P}$ if and only if

$$\sup_{\|s\|=1} a_i^T (x + Es) \le b_i, \quad i = 1, \ldots, m,$$

where a_i^T is the ith row of A; or, equivalently,

$$a_i^T x + \|Ea_i\| \le b_i, \quad i = 1, \ldots, m.$$

Introducing the notation

$$h(E) = (\|Ea_1\|, \ldots, \|Ea_m\|)^T \in \Re^m, \tag{2.3}$$

we have

$$\mathcal{E}(x, E) \subset \mathcal{P} \Longleftrightarrow b - Ax - h(E) \ge 0. \tag{2.4}$$

Let V_n be the volume of the n-dimensional unit ball. Then the volume of the ellipsoid $\mathcal{E}(x, E)$ defined in (2.2) is

$$\mathrm{Vol}(\mathcal{E}) \equiv V_n \det E.$$

It is evident that $\mathcal{E}(x^*, E^*)$ is the maximum volume ellipsoid contained in $\mathcal{P}$ if and only if $(x^*, E^*) \in \Re^n \times \mathcal{S}^n_{++}$ solves the following optimization problem:

$$\begin{array}{ll} \min & -\log\det E \\ \text{subject to (s.t.)} & b - Ax - h(E) \ge 0 \\ & (E \succ 0), \end{array} \tag{2.5}$$

where $E \succ 0$ means that E is symmetric positive definite. (The constraint in parentheses may not need to be explicitly enforced.) It is well known that the optimization problem (2.5) is a convex program with a unique solution $(x^*, E^*) \in \Re^n \times \mathcal{S}^n_{++}$. Moreover, this solution is uniquely determined by the first-order optimality, or Karush–Kuhn–Tucker (KKT), conditions for the problem which can be derived as follows.

The Lagrangian function of the convex program (2.5) is

$$L(x, E, u) = -\log\det E - u^T (b - Ax - h(E)),$$

where $u \in \Re^m$ is the vector of Lagrange multipliers and $u \geq 0$. The KKT conditions consist of the equations $\nabla_x L = 0$, $\nabla_E L = 0$, feasibility, and complementarity. Using the differentiation formulas

$$\nabla[\log\det E] = E^{-1} \quad \text{and} \quad \nabla h_i(E) = \frac{E a_i a_i^T + a_i a_i^T E}{2h_i(E)}$$

and introducing the notation $U := \mathrm{Diag}(u)$ and

$$Y \equiv Y(E,u) := \mathrm{Diag}(h(E))^{-1} U, \tag{2.6}$$

we can write the KKT conditions as

$$A^T u = 0, \tag{2.7a}$$
$$E^{-1} - [E(A^T Y A) + (A^T Y A)E]/2 = 0, \tag{2.7b}$$
$$z - (b - Ax - h(E)) = 0, \tag{2.7c}$$
$$Uz = 0, \tag{2.7d}$$
$$u, z \geq 0, \tag{2.7e}$$

where $E \succ 0$ and z is a slack variable.

3. Formulations and primal-dual algorithms. In this section, we propose formulations and algorithms for effectively solving the MaxVE problem in practice. In constructing practically efficient algorithms, we consider the following three guidelines:

1. The algorithms should not carry the matrix-valued variable E as a completely independent variable because it would require too much computation (given that $n^2 \gg m$).
2. The algorithms should be primal-dual algorithms because of their proven practical efficiency in numerous cases.
3. The algorithms should have theoretical guarantees to be well defined and well behaved.

The first objective above can be achieved by eliminating the matrix variable E. The elimination may occur either at the beginning of a formulation or at the time of solving linear systems during iterations. In this paper, we will take the former approach.

3.1. Formulations without matrix variable. We now describe three formulations, first proposed in [26], for the MaxVE problem which are free of the matrix variable E. The key idea in these formulations is to eliminate the matrix-valued variable E from the system by solving (2.7b) for E. As can be verified easily, a solution to (2.7b) is

$$E(y) = (A^T Y A)^{-1/2}, \tag{3.1}$$

where $y = \mathrm{diag}(Y)$ and Y is defined in (2.6). We will later demonstrate that this solution is unique in $\mathcal{S}^n_{++}$. Upon the substitution of $E(y)$ into the definition of $h(y)$ (recall that $h_i(E) = \|Ea_i\|$), the vector $h(E)$ becomes a function of y that we will denote, with a slight abuse of notation, as $h(y)$; namely,

$$h(y) \equiv h(E(y)). \tag{3.2}$$

In [26], after substituting (3.1) and (3.2) into the KKT system, deleting (2.7b), and adding (2.6) written in a different form, i.e.,

$$u = g(y) := Yh(y), \tag{3.3}$$

the author obtained the following system:

$$F_0(x, y, u, z) = 0, \quad y, u, z \geq 0, \tag{3.4}$$

where $x \in \Re^n$, $y, u, z \in \Re^m$, and the function $F_0 : \Re^{n+3m} \to \Re^{n+3m}$ is

$$F_0(x, y, u, z) = \begin{bmatrix} A^T u \\ Ax + h(y) + z - b \\ u - g(y) \\ Uz \end{bmatrix}. \tag{3.5}$$

Moreover, it is proposed in [26] to eliminate the variable u from the above system using (3.3). The resulting system is

$$F_1(x, y, z) = 0, \quad y, z \geq 0, \tag{3.6}$$

where the function $F_1 : \Re^{n+2m} \to \Re^{n+2m}$ is

$$F_1(x, y, z) = \begin{bmatrix} A^T g(y) \\ Ax + h(y) + z - b \\ Zg(y) \end{bmatrix}. \tag{3.7}$$

In (3.5) and (3.7), we have used the notation $U = \mathrm{Diag}(u)$ and $Z = \mathrm{Diag}(z)$, respectively.

In addition, the complementarity conditions $Uz = 0$ are clearly equivalent to the conditions $Yz = 0$ because $U = Y\mathrm{Diag}(h(y))$ and $h(y) > 0$ at the solution. Based on this observation, a third system is proposed in [26]:

$$F_2(x, y, z) = 0, \quad y, z \geq 0, \tag{3.8}$$

where the function $F_2 : \Re^{n+2m} \to \Re^{n+2m}$ is

$$F_2(x, y, z) = \begin{bmatrix} A^T g(y) \\ Ax + h(y) + z - b \\ Yz \end{bmatrix}. \tag{3.9}$$

The three systems (3.4), (3.6), and (3.8) are all free of the matrix-valued variable E and will form the bases for our algorithm construction.[1] However, in obtaining them we have applied nonlinear transformations whose properties need to be investigated. A most important question is whether or not these transformations preserve the uniqueness of solutions. We will answer this question and others in a subsequent section.

[1] In [26], some additional systems were also derived that we have found to be less satisfactory.

3.2. Primal-dual algorithmic framework. The primal-dual algorithms to be proposed can be motivated from the view of the damped Newton's method applied to the so-called perturbed complementarity conditions. Another useful perspective is to view them as path-following algorithms. In this construction, one replaces the zero right-hand side of relevant complementarity conditions with μw^0, where $\mu > 0$ and $w^0 \in \Re^m_{++}$, and applies the Newton method to the resulting "perturbed" system while decreasing the parameter μ to zero. Specifically, the perturbed systems for (3.6) and (3.8) have the form

$$F(x, y, z) = \begin{bmatrix} 0 \\ 0 \\ w \end{bmatrix}, \quad y, z > 0, \tag{3.10}$$

where F can be either F_1 or F_2, and for some $w^0 \in \Re^m_{++}$

$$w = \mu w^0, \quad \mu > 0.$$

Normally, one chooses $w^0 = e$, where e is the vector of all ones.

We will prove later that each of the perturbed systems has a unique solution for every $\mu > 0$, and as $\mu \to 0$ the corresponding solutions will converge to the (same) solution of the unperturbed systems from which the solution to the MaxVE problem can be easily constructed.

We now present our primal-dual interior-point algorithmic framework for the systems (3.6) and (3.8). The framework for the system (3.4) would be the same except that an extra variable $u \in \Re^m$ is present. In the rest of the paper, we will concentrate only on the formulations (3.6) and (3.8) but omit (3.4) because, being so closely related to (3.6), system (3.4) shares almost identical theoretical properties with (3.6), while in our tests it seems to produce algorithms with performance inferior to that of their counterparts based on (3.6) and (3.8).

ALGORITHM 1 (primal-dual interior-point algorithm).

Given x^0 in the interior of $\mathcal{P}$ and $y^0, z^0 \in \Re^m_{++}$, set $k = 0$.

Step 1. *Choose $\sigma_k \in (0, 1)$, set μ_k to $\sigma_k \frac{g(y^k)^T z^k}{m}$ for $F = F_1$ or to $\sigma_k \frac{(y^k)^T z^k}{m}$ for $F = F_2$.*

Step 2. *Solve for (dx, dy, dz) from*

$$F'(x^k, y^k, z^k) \begin{bmatrix} dx \\ dy \\ dz \end{bmatrix} = \mu_k \begin{bmatrix} 0 \\ 0 \\ e \end{bmatrix} - F(x^k, y^k, z^k). \tag{3.11}$$

Step 3. *Choose a step length $\alpha_k \in (0, 1]$ and update*

$$(x^{k+1}, y^{k+1}, z^{k+1}) = (x^k, y^k, z^k) + \alpha_k (dx, dy, dz)$$

such that $x^{k+1} \in \mathcal{P}$, $y^{k+1} > 0$ and $z^{k+1} > 0$.

Step 4. *If $\|F(x^{k+1}, y^{k+1}, z^{k+1})\| \le \epsilon$, stop; else increment k and go to Step 1.*

In addition to the initial guesses, this algorithmic framework has two essential parameters, σ_k and α_k, that need to be specified at each iteration. The main computation required is to solve the linear system (3.11) at every iteration.

When $F = F_1$, the coefficient matrix in the linear system (3.11), i.e., the Jacobian matrix of $F_1(x, y, z)$, is of the form

$$F_1'(x,y,z) = \begin{bmatrix} 0 & A^T g'(y) & 0 \\ A & h'(y) & I \\ 0 & Zg'(y) & \mathrm{Diag}(g(y)) \end{bmatrix}, \tag{3.12}$$

where $g'(y)$ and $h'(y)$ are the Jacobian matrices of $g(y)$ and $h(y)$, respectively. A direct differentiation shows that

$$g'(y) = H(y) + Yh'(y) \tag{3.13}$$

and (see also [26])

$$h'(y) = -\frac{1}{2}H(y)^{-1}[Q(y) \circ Q(y)], \tag{3.14}$$

where

$$H \equiv H(y) := \mathrm{Diag}(h(y)), \;\; Q \equiv Q(y) = A(A^T Y A)^{-1} A^T. \tag{3.15}$$

It is worth noting that $Y^{1/2}Q(y)Y^{1/2}$ is an orthogonal projection matrix.

On the other hand, when $F = F_2$ we have

$$F_2'(x,y,z) = \begin{bmatrix} 0 & A^T g'(y) & 0 \\ A & h'(y) & I \\ 0 & Z & Y \end{bmatrix}. \tag{3.16}$$

An efficient way to solve the linear system (3.11) is the following block Gaussian elimination procedure: first eliminating dz and dy, then solving for dx, finally computing dy and dz by back substitutions. We now formally describe the procedure for $F = F_1$. To simplify notation, we define the following two $m \times m$ matrices:

$$N \equiv N(y) := g'(y) \tag{3.17}$$

and

$$M_1 \equiv M_1(y,z) := -h'(y) + [YH(y)]^{-1} ZN(y). \tag{3.18}$$

For now we will assume that M_1 is nonsingular, and we will prove this fact later.

The aforementioned block Gaussian elimination reduces $F_1'(x, y, z)$ into a lower triangular matrix, which is equivalent to, when $F = F_1$, premultiplying (3.11) by the upper triangular elimination matrix

$$T_1(y,z) = \begin{bmatrix} I & A^T N M_1^{-1} & -A^T N M_1^{-1} (YH)^{-1} \\ 0 & I & -(YH)^{-1} \\ 0 & 0 & I \end{bmatrix}.$$

It is straightforward to verify that

$$T_1(y,z) F_1'(x,y,z) = \begin{bmatrix} A^T N M_1^{-1} A & 0 & 0 \\ A & -M_1 & 0 \\ 0 & ZN & YH \end{bmatrix} \tag{3.19}$$

and for any vectors $r_1 \in \Re^n$ and $r_2, r_3 \in \Re^m$

$$T_1(y,z)\begin{pmatrix} r_1 \\ r_2 \\ r_3 \end{pmatrix} = \begin{pmatrix} r_1 + A^T N M_1^{-1}\left(r_2 - (YH)^{-1} r_3\right) \\ r_2 - (YH)^{-1} r_3 \\ r_3 \end{pmatrix}. \tag{3.20}$$

Clearly, the linear system

$$F_1'(x,y,z)\begin{pmatrix} dx \\ dy \\ dz \end{pmatrix} = \begin{pmatrix} r_1 \\ r_2 \\ r_3 \end{pmatrix}$$

is equivalent to the linear system where the coefficient matrix is the one in (3.19) and the right-hand side is that of (3.20). This linear system can be formally solved by the following procedure:

$$dx = [A^T N M_1^{-1} A]^{-1}\left(r_1 + A^T N M_1^{-1}(r_2 - (YH)^{-1} r_3)\right), \tag{3.21a}$$

$$dy = -M_1^{-1}\left(r_2 - (YH)^{-1} r_3 - A dx\right), \tag{3.21b}$$

$$dz = (YH)^{-1}(r_3 - ZN dy). \tag{3.21c}$$

This solution procedure requires $\mathcal{O}(m^3)$ operations (recall that $m > n$), with the bulk of the computation involving the $m \times m$ matrix M_1.

Similarly, the linear system (3.11) corresponding to $F = F_2$ can be formally solved by the following procedure:

$$dx = [A^T N M_2^{-1} A]^{-1}(r_1 + A^T N M_2^{-1}\left(r_2 - Y^{-1} r_3\right)), \tag{3.22a}$$

$$dy = -M_2^{-1}\left(r_2 - Y^{-1} r_3 - A dx\right), \tag{3.22b}$$

$$dz = Y^{-1}(r_3 - Z dy), \tag{3.22c}$$

where

$$M_2 \equiv M_2(y,z) := -h'(y) + Y^{-1} Z. \tag{3.23}$$

This procedure also requires $\mathcal{O}(m^3)$ operations in terms of the order but less linear algebra computation than required by procedure (3.21a)–(3.21c).

Of course, we still need to establish in theory that the proposed primal-dual algorithms are well defined. To this end, we need to show that the matrices $F_i'(x,y,z)$ are nonsingular for any $y, z > 0$, and the matrices M_i and $A^T N M_i^{-1} A$ are also nonsingular for both $i = 1$ and 2. These results will be presented next.

4. Theoretical results. In this section, we give theoretical results regarding the well-definedness of the proposed algorithms, the uniqueness of solutions in our formulations, as well as the existence and convergence of solution paths. We note that the formulations introduced in the last section are obtained by applying some nonlinear transformations. Therefore we need to show that these nonlinear transformations preserve the uniqueness of solution. We also mention that when $F = F_2$, the system in (3.10) is not equivalent to the optimality conditions of a convex program. Hence, it is not evident that solution paths defined by (3.10) should always exist for $F = F_2$.

4.1. Well-definedness of algorithms. We will show in this subsection that the proposed primal-dual algorithmic framework and the solution procedures (3.21a)–(3.21c) and (3.22a)–(3.22c) are well defined for both $F = F_1$ and $F = F_2$. (Following the same approach, one can also easily verify similar results for $F = F_0$.)

We recall that throughout the paper we have assumed that A has full rank with no zero rows. The main result of this subsection is the following theorem.

THEOREM 4.1 (nonsingularity of Jacobian). *For any $y, z > 0$, the Jacobian matrices $F_i'(x, y, z)$ are nonsingular for $i = 1, 2$. Moreover, both the procedures* (3.21a)–(3.21c) *and* (3.22a)–(3.22c) *are well defined.*

Proof. The theorem follows directly from Lemma 4.4 below. □

Now we prove three technical results that will lead to the proof of Theorem 4.1.

LEMMA 4.2. *Let $P \in \Re^{n\times n}$ be an orthogonal projection matrix; i.e., P satisfies $P^T = P$ and $P^2 = P$. Then the symmetric matrix*

$$G_\gamma = I\circ P - \gamma P\circ P \tag{4.1}$$

is positive semidefinite for any $\gamma \le 1$. Moreover, if $\mathrm{diag}(P) > 0$*, then G_γ is positive definite for any $\gamma < 1$.*

Proof. We note that since $I \succeq P \succeq 0$, i.e., both P and $I - P$ are symmetric positive semidefinite, so are $P\circ P$ and $(I - P)\circ P$ because the Hadamard products of positive semidefinite matrices are also positive semidefinite (see, for example, [6]). $G\gamma$ is obviously positive semidefinite for $\gamma \le 0$. Using the identity

$$G_\gamma = \gamma(I - P)\circ P + (1 - \gamma)I\circ P,$$

we see that G_γ is a convex combination of two positive semidefinite matrices for $\gamma \in [0, 1]$, and hence is positive semidefinite. The second statement follows from the conditions $\mathrm{diag}(P) > 0$ and $\gamma < 1$ which ensure that the second term above is positive definite. □

LEMMA 4.3. *For any $y > 0$, the matrix $N(y) \equiv g'(y)$ is similar to a symmetric positive definite matrix, and thus is nonsingular.*

Proof. We first note $h(y) > 0$ whenever $y > 0$. In view of (3.17), (3.13), and (3.14),

$$\begin{aligned} N &= H - (2H)^{-1}Y[Q\circ Q] \\ &= H^{-1}\left(HYH - \frac{1}{2}Y[Q\circ Q]Y\right)Y^{-1} = H^{-1}GY^{-1} \\ &= [H^{-1/2}Y^{1/2}]\left([HY]^{-1/2}G[YH]^{-1/2}\right)[H^{-1/2}Y^{1/2}]^{-1}, \end{aligned}$$

where

$$G := HYH - \frac{1}{2}Y[Q\circ Q]Y. \tag{4.2}$$

Therefore, N is similar to $[HY]^{-1/2}G[YH]^{-1/2}$, which is positive definite if and only if the matrix G is positive definite since both Y and H are positive diagonal matrices.

Recall that by our notation $Q = A(A^TYA)^{-1}A^T$, $H = \mathrm{Diag}(h(y))$, and

$$h(y) \equiv h(E(y)) = (\mathrm{diag}(Q(y))^{1/2},$$

where the square root is taken elementwise. We have

$$HYH = (I\circ Q)Y = I\circ (Y^{1/2}QY^{1/2}).$$

In addition, since $y_i Q_{ij}^2 y_j = \left(\sqrt{y_i}\, Q_{ij} \sqrt{y_j}\right)^2$, we have

$$Y[Q \circ Q]Y = \left(Y^{1/2} Q Y^{1/2}\right) \circ \left(Y^{1/2} Q Y^{1/2}\right).$$

Therefore we can write

$$G = I \circ P - \frac{1}{2} P \circ P,$$

where the matrix

$$P = Y^{1/2} Q Y^{1/2} = Y^{1/2} A (A^T Y A)^{-1} A^T Y^{1/2}$$

is an orthogonal projection matrix. Since the vector y is positive and the matrix A has no zero rows, we have $\operatorname{diag}(P) > 0$. It follows from Lemma 4.2 with $\gamma = 1/2$ that G is indeed positive definite. This completes the proof. ∎

The relationships

$$N = H^{-1} G Y^{-1} \quad \text{and} \quad N^{-1} = Y G^{-1} H \tag{4.3}$$

that were used in the proof of Lemma 4.3 will be useful later.

LEMMA 4.4. *For $y, z > 0$, there hold the following statements:*

1. *The matrix M_1 is similar to a symmetric positive definite matrix, and $A^T N M_1^{-1} A$ is symmetric positive definite.*
2. *The matrix M_2 is similar to a symmetric positive definite matrix, and $A^T N M_2^{-1} A$ is nonsingular.*

Proof. To prove the first statement, it suffices to prove that the matrix $M_1 N^{-1}$ is symmetric positive definite. Using the definitions of M_1, N, and the formula for g' (see (3.18), (3.17), and (3.13), respectively), and the relationships (4.3), we have $h' = Y^{-1}(N - H)$ and

$$\begin{aligned} M_1 N^{-1} &= ((YH)^{-1} Z N - h') N^{-1} \\ &= (YH)^{-1} Z - Y^{-1}(N - H) N^{-1} \\ &= (YH)^{-1} Z - Y^{-1} + Y^{-1} H N^{-1} \\ &= (YH)^{-1} Z - Y^{-1} + Y^{-1} H (Y G^{-1} H) \\ &= (YH)^{-1} Z - Y^{-1} + H G^{-1} H \\ &= (YH)^{-1} Z + H \left(G^{-1} - (HYH)^{-1}\right) H. \end{aligned}$$

Then it suffices to show that $G^{-1} - (HYH)^{-1}$ is symmetric positive definite since H, Y, and Z are all positive diagonal matrices. While the symmetry is obvious, the positive definiteness follows from the fact that G equals HYH minus a positive semidefinite matrix; see (4.2); hence $G \prec YHY$ and $G^{-1} \succ (YHY)^{-1}$ (see [5], for example).

To prove the second statement, we use the formula for $h'(y)$ in (3.14) to obtain

$$M_2 = Y^{-1} Z - h' = H^{-1} \left(HY^{-1} Z + \frac{1}{2} Q \circ Q\right),$$

which is the product of two symmetric positive definite matrices, implying that M_2 is similar to a symmetric positive definite matrix. Since both M_2 and N are nonsingular, so is $A^T N M_2^{-1} A$. This completes the proof. ∎

4.2. Uniqueness of solution. Since we have utilized nonlinear transformations in the elimination of variables $E = E(y)$ and $u = g(y)$ from the KKT system (2.7a)–(2.7d), we need to establish a rigorous equivalence of our formulations (3.6) and (3.8) to the original KKT system. The main result is the following.

THEOREM 4.5 (uniqueness of solution). *The systems* (3.6) *and* (3.8) *have the same, unique solution* (x^*, y^*, z^*) *such that* $y^*, z^* \geq 0$. *Moreover, let* $u^* = g(y^*)$ *and* $E^* = E(y^*)$. *Then* (x^*, E^*, u^*, z^*) *is the unique solution of the KKT conditions* (2.7a-)–(2.7e).

Proof. The conclusions follow directly from Lemmas 4.6 and 4.7, given below, and the uniqueness of the solution to the MaxVE problem. ∎

We now prove the two technical lemmas.

LEMMA 4.6. *Let* $C \in \mathcal{S}^n_{++}$; *then the matrix equation*

$$X^{-1} = \frac{1}{2}(CX + XC) \tag{4.4}$$

has a unique solution $X^* = C^{-1/2}$ *in* $\mathcal{S}^n_{++}$. *Moreover, the mapping:* $C \to X^*$ *defined implicitly through* (4.4) *is homeomorphic between* $\mathcal{S}^n_{++}$ *and itself.*

Proof. One can easily verify that both X^* and $-X^*$ are solutions to (4.4). This implies that the matrix equation (4.4) does not in general have a unique solution in $\Re^{n\times n}$.

Suppose that $\hat{X} \in \mathcal{S}^n_{++}$ is a solution to (4.4) and U is an orthogonal matrix that diagonalizes $\hat{X}$, i.e., $U^T\hat{X}U = \Sigma$, where Σ is a positive diagonal matrix. Premultiplying both sides of (4.4) by U^T and postmultiplying them by U, we obtain

$$\Sigma^{-1} = \frac{1}{2}(D\Sigma + \Sigma D),$$

where $D = U^T C U$. Comparing the elements on both sides, we have

$$\frac{1}{2}D_{ij}(\Sigma_{ii} + \Sigma_{jj}) = \begin{cases} 0, & i \neq j, \\ 1/\Sigma_{ii}, & i = j. \end{cases}$$

Since $\mathrm{diag}(\Sigma) > 0$, we must have (i) $D_{ij} = 0$ for $i \neq j$ and (ii) $\Sigma_{ii} = D_{ii}^{-1/2}$. The first relationship says that $D = U^T C U$ is also diagonal. The second relationship says that $\Sigma = D^{-1/2}$, that is, $\hat{X} = C^{-1/2} \equiv X^*$. Consequently, X^* is the only solution of (4.4) in $\mathcal{S}^n_{++}$.

The last statement of the lemma is evident in view of the explicit relationships $X^* = C^{-1/2}$ and $C = (X^*)^{-2}$. ∎

LEMMA 4.7. *Let* $g(y) \equiv Yh(y)$. *Then the mapping* $g : \Re^m_{++} \to \Re^m_{++}$ *is homeomorphic between* $\Re^m_{++}$ *and its image under* g, *i.e.,* $g(\Re^m_{++}) \subset \Re^m_{++}$.

Proof. It is straightforward to verify that the function $g(y)$ is continuously differentiable in $\Re^m_{++}$, whose derivative is represented by the matrix $g'(y) \equiv N(y)$. By Lemma 4.3, $N(y)$ is nonsingular in $\Re^m_{++}$. With these properties, the lemma is a direct consequence of the inverse function theorem. ∎

4.3. Existence and convergence of solution paths. To justify our algorithms as the path-following type, we will show that (i) the perturbed system (3.10) with either $F = F_1$ or $F = F_2$ permits a unique solution for any given $w^0 \in \Re^m_{++}$ and each $\mu > 0$, and hence the solution set forms a path; and (ii) as $\mu \to 0$ the path converges to the unique solution of the unperturbed system. Although it is straightforward to establish these results in the case of $F = F_1$, it is much more involved in

the case of $F = F_2$ since the perturbed system (3.10) for $F = F_2$ does not correspond to the optimality conditions of a convex program.

Following the conventional terminology in the literature of interior-point methods, we will refer to the collection of solutions to the system (3.10) for $w^0 = e$ and $\mu > 0$ as the *central path* of the system, where $e \in \Re^m$ is the vector of all ones. Our analysis in this subsection applies to not only the central path but also to so-called weighted paths where $w^0 > 0$ is not equal to e.

The existence of paths for $F = F_1$ follows a standard argument as given below.

PROPOSITION 4.8 (existence and convergence of path for $F = F_1$). *For any $w^0 \in \Re^m_{++}$ and $\mu > 0$, the system* (3.10) *with $F = F_1$ has a unique solution $(x(\mu), y(\mu), z(\mu))$ such that $y(\mu), z(\mu) > 0$. Moreover,*

$$\lim_{\mu \to 0} (x(\mu), y(\mu), z(\mu)) = (x^*, y^*, z^*),$$

where (x^*, y^*, z^*) is the solution of (3.6).

Proof. The proof follows from a standard argument which we will outline as follows. It is well known that the system of the "perturbed" KKT (PKKT) conditions

$$A^T u = 0, \tag{4.5a}$$

$$E^{-1} - [E(A^T Y A) + (A^T Y A)E]/2 = 0, \tag{4.5b}$$

$$z - (b - Ax - h(E)) = 0, \tag{4.5c}$$

$$Uz = w, \tag{4.5d}$$

$$u, z > 0, \tag{4.5e}$$

has a unique solution for any $w > 0$, where Y is defined as in (2.6), because it is equivalent to the condition that the gradient of the barrier function $B_w(x, E)$ equals zero, where

$$B_w(x, E) = -\log\det(E) - \sum_{i=1}^{m} w_i \log\left(b_i - a_i^T x - h_i(E)\right). \tag{4.6}$$

This barrier function is strongly convex and has a unique stationary point $(x(\mu), E(\mu))$ corresponding to $w = \mu w^0$ for a fixed $w^0 \in \Re^m_{++}$ and any $\mu > 0$, which, together with the dual variable $u(\mu)$ and the slack variable $z(\mu)$, satisfies (4.5a)–(4.5e) for $w = \mu w^0$. This can be seen as follows. From (4.5c) and (4.5d), we obtain $u = \mathrm{Diag}(b - Ax - h(E))^{-1} w$. Substituting the expressions of y and u into (4.5a) and (4.5b), we obtain the partial gradient of $B_w(x, E)$ with respect to x and E, respectively. It is well known that $(x(\mu), E(\mu), u(\mu), z(\mu))$ converges to the unique solution (x^*, E^*, u^*, z^*) of the (unperturbed) KKT system as $\mu \to 0$. Due to the homeomorphic relationships between the PKKT conditions and the conditions in (3.10) with $F = F_1$, we know that $(x(\mu), y(\mu), z(\mu))$, where $y(\mu) = \mathrm{Diag}(h(E(\mu)))^{-1} u(\mu)$ is also the unique solution of (3.10) with $F = F_1$. Moreover, the path $\{(x(\mu), y(\mu), z(\mu)) : \mu > 0\}$ converges to (x^*, y^*, z^*), where $y^* = \mathrm{Diag}(h(E^*))^{-1} u^*$. ∎

We now consider the existence of solution to the system (3.10) when $F = F_2$, that is, the existence of solution to the system

$$A^T g(y) = 0, \tag{4.7a}$$

$$Ax + h(y) + z - b = 0, \tag{4.7b}$$

$$Yz = w, \tag{4.7c}$$

$$y, z > 0, \tag{4.7d}$$

where $w \in \Re^m_{++}$ and $g(y)$ is defined as in (3.3). The situation here is more complicated because this system is no longer equivalent to the PKKT conditions (4.5a)–(4.5e) when $w > 0$, even though they are equivalent when $w = 0$. As such, we can no longer use the standard argument used in the proof of Proposition 4.8. The question is whether or not the following holds:

$$\{0 \in \Re^n\} \times \{0 \in \Re^m\} \times \Re^m_{++} \subset \ \mathcal{R}(F_2),$$

where

$$\mathcal{R}(F_2) := F_2(\Re^n \times \Re^m_{++} \times \Re^m_{++})$$

is the range of the function F_2 corresponding to the domain $\Re^n \times \Re^m_{++} \times \Re^m_{++}$. In particular, we want to know if the vectors $(0, 0, \mu e)$ for $\mu > 0$ are in the range of F_2, in other words whether a central path exists for the system (3.10) in the case of $F = F_2$.

The answer to the above question is affirmative and given in Theorem 4.14 which we will prove now. There is a strong possibility that we can prove this theorem by identifying and verifying a set of conditions under which an existing general result is applicable to problem (3.8)—an instance of the so-called nonlinear mixed complementarity problem for which a number of potentially applicable results exist (for example, in [15]). However, we choose to provide an elementary and self-contained proof in this paper. We start with the following proposition stating some useful facts.

PROPOSITION 4.9. *The following facts hold:*

1. *Both F_1 and F_2 are locally homeomorphic at any point $(x, y, z) \in \Re^n \times \Re^m_{++} \times \Re^m_{++}$.*
2. *If $(\hat{x}, \hat{y}, \hat{z})$ is the solution to the system (3.10) with $F = F_1$ and $w = \hat{w}$, then $(\hat{x}, \hat{y}, \hat{z})$ also satisfies (3.10) with $F = F_2$ (i.e., (4.7a)–(4.7d)) and $w = \mathrm{Diag}(h(\hat{y}))^{-1}\hat{w}$.*

We note that the local homeomorphism of F_i implies that corresponding to any point $(u, v, w) \in F_i(\Re^n \times \Re^m_{++} \times \Re^m_{++})$, $i = 1, 2$, there is a unique point $(x, y, z) \in \Re^n \times \Re^m_{++} \times \Re^m_{++}$ such that $F_i(x, y, z) = (u, v, w)$.

If one were able to choose $\hat{w}$ such that $\mathrm{Diag}(h(\hat{y}))^{-1}\hat{w} = \mu e$, then the point $(0, 0, \mu e)$ would be in the range of F_2. However, since $\hat{y}$ is dependent on $\hat{w}$, it is not clear whether or not such a vector $\hat{w}$ exists, let alone how to find it. Nevertheless, we do find that points of the form $(0, 0, w)$ with $w = \mathrm{Diag}(h(\hat{y}))^{-1}\hat{w}$ are in the range of F_2.

LEMMA 4.10. *Let $x \in \Re^n$, $E \in \mathcal{S}^n_{++}$, and $z \in \Re^m_+$ satisfy the equation*

$$Ax + h(E) + z = b. \tag{4.8}$$

Then there exists a constant $\gamma > 0$, independent of x, E, and z, such that

$$\max(\|x\|, \|E\|, \|z\|) \le \gamma.$$

Proof. Equation (4.8) implies that $x \in \mathcal{P}$, where $\mathcal{P}$ is the given polytope; hence such x's must be uniformly bounded above. Consequently, $b - Ax$ for $x \in \mathcal{P}$ is also uniformly bounded above, which in turn implies that both z and $h(E)$ are uniformly bounded above because they are both nonnegative and they sum up to $b - Ax$. Since $h_i(E) = (a_i^T E^2 a_i)^{1/2}$ and, by our assumption, the set $\{a_1, a_2, \ldots, a_m\}$ spans $\Re^n$, the uniform boundedness of $h(E)$ implies that of E. This completes the proof. ∎

LEMMA 4.11. *Let the barrier function $B_w(x, E)$ be defined as in (4.6), and let $\mathcal{W}$ be a bounded set with its closure $\mathrm{cl}(\mathcal{W})$ in $\Re^m_{++} \cup \{0\}$. For any $w \in \Re^m_{++}$, define*

$$(x_w, E_w) := \arg\min B_w(x, E) \tag{4.9}$$

and for $w = 0 \in \Re^m$ *define* $(x_w, E_w) := (x^*, E^*)$ *as the solution of the MaxVE problem* (2.5). *Then*

$$\beta_{\mathcal{W}} := \inf_{w \in \mathrm{cl}(\mathcal{W})} \{\log\det(E_w)\} > -\infty.$$

Proof. Since the pair (x_w, E_w), $E_w \succ 0$, is the unique minimizer of $B_w(x, E)$, there exists some $(u_w, z_w) \in \Re^m_{++} \times \Re^m_{++}$ such that together they satisfy (4.5a)–(4.5e). It is well known that the quadruple (x_w, E_w, u_w, z_w) is a continuous function of w in $\Re^m_{++}$ and that (x_w, E_w, u_w, z_w) converges to (x^*, E^*, u^*, z^*) as w converges to 0 from the interior of $\Re^m_{++}$. Hence, the composite function $\log\det(E_w)$ of w is a continuous function of w in $\Re^m_{++} \cup \{0\}$ and must attain its minimum on the compact set $\mathrm{cl}(\mathcal{W}) \subset \Re^m_{++} \cup \{0\}$. This proves the lemma. ∎

LEMMA 4.12. *Let* $\mathcal{R}(F_2)$ *be the range of the function* F_2 *corresponding to the domain* $\Re^n \times \Re^m_{++} \times \Re^m_{++}$, *and let* $\mathcal{W}$ *be a bounded set in* $\Re^m_{++}$ *such that its closure* $\mathrm{cl}(\mathcal{W}) \subset \Re^m_{++} \cup \{0\}$. *Let*

$$\{0 \in \Re^n\} \times \{0 \in \Re^m_{++}\} \times \mathcal{W} \subset \ \mathcal{R}(F_2),$$

and let $(x(w), y(w), z(w))$ *be the solution to* (4.7a)–(4.7c) *corresponding to* $w \in \mathcal{W}$. *Then the set* $\{y(w) : w \in \mathcal{W}\}$ *is bounded.*

Proof. The triple $(x(w), y(w), z(w))$ being the solution to (4.7a)–(4.7c) implies that the quadruple

$$(x_{w'}, E_{w'}, u_{w'}, z_{w'}) := (x(w), E(y(w)), g(y(w)), z(w))$$

is the solution to (4.5a)–(4.5e) with the right-hand side of (4.5d) being replaced by $w' = \mathrm{Diag}(h(y(w)))w$. It is worth noting that the pair $(x_{w'}, E_{w'})$ also satisfies (4.9) with $w = w'$. Evidently, we have

$$E_{w'} \equiv E(y(w)).$$

Define the set

$$\mathcal{W}' := \{w' = \mathrm{Diag}(h(y_w))w : w \in \mathcal{W}\},$$

which is bounded because both $\mathcal{W}$ and the set of $\{h(y(w)) : w \in \mathcal{W}\}$ are bounded. It follows from Lemma 4.10 that the set

$$\{E_{w'} : w' \in \mathcal{W}'\} \equiv \{E(y(w)) : \ w \in \mathcal{W}\}$$

is bounded. Hence, the eigenvalues of $E(y(w))$ are uniformly bounded above. On the other hand, Lemma 4.11 implies that

$$\log\det(E(y(w))) \geq \beta_{\mathcal{W}}.$$

As a result, the eigenvalues of $E(y(w))$ are also uniformly bounded away from zero in the set $\mathcal{W}$. Consequently, the components of $h(y(w))$ are uniformly bounded above and away from zero in the set $\mathcal{W}$ because $h_i(y(w)) = (a_i^T E(y(w)) a_i)^{1/2}$ and the rows a_i^T of A are all nonzero for $i = 1, \ldots, m$.

We note that the vector $\mathrm{Diag}[h(y(w))]^2 y(w)$ is the diagonal of the orthogonal projection matrix $Y(w)^{1/2} A [A^T Y(w) A]^{-1} A^T Y(w)^{1/2}$ and therefore is componentwise bounded above by unity; namely,

$$y_i(w) \leq \frac{1}{h_i(y(w))^2}, \quad i = 1, 2, \ldots, m.$$

Since $h(y(w))$ is uniformly bounded away from zero for $w \in \mathcal{W}$, we conclude that $y(w)$ is uniformly bounded above for $w \in \mathcal{W}$. This completes the proof. ∎

LEMMA 4.13. *Let* $\mathcal{R}(F_2)$ *be defined as in Lemma* 4.12. *Then*

$$\{0 \in \Re^n\} \times \{0 \in \Re^m\} \times \Re^m_{++} \subset \mathcal{R}(F_2).$$

Proof. From the second statement of Proposition 4.9, we know that there exists a triple $(0, 0, w^\alpha) \in \mathcal{R}(F_2)$ for some $w^\alpha \in \Re^m_{++}$. Now for any given $w^\beta \in \Re^m_{++}$, we are to show that $(0, 0, w^\beta) \in \mathcal{R}(F_2)$.

Let us define the line segment between w^α and w^β,

$$w(t) = (1 - t)w^\alpha + t\, w^\beta,$$

and the number

$$\hat{t} = \sup\{t \in [0, 1] : \{(0, 0, w(t')) : t' \in [0, t]\} \subset \mathcal{R}(F_2)\}.$$

Since $(0, 0, w(0)) \in \mathcal{R}(F_2)$ and F_2 is homeomorphic between $\Re^n \times \Re^m_{++} \times \Re^m_{++}$ and $\mathcal{R}(F_2)$, we must have $\hat{t} > 0$. If $\hat{t} = 1$, we already have $w^\beta \in \mathcal{R}(F_2)$ and we are done.

Now suppose $\hat{t} < 1$. This implies that $(0, 0, w(\hat{t})) \notin \mathcal{R}(F_2)$; otherwise by the local homeomorphism of F_2 the number $\hat{t}$ would not have been a supremum. Consider the set

$$\mathcal{W} := \{w(t) : t \in [0, \hat{t})\} \subset \mathcal{R}(F_2),$$

which is clearly bounded with its closure $\mathrm{cl}(\mathcal{W})$ in $\Re^m_{++}$. It follows from Lemmas 4.10 and 4.12, that the set

$$\{(x(w), y(w), z(w) : w \in \mathcal{W}\}$$

is also bounded. Let us denote $x(w(t))$ by $x(t)$, and so on. Then there must exist a sequence $\{t_k\}_{k=1}^\infty$ such that $t_k \to \hat{t}$ and $(x(t_k), y(t_k), z(t_k)) \to (\hat{x}, \hat{y}, \hat{z})$ for some $(\hat{x}, \hat{y}, \hat{z}) \in \Re^n \times \Re^m_+ \times \Re^m_+$. (Otherwise, a convergent subsequence can be selected.)

Since the function F_2 is continuous, we have $F_2(\hat{x}, \hat{y}, \hat{z}) = (0, 0, w(\hat{t}))^T$, meaning that $(0, 0, w(\hat{t})) \in \mathcal{R}(F_2)$. This is a contradiction. Thus the assumption $\hat{t} < 1$ is false, and we have proved the lemma. ∎

Finally we prove the existence and convergence of solution paths, including the central path, leading to the solution of the original MaxVE problem in the sense specified in the following theorem.

THEOREM 4.14 (existence and convergence of path for $F = F_2$). *For any* $w^0 \in \Re^m_{++}$ *and* $\mu > 0$, *the system* (3.10) *with* $F = F_2$ *and* $w = \mu w^0$ *has a unique solution* $(x(\mu), y(\mu), z(\mu))$. *Moreover,*

$$\lim_{\mu \to 0}(x(\mu), y(\mu), z(\mu), u(\mu), E(\mu)) = (x^*, y^*, z^*, u^*, E^*),$$

where (x^*, y^*, z^*) *satisfies the system* (3.8), *and* (x^*, E^*, u^*, z^*) *satisfies the KKT system* (2.7a)–(2.7e). *Consequently,* (x^*, E^*) *solves the MaxVE problem* (2.5).

Proof. The first statement follows directly from Lemma 4.13 and the fact that F_2 is homeomorphic in $\Re^n \times \Re^m_{++} \times \Re^m_{++}$.

By Lemmas 4.10 and 4.12, the quantities $x(\mu), y(\mu), z(\mu), u(\mu)$, and $E(\mu)$ are all bounded as $\mu \to 0$. Hence, they must have accumulation points as $\mu \to 0$, say,

x^*, y^*, z^*, u^*, and E^*. Clearly, these accumulation points satisfy the two systems in the theorem. Since these systems permit only unique solutions, we conclude that all accumulation points of $x(\mu)$ as $\mu \to 0$ must coincide, and the same is true for other quantities as well; namely, accumulation points are actually the limit point. Obviously, x^* and E^* solve the optimization problem (2.5) because they, together with u^* and z^*, satisfy the optimality conditions (2.7a)–(2.7e). This proves the theorem. □

4.4. Issues of algorithmic convergence. So far polynomial convergence theory for primal-dual interior point algorithms has been established only for convex conic programming in symmetric cones (see [18], for example), with the exception of Nemirovskii [16]. Given the highly nonlinear formulations upon which we build our primal-dual interior-point algorithms, it seems unlikely that polynomial convergence could be proven for our primal-dual algorithms unless some new technique is discovered.

On the other hand, performing some nonpolynomial, global convergence analysis for the proposed algorithmic framework appears to be a worthy task. Given the good theoretical properties we have already established for our formulations, we do not see any fundamental difficulty in proving global and fast local convergence for some parameter choices in the proposed algorithmic framework (for example, following the approach in [3]). Such an analysis, however, would be rather lengthy and technical. To keep the current paper focused and within a reasonable length, we will not attempt a convergence analysis in this paper.

5. Khachiyan–Todd algorithm and modification. We will introduce two other algorithms, the Khachiyan and Todd algorithm [11] and a modification of it, and will later compare them with algorithms proposed in section 3.

Given a set of inequalities $Ax \le b$ and a strictly interior point x^0, using the change of variable $x = v + x^0$, we can rewrite the inequalities as $Av \le b - Ax^0$. By multiplying both sides by the positive diagonal matrix $\mathrm{Diag}(b - Ax^0)^{-1}$, we obtain the following polytope:

$$\mathcal{P} = \{v \in \Re^n : Cv \le e\}, \tag{5.1}$$

where $C \equiv \mathrm{Diag}(b - Ax^0)^{-1}A \in \Re^{m\times n}$ and e is the vector of all ones in $\Re^m$. We will use this form of polytopes in this section as it was used by Khachiyan and Todd in [11].

In the formulation (2.5), the matrix-valued variable E appears in the constraints in a nonlinear manner. In an alternative formulation given below, through the change of variables $B = E^2$ one can have the unknown matrix B appear linearly in the constraints. Indeed, after substituting E^2 by B and using the form (5.1), we can rewrite the problem (2.5) into

$$\begin{array}{ll} \min & -\log\det B \\ \text{s.t.} & c_i^T B c_i \le (1 - c_i^T x)^2, \quad i = 1, \ldots, m, \\ & (Cx < e, \;\; B \succ 0). \end{array} \tag{5.2}$$

While the constraints of (5.2) are linear with respect to the matrix variable B, they are no longer linear or convex with respect to the vector variable x.

5.1. Khachiyan and Todd's algorithm. Khachiyan and Todd's algorithm [11] for the MaxVE problem has a good complexity bound and also takes the advantage

of the special structure of the MaxVE problem. It is a suitable candidate for the purpose of performance comparison.

To make use of the simplicity of linear constraints, Khachiyan and Todd introduced the following subproblem, or auxiliary problem $AP(a)$, from (5.2):

$$\begin{array}{ll} \min & -\log\det B \\ \text{s.t.} & c_i^T B c_i \le (1 - c_i^T x)(1 - c_i^T a), \quad i = 1, \ldots, m, \\ & (B \succ 0) \end{array} \tag{5.3}$$

for a fixed $a \in \Re^n$, where $Ca < e$. Note that now the constraints are linear in both B and x. The key idea here is to solve subproblems $AP(a)$ iteratively until x and a become sufficiently close to each other so (5.3) becomes a good approximation of (5.2). Khachiyan and Todd use a primal barrier method to solve the subproblem $AP(a)$. Their barrier function has the form

$$F_t(x, B|\ a) = -\log\det B - t\sum_{i=1}^{m} \log\left((1 - c_i^T x)(1 - c_i^T a^k) - c_i^T B c_i\right),$$

where a is fixed and t is the barrier parameter. The Khachiyan and Todd (KT) algorithm can be summarized as follows.

ALGORITHM 2 (Khachiyan and Todd's algorithm).

Step 1. *Let a^0 be a strictly interior point of $\mathcal{P}$, $B^0 \succ 0$, $\epsilon > 0$, and $k = 0$.*

Step 2. *Solve the subproblem $AP(a^k)$ by using Newton's method to minimize the barrier function $F_t(x, B|\ a^k)$ for a sequence of $t \downarrow 0$. The solution of $AP(a^k)$ is (x^k, B^k).*

Step 3. *If $\|x^k - a^k\| \le \epsilon$, then stop; else let $a^{k+1} = (a^k + x^k)/2$, increment k, and go to Step* 2.

Khachiyan and Todd prove that to attain a sufficient accuracy only a small number of subproblems need to be solved, and they derive a linear system of size $n + m$ for calculating the Newton direction. Since the updates to the matrix-valued variable B are parameterized by a vector-valued variable, they are able to reduce the complexity of the algorithm. However, the drawback of their algorithm is that the barrier method used to solve the subproblem is not efficient in practice. Particularly, as we can see from the algorithmic framework, three layers of loops are involved in the KT algorithm: the loop for the subproblem parameter a, the loop for the barrier parameter t, and the iterations for a fixed a and a fixed t.

5.2. A modification of the KT algorithm. Since primal barrier methods are generally less efficient than primal-dual, interior-point methods, in order to speed up the KT algorithm we modify it by applying a primal-dual interior-point method to the subproblems in Step 2 of the KT algorithm, while keeping the outer iterations intact.

Following Khachiyan and Todd's approach, we transform the subproblem $AP(a)$ into the standard form $AP(0)$:

$$\begin{array}{ll} \min & -\log\det B \\ \text{s.t.} & c_i^T B c_i + c_i^T x \le 1, \quad i = 1, \ldots, m, \\ & (B \succ 0) \end{array} \tag{5.4}$$

by the change of variables $x \Rightarrow x + a$ and the change of data $c_i/(1 - c_i^T a) \Rightarrow c_i$ for $i = 1, \ldots, m$.

The optimality conditions, or KKT conditions, of problem $AP(0)$ are as follows:

$$C^T y = 0, \tag{5.5a}$$

$$B^{-1} - C^T Y C = 0, \tag{5.5b}$$

$$Cx + \mathrm{diag}(CBC^T) + z - e = 0, \tag{5.5c}$$

$$Yz = 0, \tag{5.5d}$$

$$y, z \geq 0, \tag{5.5e}$$

where $y \in \Re^m$ is the vector of Lagrangian multipliers, $z \in \Re^m$ consists of slack variables, and $C \in \Re^{m \times n}$ with c_i^T as its ith row.

Following the same strategy used earlier, we eliminate the matrix variable B from the system using the substitution $B(y) = (C^T Y C)^{-1}$ that is the solution to (5.5b). We also replace the zero right-hand side of (5.5d) by μe. The resulting system that defines the central path is

$$F_3(x, y, z) := \begin{pmatrix} C^T y \\ Cx + \mathrm{diag}(Q(y)) + z - e \\ Yz \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \mu e \end{pmatrix}, \tag{5.6}$$

where $y, z > 0$, and $Q(y) = C(C^T Y C)^{-1} C^T$. Clearly, (5.6) is a square, nonlinear system of $n + 2m$ variables. The Jacobian matrix of $F_3(x, y, z)$ is

$$F_3'(x, y, z) = \begin{bmatrix} 0 & C^T & 0 \\ C & -Q \circ Q & I \\ 0 & Z & Y \end{bmatrix}.$$

To solve the Newton linear system

$$F_3'(x, y, z) \begin{pmatrix} dx \\ dy \\ dz \end{pmatrix} = \begin{pmatrix} r_1 \\ r_2 \\ r_3 \end{pmatrix} := \begin{pmatrix} 0 \\ 0 \\ \mu e \end{pmatrix} - F_3(x, y, z),$$

we use the following block Gaussian elimination procedure:

$$\begin{aligned} dx &= (C^T M^{-1} C)^{-1}(r_1 + C^T M^{-1}(r_2 - Y^{-1} r_3)), \\ dy &= M^{-1}(Cdx - r_2 + Y^{-1} r_3), \\ dz &= Y^{-1}(r_3 - Z dy), \end{aligned}$$

where the matrix $M := Q \circ Q + Y^{-1} Z$ is symmetric positive definite.

The primal-dual algorithm for solving the subproblem $AP(0)$ falls into the same framework of Algorithm 1.

6. Numerical results. In this section, we report our numerical results on the four algorithms: the KT algorithm, the modified KT, or MKT, algorithm, and the two direct primal-dual interior-point algorithms based on the systems (3.6) and (3.8) which we name F1PD and F2PD, respectively. The numerical tests were performed on three sets of test problems with a total of 200 problems. Our implementations of the four algorithms are in Matlab. All the experiments were run on an SGI Origin2000 computer with multiple 300-MHz R12000 processors. However, our programs use only a single processor at a time.

6.1. Implementation details. In describing the implementation details, we first give some features common to all the algorithms and then other features specific to individual algorithms.

For all the algorithms, the input data for a polytope include the matrix A, the vector b, and a strictly interior point x^0 such that $Ax^0 < b$ which will serve as the starting point for the center of the initial ellipsoid. In our implementations, the point x_0 is selected to be the solution to an auxiliary linear program $\max\{\tau : Ax + \tau e \leq b\}$. Other choices are certainly possible such as the analytic center of the polytope. However, it was not our intention to use the best possible starting point.

Scaling is an important issue in numerical computation. In our implementations, we always first transform the inequality $Ax \leq b$ into the form $Cv \leq e$ using the change of variables and the row scaling as described at the beginning of section 5. After the transformation, the starting point x^0 is transformed into the origin, and the transformed polytope is better scaled.

In all the algorithms, the stopping tolerance is set to $\epsilon = 10^{-4}$. In the case of the KT and MKT algorithms, we stop the outer iterations whenever the relative change between the current and previous centers is less than or equal to ϵ. In the case of the F1PD and F2PD algorithms, we stop whenever the residual norm of F_i, $i = 1$ or 2, becomes less than or equal to ϵ.

We now describe some algorithm-specific features.

- The KT and MKT algorithms: Both algorithms have the same outer loop with the center varying. The initial center is the origin and the initial value for the matrix variable B is $B^0 = \rho I$, where I is the identity matrix and ρ is chosen such that the corresponding ball, centered at the origin with radius ρ, lies entirely inside the polytope. During the outer iterations, we use a warm-start strategy in which a later iteration always starts from the solution of the previous iteration.
- The KT algorithm: In the subproblems, the barrier parameter t is set to 0.5 initially and then decreased by a factor of 10 whenever the subproblem stopping criterion is met. For a fixed t value, the subproblem stopping criterion is that the gradient norm of the corresponding barrier function becomes less than or equal to t. This way, the stopping criterion becomes progressively more stringent as t approaches zero. We found that this adaptive strategy made the algorithm run significantly faster. To prevent the loss of symmetry during the computation, we set $B = (B + B^T)/2$ after B is updated at every iteration. We update an iterate for (x, B) by a damped Newton step to ensure that the updated ellipsoid remains inside the polytope. Specifically, the step length is 0.75 times the largest allowable step that keeps the updated ellipsoid inside the polytope.
- The primal-dual algorithms: The primal-dual algorithmic framework (i.e., Algorithm 1) encompasses the F1PD and F2PD algorithms, and the subproblem solver of the MKT algorithm. The initial values for the primal-dual algorithms are set as follows: the initial center is $x = 0$; the initial multiplier value is $y = e$; and the initial slack variable z, say, in the equation $z - g = 0$, is set as $z_i = \max(0.1, g_i)$. In addition to the initial values, there are two critical parameters in these algorithms: the so-called centering parameter σ^k and the step length α^k. In our implementations, we choose $\sigma^k = \min\{0.5, g(y^k)^T z^k/m\}$ for F1PD or $\sigma^k = \min\{0.5, (y^k)^T z^k/m\}$ for F2PD, and $\alpha^k = \min(1, \tau\hat{\alpha})$, where $\tau \in (0, 1)$ and $\hat{\alpha}$ is the maximum

length such that updated iterate for (x, y, z) reaches the boundary of the set $\mathcal{P} \times \Re^m_{++} \times \Re^m_{++}$. We use $\tau = 0.75$ for the F1PD and F2PD algorithms, and a more aggressive value $\tau = 0.9$ for the subproblem solver of the MKT algorithm because the subproblem (5.3) is not as nonlinear as its counterparts are in the F1PD and F2PD algorithms.

The parameter settings given above are rather generic and unsophisticated. For example, a line search scheme for determining step length could be a more effective and theoretically sound strategy. However, given our purpose of identifying the most robust and efficient algorithm, we consider our current settings to be appropriate and sufficient.

6.2. Test problems. Three sets of test problems were used in our numerical experiments, consisting of 47, 143, and 10 problems, respectively. The total number of test problems is 200. (All the test problems, as well as detailed problem information and numerical results, are available from http://www.caam.rice.edu/~zhang/mve.)

Test sets 1 and 2 are obtained from an implementation of Lenstra's algorithm for an integer programming feasibility problem [12, 13]. This algorithm searches on a tree of subproblems and applies ellipsoidal approximation on each one of them. The polytopes in sets 1 and 2 are taken from some branches of the search trees for two different integer programming feasibility problems. The problem sizes in sets 1 and 2 are relatively small with $m \leq 288$ and $n \leq 80$. Nevertheless, our numerical experience has indicated that some of the problems are nontrivial to solve.

In order to test the ability of our algorithms for solving larger problems, we generated a set of 10 random problems that is called set 3. The largest problem in this set has $m = 1200$ and $n = 500$. For each problem, we first use the Matlab function `sprandn` to generate a sparse random matrix B and then use the `rand` function to generate a right-hand side vector $c > 0$, an upper-bound vector $ub > 0$, and a lower-bound vector $lb < 0$. Together, they form a polytope

$$\{x \in \Re^n : Bx \leq c, \;\; lb \leq x \leq ub\},$$

where $B \in \Re^{k \times n}$ and $c \in \Re^k$ and $lb, ub \in \Re^n$. By construction, the origin $x = 0$ is strictly interior to the polytope. Then we rewrite the polytope into the standard form

$$\{x \in \Re^n : Ax \leq b\},$$

where $A \in \Re^{m \times n}, b \in \Re^m$, with $m = k + 2n$. The matrix A is constructed, in an obvious manner, from the matrix B and the identity matrix in $\Re^n$, and the vector b is constructed from the vectors $c \in \Re^k$ and $lb, ub \in \Re^n$. The problems in set 3 are sparse.

6.3. Test results. Test results on sets 1 and 2, totaling 190 problems are summarized in Table 6.1. Six rows of numbers are presented in Table 6.1. For each test set, in the first row we list the test set number, the number of test problems in the set, the total number of iterations, and the total amount of CPU time in seconds taken by each algorithm for solving the entire set of test problems; then in the last two rows for each category we give the algebraic mean and the standard deviation (std) of the set.

We note that the iteration numbers for the KT and the MKT algorithms are the numbers of innermost, Newton iterations that involve solving systems of linear equations. These innermost iterations are comparable to the iterations of the primal-dual algorithms in terms of complexity of linear algebra computation. Specifically, all of the iterations require either solving $m \times m$ linear systems or inverting $m \times m$

TABLE 6.1
Summary of results on tests 1 *and* 2.

Test set	No. of probs	KT		MKT		F1PD		F2PD	
		iter	time	iter	time	iter	time	iter	time
1	47	19416	3340	2655	240	692	124	694	77
mean		413.1	71.1	56.5	5.1	14.7	2.6	14.8	1.6
std		17.7	38.6	5.8	2.9	1.9	1.8	1.6	1.0
2	143	56783	3567	9720	429	2448	168	2058	104
mean		397.1	24.9	68.0	3.0	18.1	1.2	14.4	0.7
std		54.3	6.0	10.1	0.8	3.6	0.4	2.3	0.2

TABLE 6.2
Results on test set 3: *Problems* 1–10.

Prob number	Size			F1PD		F2PD	
	m	n	nnz	iter	time	iter	time
1	600	100	7426	31	97	22	30
2	600	150	8408	30	107	23	39
3	600	200	7669	53	203	29	58
4	600	250	5022	60	249	31	73
5	800	100	5914	34	235	22	63
6	800	200	8029	34	271	24	91
7	800	300	8933	58	549	32	165
8	1000	300	11993	40	675	28	245
9	1000	400	8433	60	1134	31	330
10	1200	500	10518	73	2917	37	703
mean	—	—	—	47.3	643.3	27.9	179.6
std	—	—	—	15.3	860.5	5.0	211.9

matrices, and hence have an $\mathcal{O}(m^3)$ complexity per iteration. Nevertheless, these algorithms do differ in terms of secondary computational tasks. For example, both KT and MKT algorithms compute matrices of the form $A^T M^{-1} A$, while F1PD and F2PD algorithms compute $A^T N M^{-1} A$, where A is $m \times n$ $(m > n)$ and M and N are $m \times m$. For both cases the leading complexity term is $\mathcal{O}(m^3)$, but the latter is more expensive than the former. Similarly, comparing (3.18) and (3.21c) with (3.23) and (3.22c) we can see that F1PD requires more linear algebra computation than F2PD.

From Table 6.1, we observe that on average the MKT algorithm is about 10 times faster than the KT algorithm, the F1PD algorithm is over 2 times faster than the MKT algorithm, and the F2PD algorithm is about 1.5 times faster than the F1PD algorithm. Moreover, the standard deviations in both iteration count and CPU time decrease monotonically in the same order: KT, MKT, F1PD, and F2PD. The results are remarkably consistent; for example, there is not a single problem which F1PD solved in less time than F2PD did.

We mention that out of the 190 test problems in test sets 1 and 2 the KT algorithm failed to converge on two: problems 22 and 120 in set 2. More conservative choices of parameters would make the KT algorithm converge on these two problems but would also adversely affect the overall performance of the algorithm. We kept the current choices of parameters for the benefit of the KT algorithm.

The test results on the randomly generated test set 3 are presented in Table 6.2. Only the F1PD and F2PD algorithms were tested on this set of larger problems because the other two algorithms would require an excessively long time to run. Since these test problems are sparse, in addition to the matrix sizes m and n, we also include the number of nonzero entries, denoted as nnz, in the matrix A. We mention that although the sparsity in A makes relevant matrix multiplications cheaper, the matrix

$Q = A(A^T Y A)^{-1} A^T$ involved in $h'(y)$ (see (3.14)) is still generally dense. As a result, it is still necessary to solve $m \times m$ dense linear systems in the algorithms.

The results in Table 6.2 indicate that given the current choices of parameters, the F2PD algorithm clearly outperforms the F1PD algorithm by a considerable margin on test set 3. Although the performance of the F1PD algorithm may be somewhat improved by selecting different parameters, we do not believe that it can in general outperform the F2PD algorithm because it requires more linear algebra calculation in each iteration for solving its version of the Newton linear system.

7. Concluding remarks. The goal of this study is to find a practically efficient algorithmic framework for solving general MaxVE problems where the number of constraints m is a small multiple of the number of variables n. Our extensive numerical results show that among the four tested algorithms, the method of choice is clearly the F2PD algorithm built on the formulation (3.8), which has been shown to have a sound theoretical foundation. We have established, among other things, the existence of a central path for this formulation even though this central path is not known to be directly connected to the optimality conditions of a barrier function.

The main advantage of the F2PD algorithm over the KT and the MKT algorithms is that, without the need for solving a number of subproblems either for fixed centers or fixed barrier parameter values, it requires fewer iterations (or linear system solutions) than the other two algorithms. We expect that the same advantage would still hold against some other untested algorithms like the one given in [1]. In addition, compared to the F1PD algorithm, the F2PD algorithm requires less linear algebra computation per iteration and seems to be more robust. These features make the F2PD algorithm particularly attractive.

We should point out that the polynomial algorithm recently proposed by Nemirovskii [16] is, much like our algorithms, a primal-dual type algorithm free of matrix variables. Such a characteristic indicates that it may also be promising as a practically efficient algorithm. This algorithm deserves further study from a computational point of view.

The algorithms considered in this paper are all of the general-purpose type. For really large-scale problems with special structures, one will likely need special-purpose algorithms that can take full advantage of the problem-specific structures, in particular sparsity, in order to solve the problems efficiently. This should be a topic of further research.

Acknowledgment. We would like to thank two anonymous referees for their thoughtful and helpful comments and suggestions which have enabled us to improve the paper significantly. Our thanks also go to Michael Todd for suggesting that we consider the modification of the Khachiyan–Todd algorithm.

REFERENCES

[1] K. M. Anstreicher, *Improved complexity for maximum volume inscribed ellipsoids*, SIAM J. Optim., 13 (2002), pp. 309–320.

[2] M. Dyer, A. Frieze, and R. Kannan, *A random polynomial-time algorithm for estimating volumes of convex bodies*, J. Assoc. Comput. Mach., 38 (1991), pp. 1–17.

[3] A. El-Bakry, R. A. Tapia, T. Tsuchiya, and Y. Zhang, *On the formulation of the primal-dual Newton interior-point method for nonlinear programming*, J. Optim. Theory Appl., 89 (1996), pp. 507–541.

[4] M. Grötschel, L. Lovász, and A. Schrijver, *Geometric Algorithms and Combinatorial Optimization*, Springer, Berlin, 1988.

[5] R. A. Horn and C. R. Johnson, *Matrix Analysis*, Cambridge University Press, Cambridge, UK, 1985.
[6] R. A. Horn and C. R. Johnson, *Topics in Matrix Analysis*, Cambridge University Press, Cambridge, UK, 1991.
[7] F. John, *Extreme problems with inequalities as subsidiary conditions*, in Studies and Essays Presented to R. Courant on His 60th Birthday, Interscience, New York, 1948, pp.187–204.
[8] R. Kannan, L. Lovász, and M. Simonovits, *Random walks and an $O^*(n^5)$ volume algorithm for convex bodies*, Random Structures Algorithms, 11 (1997), pp. 1–50.
[9] L. Khachiyan, *A polynomial algorithm in linear programming*, Dokl. Akad. Nauk SSSR, 244 (1979), pp. 1093–1096.
[10] L. Khachiyan, *Rounding of polytopes in the real number model of computation*, Math. Oper. Res., 21 (1996), pp. 307–320.
[11] L. Khachiyan and M. Todd, *On the complexity of approximating the maximal inscribed ellipsoid for a polytope*, Mathematical Programming, 61 (1993), pp. 137–159.
[12] H. W. Lenstra, Jr., *Integer programming with a fixed number of variables*, Math. Oper. Res., 8 (1983), pp. 538–548.
[13] L. Lovász, *An Algorithmic Theory of Numbers, Graphs, and Convexity*, CBMS-NSF Regional Conf. Ser. in Appl. Math. 50, SIAM, Philadelphia, 1986
[14] L. Lovász and M. Simonovits, *On the randomized complexity of volumes and diameters*, in Proceedings of the 33rd Annual Symposium on Foundations of Computer Science, Pittsburgh, PA, 1992, pp. 482–491.
[15] R. D. C. Monteiro and J. S. Pang, *Properties of an interior-point mapping for nonlinear mixed complementarity problems*, Math. Oper. Res., 21 (1996), pp. 629–654.
[16] A. Nemirovskii, *On self-concordant convex-concave functions*, Optim. Methods Softw., 11/12 (1999), pp. 303–384.
[17] Y. Nesterov and A. Nemirovskii, *Interior-Point Polynomial Algorithms in Convex Programming*, SIAM Stud. Appl. Math. 13, SIAM, Philadelphia, 1994.
[18] Yu. E. Nesterov and M. J. Todd, *Primal-dual interior-point methods for self-scaled cones*, SIAM J. Optim., 8 (1998), pp. 324–364.
[19] A. Schrijver, *Theory of Linear and Integer Programming*, John Wiley and Sons, Chichester, UK, 1986.
[20] S. Silvey and D. Titterington, *A geometric approach to optimal design theory*, Biometrika, 60 (1973), pp. 21–32.
[21] S. Tarasov, L. Khachiyan, and I. Erlich, *The method of inscribed ellipsoid*, Soviet Math. Dokl., 37 (1988), pp. 226–230.
[22] D. Titteringto, *Optimal design: Some geometric aspects of d-optimality*, Biometrika, 62 (1975), pp. 313–320.
[23] L. Vandenberghe, S. Boyd, and S.-P. Wu, *Determinant maximization with linear matrix inequality constraints*, SIAM J. Matrix Anal. Appl., 19 (1998), pp. 499-533.
[24] E. Welzl, *Smallest enclosing disks, balls and ellipsoids*, in New Results and New Trends in Computer Sciences, Lecture Notes in Comput. Sci. 555, H. Maurer, ed., Springer, Berlin, 1991, pp. 359–370.
[25] Y. Ye, *A new complexity result on minimization of a quadratic function with a sphere constraint*, in Recent Advances in Global Optimization, C. Floudas and P. Pardalos, eds., Princeton University Press, Princeton, NJ, 1992, pp. 19–31.
[26] Y. Zhang, *An Interior-Point Algorithm for the Maximum-Volume Ellipsoid Problem*, Technical report TR98-15, CAAM Department, Rice University, Houston, TX, 1998 (revised 1999).

SIAM J. OPTIM.
Vol. 14, No. 1, pp. 77–114

LARGE-SCALE MOLECULAR OPTIMIZATION FROM DISTANCE MATRICES BY A D.C. OPTIMIZATION APPROACH*

LE THI HOAI AN† AND PHAM DINH TAO†

Abstract. A so-called DCA method based on a d.c. (difference of convex functions) optimization approach (algorithm) for solving large-scale distance geometry problems is developed. Different formulations of equivalent d.c. programs in the l_1-approach are stated via the Lagrangian duality without gap relative to d.c. programming, and new nonstandard nonsmooth reformulations in the l_∞-approach (resp., the $l_1 - l_\infty$-approach) are introduced. Substantial subdifferential calculations permit us to compute sequences of iterations in the DCA quite simply. The computations actually require matrix-vector products and only one Cholesky factorization (resp., with an additional solution of a convex program) in the l_1-approach (resp., the $l_1 - l_\infty$-approach) and allow the exploitation of sparsity in the large-scale setting. Two techniques—respectively, using shortest paths between all pairs of atoms to generate the complete dissimilarity matrix and the spanning trees procedure—are investigated in order to compute a good starting point for the DCA. Finally, many numerical simulations of the molecular optimization problems with up to 12567 variables are reported, which prove the practical usefulness of the nonstandard nonsmooth reformulations, the globality of found solutions, and the robustness and efficiency of our algorithms.

Key words. reformulations, d.c. programming, d.c. algorithm (DCA), Lagrangian duality, subdifferential calculations, distance geometry problem, dissimilarity geometry problem

AMS subject classifications. 05C10, 49M27, 51K99, 65K05, 65K10, 90C30, 90C35

DOI. 10.1137/S1052623498342794

1. Introduction.

In recent years there has been much active research in molecular optimization, especially in the protein-folding framework, which is one of the most important problems in biophysical chemistry. Molecular optimization problems arise also in the study of clusters (molecular cluster problems) and of large confined ionic systems in plasma physics [23]. The determination of a molecular conformation can be tackled by either minimizing a potential energy function (if the molecular structure corresponds to the global minimizer of this function) or solving the distance geometry problem [5], [11] (when the molecular conformation is determined by distances between pairs of atoms in the molecule). Both methods are concerned with global optimization problems.

In this paper we are interested in the large-scale molecular conformation via the distance geometry problem, which can be formulated as follows: find positions $x^1, \ldots, x^n$ of n atoms in $\mathbb{R}^3$ such that

$$\|x^i - x^j\| = \delta_{ij} \qquad \text{for } (i,j) \in \mathcal{S}, \tag{1.1}$$

where $\mathcal{S}$ is a subset of the atom pairs, δ_{ij} with $(i,j) \in \mathcal{S}$ is the given distance between atoms i and j, and $\|\cdot\|$ denotes the Euclidean norm. Usually, a small subset of pairwise distances is known; i.e., $\mathcal{S}$ is sparse.

*Received by the editors July 22, 1998; accepted for publication (in revised form) December 26, 2002; published electronically July 18, 2003. This work was partially supported by the computer resources financed by "Contrat de Plan Interrégional du Bassin Parisien—Pôle interrégional de modélisation en Sciences pour Ingénieurs."

http://www.siam.org/journals/siopt/14-1/34279.html

†Laboratory of Modelling, Optimization & Operations Research, National Institute for Applied Sciences-Rouen, BP 8, F-76 131 Mont Saint Aignan Cedex, France (lethi@insa-rouen.fr, pham@insa-rouen.fr)

The above formulation corresponds to the *exact* distance geometry problem. By the error in the theoretical or experimental data, there may not exist any solution to this problem, for example, when the triangle inequality

$$\delta_{ij} \leq \delta_{ik} + \delta_{kj}$$

is violated for atoms i, j, k. Then an ε-optimal solution of (1.1), namely, a configuration $x^1, \ldots, x^n$ satisfying

$$|\|x^i - x^j\| - \delta_{ij}| \leq \varepsilon \qquad \text{for } (i,j) \in \mathcal{S}, \tag{1.2}$$

is useful in practice.

The *general* distance geometry problem then is to find positions $x^1, \ldots, x^n$ in $\mathbb{R}^3$ verifying

$$l_{ij} \leq \|x^i - x^j\| \leq u_{ij} \qquad \text{for } (i,j) \in \mathcal{S}, \tag{1.3}$$

where l_{ij} and u_{ij} are lower and upper bounds of the distance constraints, respectively.

In what follows, $\mathcal{M}_{n,p}(\mathbb{R})$ denotes the space of real matrices of order $n \times p$, and for $X \in \mathcal{M}_{n,p}(\mathbb{R})$, X_i (resp., X^i) is its ith row (resp., ith column). By identifying a set of positions $x^1, \ldots, x^n$ with the matrix X (i.e., $(X^T)^i = (X_i)^T = x^i$ for $i = 1, \ldots, n$), we can advantageously express the exact and/or general distance geometry problems in the matrix space $\mathcal{M}_{n,p}(\mathbb{R})$:

$$\text{(EDP)} \qquad 0 = \min \left\{ \sigma(X) := \frac{1}{2} \sum_{(i,j) \in S, i<j} w_{ij}\theta_{ij}(X) : \quad X \in \mathcal{M}_{n,p}(\mathbb{R}) \right\},$$

where $w_{ij} > 0$ for $i \neq j$ and $w_{ii} = 0$ for all i. The pairwise potential $\theta_{ij} : \mathcal{M}_{n,p}(\mathbb{R}) \longrightarrow \mathbb{R}$ is defined for problem (1.1) by either

$$\theta_{ij}(X) = \left(\delta_{i,j}^2 - \|X_i^T - X_j^T\|^2\right)^2 \tag{1.4}$$

or

$$\theta_{ij}(X) = \left(\delta_{ij} - \|X_i^T - X_j^T\|\right)^2, \tag{1.5}$$

and for problem (1.3) by

$$\theta_{ij}(X) = \min^2 \left\{ \frac{\|X_i^T - X_j^T\|^2 - l_{ij}^2}{l_{ij}^2}, 0 \right\} + \max^2 \left\{ \frac{\|X_i^T - X_j^T\|^2 - u_{ij}^2}{u_{ij}^2}, 0 \right\}. \tag{1.6}$$

Note that for simplicity $(X_j)^T$ is written throughout the paper as X_j^T. In the molecular optimization problem, p is equal to 3. It is easy to see that, except for θ_{ij} defined by (1.4), where the objective function is infinitely differentiable, problems (EDP), with θ_{ij} given by (1.5) and (1.6), are nondifferentiable optimization problems. However, they are all d.c. programs.

Observe that X is a solution of the distance geometry problem if and only if it is a global minimizer of problem (EDP) and $\sigma(X) = 0$.

When all pairwise distances are available and a solution exists, the exact distance geometry problem (1.1) can be solved by a polynomial time algorithm (Blumenthal [3], Crippen and Havel [5]). However, in practice one knows only a subset of the distances, and it is well known (Saxe [37]) that p-dimensional distance geometry problems are strongly NP-complete with $p = 1$ and strongly NP-hard for all $p > 1$. The visible

sources of difficulties of these problems are

- the question of the existence of a solution,
- the nonuniqueness of solutions,
- the presence of a large number of local minimizers,
- the large scale of problems that arise in practice.

Several methods have been proposed for solving the distance geometry problems (1.1) and/or (1.3). De Leeuw [7], [8] proposed the well-known majorization method for solving the Euclidean metric multidimensional scaling (MDS) problem, which includes (EDP) with θ_{ij} given by (1.5). Crippen and Havel [5] used the function θ_{ij} defined in (1.6) for solving (1.3) by the EMBED algorithm. Their method consists of solving a sequence of exact distance geometry problems where all pairwise distances are included. It relies on the SVD or alternative Cholesky decomposition with diagonal pivoting. Current implementations of the EMBED algorithm use a local minimizer of problem (EDP), (1.4) as a starting point for a simulated annealing. Glunt, Hayden, and Raydan [9] studied a special gradient method for determining a local minimizer of (1.1), with θ_{ij} defined in (1.5). From a graph-theoretic viewpoint, Hendrickson [13] developed an algorithm to solve (1.1), where θ_{ij} is given by (1.4). His method works well for his test problems in which a protein contains at most 124 amino acids (at most 777 atoms). The protein actually has 1849 atoms, but some simple structure exploitation allowed the author to start the numerical method with only 777 atoms. With a smoothing technique and a continuation approach based on the Gaussian transform of the objective function and on the trust region method, Moré and Wu [22] proposed an algorithm for solving (1.1), with θ_{ij} defined by (1.4). By the Gaussian transform, the original function becomes a smoother function with fewer local minimizers. Computational experiments with up to 648 variables ($n = 216$) in [22] proved that the continuation method is more reliable and efficient than the multistart approach, a standard procedure for finding the global minimizer to (EDP). Also by the Gaussian transform, Moré and Wu [24] considered the general distance geometry problem with the function θ_{ij} defined by (1.6). A stochastic/perturbation algorithm was proposed by Zou, Bird, and Schnabel [41] for both general and exact distance geometry problems. This is a combination of a stochastic phase that identifies an initial set of local minimizers and a more deterministic phase that moves from a low to an even lower local minimizer. The numerical experiments presented there (with the same data as in Moré and Wu [22] and Hendrickson [13]) showed that this approach is promising. It is worth noting that (EDP) is intimately related to the Euclidean distance matrix completion problem [1], [15]. This problem has been formulated as a semidefinite programming problem and considered by Alfakih, Khandami, and H. Wolkowicz [1] with an adapted interior-point method.

In the convex analysis approach to nondifferentiable nonconvex programming, the d.c. (difference of convex functions) optimization and its solution algorithms (DCA) developed by Pham Dinh and Le Thi (see [16], [17], [30], [31], [32], [33], and references therein) constitute a natural and logical extension of Pham Dinh's earlier works concerning convex maximization and its subgradient algorithms (see [26], [27], [28], [29], and references therein). The majorization algorithm is a suitable adaptation of the above subgradient methods for maximizing a seminorm over the unit ball of another one. Our method in this paper, based on the d.c. optimization approach, aims at solving the exact geometry distance problem (1.1) with different standard and nonstandard formulations as nonsmooth d.c. programs. More precisely, we are concerned with (EDP), (1.5) and problems (1.7), (1.8), and (1.9) given below.

The purpose of this paper is to demonstrate that the DCA can be suitably adapted for devising efficient algorithms for solving large-scale exact distance geometry problems. We propose various versions of DCA that are based on different formulations for this problem. The DCA is a primal-dual subgradient method for solving a general d.c. program that consists of the minimization of the d.c. on the whole space. (The convex constraint set is incorporated into the objective function by using its indicator function.) Featured as a descent method without line-search, it is at present one of a few algorithms in the local approach which has been successfully applied to many large-scale d.c. optimization problems and proved to be more robust and efficient than related standard methods. Due to its local character it cannot guarantee the globality of computed solutions for general d.c. programs. However, we observe that with a suitable starting point it converges quite often to a global solution (see, e.g., [16], [17], [18], [32]). This property motivates us to investigate a technique for computing a "good" starting point for the DCA in the solution of (EDP), with θ_{ij} defined by (1.5). The idea of this technique arose from two facts:

- When all pairwise distances are known, the DCA applied to (EDP), (1.5) is very simple. Although the DCA is not a polynomial time algorithm, it works very well in practice, because it has an explicit form and requires only matrix-vector products.
- In the general case where only a small subset of distances is known, one can approximate a solution of (EDP), (1.5) by using a dense set of constraints, which is extrapolated from the given distances and then works with this set.

The so-called EDCA, a variant of the DCA, is composed of the following two phases. In Phase 1 we *complete* the matrix of distances by using the *shortest path* between all pairs of atoms and then apply the DCA to the new problem where all pairwise "distances" (rather, dissimilarities) are known. In Phase 2 we solve the original problem by applying the DCA from the point obtained in Phase 1.

This two-phase algorithm EDCA has some advantages. First, we work with both (*dense* and *sparse*) sets of constraints. The use of a complete matrix which is an *approximate* distance matrix (called a *dissimilarity* matrix) aims at finding a good initial point for the DCA applied to the original problem: such a starting point is computed by DCA applied to the resulting problem (3.14) with the complete dissimilarity matrix. By contrast, the existing methods work only on either a full set of constraints (see, e.g., [5]) or a sparse set of constraints [24], [41].

As an alternative to Phase 1, we also propose Procedure SP (section 4), which is an adaptation of the *inexpensive approach using spanning trees* (Algorithm Struct of Moré and Wu [24]) to compute acceptable starting points. In our experiments Procedure SP is more (resp., less) efficient than Phase 1 when the number of distance constraints is small (resp., large).

Another important issue in our d.c. optimization approach is that we can exploit the *nice* effect of d.c. decompositions of problem (EDP), (1.5). In fact, by using the *Lagrangian duality with zero gap* relative to the problem of maximization of a gauge ψ on the unit ball defined by a gauge ϕ (see [16], [31]), we are able to obtain different d.c. formulations of Problem (1.1), and then different d.c. optimization algorithms are introduced in Phases 1 and 2. In particular, we are interested in the following convex maximization problem:

$$\max\left\{\xi(X) := \sum_{(i,j)\in\mathcal{S}} w_{ij}\|X_i^T - X_j^T\| : X \in C\right\}, \tag{1.7}$$

with

$$C := \left\{ X \in \mathcal{M}_{n,p}(\mathbb{R}) : \sum_{(i,j)\in\mathcal{S}, i<j} w_{ij}\delta_{ij}\|X_i^T - X_j^T\|^2 \leq \sum_{(i,j)\in\mathcal{S}, i<j} w_{ij}\delta_{ij}^2 \right\}.$$

It appears that the norms l_1 and l_2 have served to model these nonconvex programs. In addition, we have introduced the nonstandard $l_1 - l_\infty$-approach to reformulating the exact distance geometry problem (section 5): we then also take into account in the objective functions the function

$$\Phi(X) := \max\left\{ \Phi_{ij}(X) := \frac{1}{2} w_{ij}[\|X_i^T - X_j^T\| - \delta_{ij}]^2 : (i,j) \in \mathcal{S}, i < j \right\}.$$

That leads to the two nonstandard d.c. programs using the l_∞-norm and the combination of the l_1 and l_∞ norms in their formulations

(1.8)
$$0 = \min\left\{ \Phi(X) := \max_{(i,j)\in\mathcal{S}, i<j} \left\{ \Phi_{ij}(X) := \frac{1}{2} w_{ij}[\|X_i^T - X_j^T\| - \delta_{ij}]^2 : X \in \mathcal{M}_{n,p}(\mathbb{R}) \right\} \right\},$$

$$0 = \min\left\{ \Phi(X) + \frac{\rho}{2} \sum_{(i,j)\in\mathcal{S}, i<j} w_{ij}\left(\delta_{ij} - \|X_i^T - X_j^T\|\right)^2 : X \in \mathcal{M}_{n,p}(\mathbb{R}) \right\}. \tag{1.9}$$

Although these d.c. programs have the same solution set, problems (1.7), (1.8), and (1.9) *cannot be transformed into equivalent smooth nonconvex programs.* Consequently the DCA seems to play here *its crucial role in nonsmooth d.c. programming.* Recall, as mentioned before, that most alternative methods for solving the distance geometry problem were applied to the (infinitely differentiable) d.c. program (EDP), (1.4) because they require some smoothness. But paradoxically problems (EDP), (1.5) and (1.7) are quite suitable to the application of the DCA: the choice of a d.c. program equivalent to problem (1.1) is crucial because DCA is far better when applied to (EDP), (1.5) than to the standard (EDP), (1.4).

As will be seen in section 5, the DCA for problems (1.8), (1.9) are not explicit as they are for Problems (EDP), (1.5) and (1.7). They are consequently more expensive. However, for certain classes of *hard* exact distance geometry problems, problem (1.9) seems to be more suitable to making DCA converge to global solutions (section 6).

Our algorithms are quite simple and easy to implement. They require only matrix-vector products and one Cholesky factorization for DCA1 (DCA applied to (EDP), (1.5)) and DCA2 (DCA applied to (1.7)). For DCA3 (DCA applied to (1.9)) we have to solve nonsmooth convex programs. We have tested our codes on the artificial distance geometry problems (Moré and Wu [22]), on the data derived from the Protein Data Bank (PDB) [2] with up to 12567 variables (the molecule contains 4189 atoms), and on the twelve test problems constructed by Hendrickson [12], [13]. These last are among the most difficult test problems for the exact distance geometry problems.

Our work relies on d.c. programming and its main tool, the DCA. A short description of the background indispensable for understanding this approach is given in section 2. In section 3 we present the use of the general scheme of the DCA to solve two equivalent d.c. programs (problems (3.6) and (3.12)) of problem (1.1). A

thorough study of these problems and their proximal regularization in their elegant matrix formulation and especially the *substantial subdifferential calculus for related convex functions* in the resulting d.c. programs allows us to express DCA1, DCA2, and their proximal regularized versions DCA1r and DCA2r in an *explicit form* and to *exploit the sparsity*. The two-phase algorithm EDCA is summarized in section 4. The *nonstandard reformulations* in the l_∞-approach and the l_1-l_∞-approach, respectively, are presented, together with their solution, algorithm DCA3, in section 5. Numerical simulations reported in section 6 demonstrate the practical usefulness of the nonstandard reformulations, the globality of the sought solutions, and the efficiency and reliability of our algorithms.

2. D.c. programming and the DCA. In this section we summarize the material needed for an easy understanding of d.c. programming and the DCA, which will be used to solve the exact distance geometry problem (EDP), with θ_{ij} given in (1.5). We are working with the space $E = \mathbb{R}^n$, which is equipped with the canonical inner product $\langle \cdot, \cdot \rangle$ and the corresponding Euclidean norm $\| \cdot \|$; thus the dual space E^* of E can be identified with E itself. We follow [35] for definitions of the usual tools of convex analysis, where functions could take the infinite values $\pm\infty$. A function $\theta : E \to \mathbb{R} \cup \{\pm\infty\}$ is said to be proper if it takes the value $-\infty$ nowhere and is not identically equal to $+\infty$. The effective domain of θ, denoted by dom θ, is

$$\text{dom } \theta = \{x \in E : \theta(x) < \infty\}.$$

The set of all lower semicontinuous proper convex functions on E is denoted by $\Gamma_0(E)$. For $g \in \Gamma_0(E)$, the *conjugate function* g^* of g is a function belonging to $\Gamma_0(E^*)$ and defined by

$$g^*(y) = \sup\{\langle x, y\rangle - g(x) : \ x \in E\}.$$

Note that $g^{**} = g$.

Let $g \in \Gamma_0(E)$, and let $x^0 \in \text{dom } g$ and $\epsilon > 0$. Then $\partial_\epsilon g(x^0)$ stands for the $\epsilon-$ *subdifferential* of g at x^0 and is given by

$$\partial_\epsilon g(x^0) = \{y^0 \in E^* : g(x) \geq g(x^0) + \langle x - x^0, y^0\rangle - \epsilon \quad \forall x \in E\},$$

while $\partial g(x^0)$ corresponding to $\epsilon = 0$ stands for the *usual (or exact) subdifferential* of g at x^0. Recall that

$$y^0 \in \partial g(x^0) \Longleftrightarrow x^0 \in \partial g^*(y^0) \Longleftrightarrow \langle x^0, y^0\rangle = g(x^0) + g^*(y^0).$$

One says that g is *subdifferentiable* at x^0 if $\partial g(x^0)$ is nonempty. It has been proved [35] that ri(dom g) $\subset$ dom $\partial g \subset$ dom g, where ri(dom g) stands for the relative interior of dom g and dom $\partial g := \{x \in E : \partial g(x) \neq \emptyset\}$.

Furthermore, the indicator function χ_C of a closed convex set C is defined by $\chi_C(x) = 0$ if $x \in C$, and $+\infty$ otherwise.

Let $\rho \geq 0$, and let C be a convex subset of E. One says that a function $\theta : C \longrightarrow \mathbb{R} \cup \{+\infty\}$ is ρ-*convex* if

$$\theta[\lambda x + (1-\lambda)x'] \leq \lambda\theta(x) + (1-\lambda)\theta(x') - \frac{\lambda(1-\lambda)}{2}\rho\|x - x'\|^2$$
$$\forall \lambda \in]0,1[, \ \forall x, x' \in C.$$

This amounts to saying that $\theta - (\rho/2)\|\cdot\|^2$ is convex on C. The *modulus of strong convexity* of θ on C, denoted by $\rho(\theta, C)$ or $\rho(\theta)$ if $C = E$, is given by

$$\rho(\theta, C) = \sup\{\rho \geq 0 : \theta - (\rho/2)\|\cdot\|^2 \text{ is convex on } C\}. \tag{2.1}$$

Clearly, θ is convex on C if and only if $\rho(\theta, C) \geq 0$. One says that θ is *strongly convex* on C if $\rho(\theta, C) > 0$.

For f_1 and f_2 belonging to $\Gamma_0(E)$, the *infimal convolution* of f_1 and f_2, denoted $f_1 \nabla f_2$, is a convex function on E, defined by (see [14])

$$f_1 \nabla f_2(x) = \inf\{f_1(x_1) + f_2(x_2) : x_1 + x_2 = x\} \quad \forall x \in E.$$

In convex analysis, this functional operation aims, as does the convolution in functional analysis, at regularizing convex functions [14]. The *proximal regularization* corresponds to $\theta = \frac{\lambda}{2}\|\cdot\|^2$.

For $f \in \Gamma_0(E)$ and $\lambda > 0$ the *Moreau–Yosida regularization* of f with parameter λ, denoted by f_λ, is the infimal convolution of f and $\frac{1}{2\lambda}\|\cdot\|^2$. The function f_λ is continuously differentiable, underapproximates f without changing the set of minimizers, and $(f_\lambda)_\mu = f_{\lambda+\mu}$. More precisely, $\nabla f_\lambda = \frac{1}{\lambda}[I - (I + \lambda\partial f)^{-1}]$ is Lipschitzian with ratio $\frac{1}{\lambda}$. The operator $(I + \lambda\partial f)^{-1}$ is called the *proximal mapping associated with* λf (see [20], [21], [36]).

For $g, h \in \Gamma_0(E)$, a general *d.c. program* is that of the form

$$\alpha = \inf\{f(x) := g(x) - h(x) : x \in E\}, \tag{P_{dc}}$$

where we adopt the convention $+\infty - (+\infty) = +\infty$ to avoid ambiguity. One says that $g - h$ is a *d.c. decomposition* (or d.c. representation) of f, and g, h are its *convex d.c. components*. If g and h are finite on E, then $f = g - h$ is said to be a finite d.c. function on E. The set of d.c. functions (resp., finite d.c. functions) on E is denoted by $\mathcal{DC}(E)$ (resp., $\mathcal{DC}_f(E)$).

Note that the finiteness of α merely implies that

$$\text{dom } g \subset \text{dom } h \quad \text{and} \quad \text{dom } h^* \subset \text{dom } g^*. \tag{2.2}$$

Such inclusions will be assumed throughout the paper.

A point x^* is said to be *a local minimizer* of $g - h$ if $g(x^*) - h(x^*)$ is finite (i.e., $x^* \in \text{dom } g \cap \text{dom } h$) and there exists a neighborhood U of x^* such that

$$g(x^*) - h(x^*) \leq g(x) - h(x) \quad \forall x \in U. \tag{2.3}$$

Under the convention $+\infty - (+\infty) = +\infty$, the property (2.3) is equivalent to $g(x^*) - h(x^*) \leq g(x) - h(x)$ $\forall x \in U \cap \text{dom } g$.

A point x^* is said to be *a critical point* of $g - h$ if $\partial g(x^*) \cap \partial h(x^*) \neq \emptyset$.

It is worth noting the richness of $\mathcal{DC}(E)$ and $\mathcal{DC}_f(E)$ (see [16], [32], [33], and references therein).

D.c. programming is a natural extension of convex maximization in which the function g is the indicator function of a nonempty closed convex set C. In the convex analysis approach to nonsmooth nonconvex optimization, convex maximization has been extensively studied since 1974 by Pham Dinh (see [26], [27], [28], [29], and references therein), who has introduced subgradient algorithms for solving convex maximization problems.

The d.c. duality (due to Toland [39], who generalized in a very elegant and natural way the early work of Pham Dinh on convex maximization programming, mentioned above) associates the d.c. program (P_{dc}) with the so-called dual d.c. program

$$(D_{dc}) \qquad \alpha = \inf\{h^*(y) - g^*(y) : y \in E^*\},$$

with the help of the functional conjugate notion, and states relationships between them. More precisely, using the fundamental characterization of a convex function $\theta \in \Gamma_0(E)$ as *the pointwise supremum of a collection of affine minorizations*,

$$\theta(x) = \sup\{\langle x, y\rangle - \theta^*(y) : y \in E^*\} \quad \forall x \in E, \tag{2.4}$$

the d.c. duality is built by replacing the function h in problem (P_{dc}) with its corresponding expression of the form (2.4).

Thanks to a *symmetry* in the d.c. duality (the bidual d.c. program is exactly the primal one) and the *d.c. duality transportation of global minimizers*—the operator ∂h (resp., ∂g^*) transports the solution set of the primal problem (P_{dc}) (resp., the solution set of the dual problem (D_{dc})) into the solution set of the dual problem (D_{dc}) (resp., the solution set of the primal problem (P_{dc}))—solving a d.c. program implies solving the dual one and *vice versa*. This may be useful if one of them is easier to solve than the other. The equality of the optimal value in the primal and dual programs can be easily translated (with the help of the ϵ-*subdifferential* of the d.c. components) into global optimality conditions; namely, x^* is a global solution to (P_{dc}) if and only if

$$\partial_\epsilon h(x^*) \subset \partial_\epsilon g(x^*) \quad \forall \epsilon \geq 0.$$

Unfortunately, as we foresee, these conditions are rather difficult to use for devising solution methods to d.c. programs.

Local d.c. optimality conditions constitute (with the d.c. duality) the basis of the DCA. In general, it is not easy to state them as one does for the global d.c. optimality conditions, and they have been found to have very few properties which are useful in practice (see, e.g., [16], [32], [33], [19]).

REMARK 2.1. *Problem* (P_{dc}) *is a "false" d.c. program if the function* $f = g - h$ *is actually convex on E. For example, the problem of minimizing a convex function* f *on E can be (equivalently) cast in the d.c. framework as that of minimizing a d.c. function* $g - h$, *where* $g = f + \theta$, $h = \theta$, *and* θ *is a finite convex function on E. In such a case it is proved that the subdifferential inclusion* $\partial h(x^*) \subset \partial g(x^*)$ *is equivalent to* $0 \in \partial f(x^*)$; *i.e.,* x^* *is a solution to the problem being considered. Different ways of generating equivalent d.c. programs by using regularization techniques can be found in* [16], [33], [19]; *see also Remark* 2.2. *These proper features of the d.c. framework are crucial in the use of the DCA for solving nonconvex problems (or false d.c. programs), as will be shown in what follows: there are as many DCA as there are d.c. decompositions.*

2.1. The DCA for general d.c. programs. The DCA consists of the construction of the two sequences $\{x^k\}$ and $\{y^k\}$ (candidates for being primal and dual solutions, respectively) that we improve at each iteration (thus, the sequences $\{g(x^k) - h(x^k)\}$ and $\{h^*(y^k) - g^*(y^k)\}$ are decreasing) in an appropriate way such that their corresponding limits x^∞ and y^∞ satisfy the local optimality condition

$$\partial h(x^\infty) \subset \partial g(x^\infty) \quad \text{and} \quad \partial g^*(y^\infty) \subset \partial h^*(y^\infty), \quad \text{i.e.,} \quad (x^\infty, y^\infty) \in \mathcal{P}_l \times \mathcal{D}_l, \tag{2.5}$$

where $\mathcal{P}_l = \{x^* \in E : \partial h(x^*) \subset \partial g(x^*)\}$ and $\mathcal{D}_l = \{y^* \in E^* : \partial g^*(y^*) \subset \partial h^*(y^*)\}$, or are critical points of $g - h$ and $h^* - g^*$, respectively.

These sequences are generated as follows: x^{k+1} (resp., y^k) is a solution to the *convex program* (P_k) (resp., (D_k)) defined by

$$(P_k) \qquad \inf\{g(x) - [h(x^k) + \langle x - x^k, y^k\rangle] : x \in E\},$$

$$(D_k) \qquad \inf\{h^*(y) - [g^*(y^{k-1}) + \langle x^k, y - y^{k-1}\rangle] : y \in E^*\}.$$

In view of the observation that (P_k) (resp., (D_k)) is obtained from (P_{dc}) (resp., (D_{dc})) by replacing h (resp., g^*) with its affine minorization defined by $y^k \in \partial h(x^k)$ (resp., $x^k \in \partial g^*(y^{k-1})$), the DCA yields the next scheme:

$$y^k \in \partial h(x^k); \quad x^{k+1} \in \partial g^*(y^k). \tag{2.6}$$

This corresponds actually to the *simplified* DCA (which will be called DCA throughout the paper for simplicity), where x^{k+1} (resp., y^k) is arbitrarily chosen in $\partial g^*(y^k)$ (resp., $\partial h(x^k)$). In the *complete* form of DCA, we impose the following natural choice:

$$x^{k+1} \in \arg\min\{g(x) - h(x) : x \in \partial g^*(y^k)\} \tag{2.7}$$

and

$$y^k \in \arg\min\{h^*(y) - g^*(y) : y \in \partial h(x^k)\}. \tag{2.8}$$

Problems (2.7) and (2.8) are equivalent to convex maximization problems (2.9) and (2.10), respectively:

$$x^{k+1} \in \arg\min\{\langle x, y^k\rangle - h(x) : x \in \partial g^*(y^k)\}, \tag{2.9}$$

$$y^k \in \arg\min\{\langle x^k, y\rangle - g^*(y) : y \in \partial h(x^k)\}. \tag{2.10}$$

The complete DCA ensures that $(x^\infty, y^\infty) \in \mathcal{P}_l \times \mathcal{D}_l$. It can be viewed as a variation of the decomposition approach of the primal and dual problems (P_{dc}), (D_{dc}). From a practical point of view, although problems (2.7) and (2.8) are simpler than (P_{dc}), (D_{dc}) (we work in $\partial h(x^{k+1})$ and $\partial g^*(y^k)$ with convex maximization problems), they remain nonconvex programs and thus are still hard to solve. In practice, except for the cases in which the convex maximization problems (2.9) and (2.10) are easy to treat, one generally uses the simplified DCA to solve d.c. programs.

The DCA was introduced by Pham Dinh in 1986 as an extension of the aforementioned subgradient algorithms (for convex maximization programming) to d.c. programming [29]. However, this field has been really developed since 1994 by the joint work of Le Thi and Pham Dinh [16], [17], [18], [19], [32], [33] for solving nonsmooth nonconvex optimization problems. To our knowledge, DCA is actually one of a few algorithms (in the convex analysis approach to d.c. programming) which allow the solution of large-scale d.c. programs.

It had been proved by Pham Dinh and Le Thi (see, e.g., [16], [32], [33], [19]) that, for the simplified DCA, the sequences $\{x^k\}$ and $\{y^k\}$ enjoy the following properties:

(1) The sequences $\{g(x^k) - h(x^k)\}$ and $\{h^*(y^k) - g^*(y^k)\}$ are decreasing and

- $g(x^{k+1}) - h(x^{k+1}) \le h^*(y^k) - g^*(y^k) - \max\{\frac{\rho(h)}{2}\|x^{k+1} - x^k\|^2, \frac{\rho(h^*)}{2}\|y^{k+1} - y^k\|^2\} \le g(x^k) - h(x^k) - \delta_k$, where $\delta_k := \max\{\frac{\rho(g)+\rho(h)}{2}\|x^{k+1} - x^k\|^2, \frac{\rho(g^*)}{2}\|y^k - y^{k-1}\|^2 + \frac{\rho(h)}{2}\|x^{k+1} - x^k\|^2, \frac{\rho(g^*)}{2}\|y^k - y^{k-1}\|^2 + \frac{\rho(h^*)}{2}\|y^{k+1} - y^k\|^2\}$;

- $g(x^{k+1}) - h(x^{k+1}) = g(x^k) - h(x^k)$ if and only if $y^k \in \partial g(x^k) \cap \partial h(x^k)$, $y^k \in \partial g(x^{k+1}) \cap \partial h(x^{k+1})$, and $[\rho(g) + \rho(h)]\|x^{k+1} - x^k\| = 0$;
- $h^*(y^{k+1}) - g^*(y^{k+1}) = h^*(y^k) - g^*(y^k)$ if and only if $x^{k+1} \in \partial g^*(y^k) \cap \partial h^*(y^k)$, $x^{k+1} \in \partial g^*(y^{k+1}) \cap \partial h^*(y^{k+1})$, and $[\rho(g^*) + \rho(h^*)]\|y^{k+1} - y^k\| = 0$. In such a case DCA terminates at the kth iteration.

(2) If $\rho(g) + \rho(h) > 0$ (resp., $\rho(g^*) + \rho(h^*) > 0$), then the series $\{\|x^{k+1} - x^k\|^2\}$ (resp., $\{\|y^{k+1} - y^k\|^2\}$) converges.

(3) If the optimal value α of problem (P_{dc}) is finite and the sequences $\{x^k\}$ and $\{y^k\}$ are bounded, then every limit point x^∞ (resp., y^∞) of the sequence $\{x^k\}$ (resp., $\{y^k\}$) is a critical point of $g - h$ (resp., $h^* - g^*$).

(4) DCA has a linear convergence for general d.c. programs.

(5) In polyhedral d.c. programs (i.e., when g or h is polyhedral convex), the sequences DCA $\{x^k\}$ and $\{y^k\}$ contain finitely many elements, and DCA has a finite convergence.

We have the same results for the complete DCA, except that in (1) (resp., (3)) we must add the following property: $\partial h(x^k) \subset \partial g(x^k)$ and $\partial g^*(y^k) \subset \partial h^*(y^k)$ (resp., $\partial h(x^\infty) \subset \partial g(x^\infty)$ and $\partial g^*(y^\infty) \subset \partial h^*(y^\infty)$).

For more details, see [16], [32], [33], [19], and the references therein.

REMARK 2.2. *In general, the qualities (cost, robustness, stability, rate of convergence, and globality of sought solutions) of the DCA depend upon the d.c. decomposition of the function f. Assertion* (2) *shows how the strong convexity of d.c. components in primal and dual problems can influence the DCA. To make the d.c. components (of the primal objective function $f = g - h$) strongly convex, we usually apply the following decomposition (proximal regularization in d.c. programming):*

$$f = g - h = \left(g + \frac{\lambda}{2}\|\cdot\|^2\right) - \left(h + \frac{\lambda}{2}\|\cdot\|^2\right). \tag{2.11}$$

In this case the d.c. components in the dual problem will be differentiable. In the same way, inf-convolution of g and h with $\frac{\lambda}{2}\|\cdot\|^2$ will make the d.c. components (in the dual program) strongly convex and the d.c. components of the primal objective function differentiable. It is worth mentioning, for instance, that by using conjointly suitable d.c. decompositions of convex functions and proximal regularization techniques [20], [21], [35], *we can obtain the proximal point algorithm and the Goldstein–Levitin–Polyak subgradient method (in convex programming) as special cases of the DCA. For a detailed study of regularization techniques in d.c. programming, see* [16], [19], [30], [33]. *Since there are as many DCA as there are d.c. decompositions, it is of particular interest to study various equivalent d.c. forms for the primal and dual d.c. programs (see section* 3*).*

The choice of the d.c. decomposition of the objective function in a d.c. program and the initial point for the DCA are open questions to be studied. Of course, such a choice depends strongly on the very specific structure of the problem being considered. In practice, for solving a given d.c. program, we try to choose g and h such that sequences $\{x^k\}$ and $\{y^k\}$ can be easily calculated, i.e., either they are in explicit form or their computations are inexpensive. On the other hand, for a method based on local optimality conditions like DCA, it is crucial to point up different equivalent reformulations of the d.c. program which do not have the same local optimality, because they may serve to restart the DCA (escaping local solutions procedure); see sections 3 and 5.

The above description of DCA does not really reveal the main features of this approach that can partly explain its qualities *(low costs, robustness, stability, rate*

of convergence and globality of sought solutions) from the computational viewpoint. For *a deeper insight into DCA*, the reader is referred to [19].

The major difficulty in nonconvex programming resides in the fact that there are, in general, no practical global optimality conditions. Thus, checking the globality of solutions computed by local algorithms is possible only in the cases where optimal values are known a priori (*for example, they are zero in exact distance geometry problems*) or by comparison with global algorithms which, unfortunately, cannot be applied to large-scale problems. A pertinent comparison of local algorithms should be based on the following aspects:

+ mathematical foundations of the algorithms,

+ rate of convergence and running-time,

+ ability to treating large-scale problems,

+ quality of computed solutions: the lower the corresponding value of the objective is, the better the local algorithm will be,

+ the degree of dependence on initial points: the larger the set (composed of starting points, which ensure convergence of the algorithm to a global solution) is, the better the algorithm will be.

DCA seems to meet these standards since it was successfully applied to a lot of different nonconvex optimization problems, to which it gave global solutions and proved to be more robust and more efficient than related standard methods, especially in the large-scale setting (see [16], [17], [18], [32], [33], [19] and references therein).

We shall apply *all these d.c. enhancement features* to solving exact distance geometry problems (1.1) that are formulated as d.c. programs.

3. Solving the distance geometry problem by DCA. This section is devoted to the formulation of the exact distance geometry problem (1.1) in terms of d.c. programs and the computation of the sequences $\{X^k\}$ and $\{Y^k\}$ generated by DCA for solving them. It will be proved that both problems (EDP), with θ_{ij} defined by (1.4) or (1.5), are d.c. programs; moreover, the objective function of the former (the usual formulation of problem (1.1) as a global optimization problem) is infinitely differentiable, while the latter is a nondifferentiable nonconvex optimization problem. Paradoxically, the second formulation is advantageous in using DCA for solving the exact distance geometry problem since the sequences $\{X^k\}$ and $\{Y^k\}$ have explicit forms. On the other hand, the zero-gap of the Lagrangian duality relative to a special convex maximization problem allows the statement of interesting d.c. programs equivalent to the exact distance geometry problem.

A thorough study of the two chosen d.c. programs has been developed in the appropriate matrix framework. Substantial calculations of subdifferentials of related convex d.c. components prove that DCA requires matrix-vector products and only one Cholesky factorization. These results justify our theoretical choice of the d.c. program to be solved. Numerical simulations presented in section 5 will prove their practical efficiency.

The first nonconvex optimization problem equivalent to the exact distance geometry problem (1.1) is

$$(\mathrm{EDP}_1) \quad 0 = \min \left\{ \sigma(X) := \frac{1}{2} \sum_{(i,j)\in\mathcal{S},\, i<j} w_{ij} \left(\|X_i^T - X_j^T\| - \delta_{ij} \right)^2 : \quad X \in \mathcal{M}_{n,p}(\mathbb{R}) \right\}$$

(recall that $p = 3$ in the molecular problem). By identifying an $n \times p$ matrix X with a $p \times n-$ vector, in what follows, we use either $\mathcal{M}_{n,p}(\mathbb{R})$ or $\mathbb{R}^{p\times n}$ for indicating the same notation. We can identify by rows (resp., columns) each matrix $X \in \mathcal{M}_{n,p}(\mathbb{R})$ with a row-vector (resp., column-vector) in $(\mathbb{R}^p)^n$ (resp., $(\mathbb{R}^n)^p$) by writing, respectively,

$$X \longleftrightarrow \mathcal{X} = (X_1, \dots, X_n), \quad X_i^T \in \mathbb{R}^p,\ \mathcal{X}^T \in (\mathbb{R}^p)^n, \tag{3.1}$$

and

$$X \longleftrightarrow \overline{\mathcal{X}} = (X^1, \dots, X^p)^T, \quad X^i \in \mathbb{R}^n,\ \overline{\mathcal{X}} \in (\mathbb{R}^n)^p. \tag{3.2}$$

The inner product in $\mathcal{M}_{n,p}(\mathbb{R})$ is defined as the inner product in $(\mathbb{R}^p)^n$ or $(\mathbb{R}^n)^p$. That is,

$$\langle X, Y\rangle_{\mathcal{M}_{n,p}(\mathbb{R})} = \langle \mathcal{X}^T, \mathcal{Y}^T\rangle_{(\mathbb{R}^p)^n} = \sum_{i=1}^{n} \langle X_i^T, Y_i^T\rangle_{\mathbb{R}^p} = \sum_{i=1}^{n} X_i Y_i^T \tag{3.3}$$

$$= \langle \overline{\mathcal{X}}, \overline{\mathcal{Y}}\rangle_{(\mathbb{R}^n)^p} = \sum_{k=1}^{p} \langle X^k, Y^k\rangle_{\mathbb{R}^n} = \sum_{k=1}^{p} (X^k)^T Y^k = Tr(X^T Y). \tag{3.4}$$

Here $\mathrm{Tr}(X^T Y)$ denotes the trace of the matrix $X^T Y$. In what follows, for simplicity we shall suppress, where no ambiguity is possible, the indices for the inner product and denote by $\|\cdot\|$ the corresponding Euclidean norm on $\mathcal{M}_{n,p}(\mathbb{R})$. Evidently, we must choose either representation in a convenient way.

The data of (EDP_1) can be succinctly represented by a graph $G(N, \mathcal{S})$. The vertices $N = \{1, \dots, n\}$ correspond to the atoms, and an edge $(i, j) \in \mathcal{S}$ connects vertices i and j if the distance δ_{ij} between the corresponding atoms is known. The weight matrix $W = (w_{ij})$ of (EDP_1) is defined by taking $w_{ij} = 0$ when $(i, j) \notin \mathcal{S}$. Throughout this paper, we assume that W is irreducible, i.e., that the graph $G(N, \mathcal{S})$ is connected. This assumption is not restrictive for problem (EDP_1) since it can be decomposed into a number of smaller problems otherwise. Then we work under the next assumptions for the two symmetric matrices $\Delta = (\delta_{ij})$ (the distance matrix) and $W = (w_{ij})$:

(a1) for $i \neq j$, $\delta_{ij} > 0$ when $(i, j) \in \mathcal{S}$ (i.e., two different atoms are not in the same position), and $w_{ii} = 0$ for all i;
(a2) for $i \neq j$, $w_{ij} = 0$ if and only if δ_{ij} is unknown, say $(i, j) \notin \mathcal{S}$;
(a3) the weight matrix W is irreducible.

We note that if we set $\delta_{ij} = 0$ for $(i, j) \notin \mathcal{S}$, then $G(N, \mathcal{S})$ is the graph associated with the distance matrix Δ too.

The case in which $w_{ij} = c$ (c is a given positive number) for all $i \neq j$ is called *the normal case*. Clearly, this case can occur if and only if the distance matrix Δ is completely defined, say, when all pairwise distances are known.

We are now in a position to present, in the matrix framework, the two d.c. programs equivalent to the exact distance geometry problem (1.1): problems (3.6) and (3.12).

3.1. D.c. formulations of problem (EDP_1): The l_1-norm approach. The objective function of (EDP_1) can be written as

$$\sigma(X) = \frac{1}{2} \sum_{(i,j)\in\mathcal{S},\, i<j} w_{ij} d_{ij}^2(X) - \sum_{(i,j)\in\mathcal{S},\, i<j} w_{ij}\delta_{ij} d_{ij}(X) + \frac{1}{2}\eta_\delta^2, \tag{3.5}$$

with $d_{ij}(X) = \|X_i^T - X_j^T\|$ and $\eta_\delta := \big[\sum_{(i,j)\in\mathcal{S},\, i<j} w_{ij}\delta_{ij}^2\big]^{1/2}$.

Under assumption (a2), although δ_{ij} is unknown for any $(i,j) \notin \mathcal{S}$, in (3.5) the summation over pairs $(i,j) \in \mathcal{S}$ can be extended to that over all pairs (i,j). This fact must be taken into account later in subsection 3.2 while computing sequences of iterations in DCA.

Then (EDP_1) is equivalent to the following problem:

$$-\frac{1}{2}\eta_\delta^2 = \min\left\{F_1(X) := \frac{1}{2}\eta^2(X) - \xi(X) : \quad X \in \mathcal{M}_{n,p}(\mathbb{R})\right\}, \tag{3.6}$$

where η and ξ are the functions defined on $\mathcal{M}_{n,p}(\mathbb{R})$ by

$$\eta(X) = \left[\sum_{i<j} w_{ij} d_{ij}^2(X)\right]^{1/2} \quad \text{and} \quad \xi(X) = \sum_{i<j} w_{ij}\delta_{ij} d_{ij}(X). \tag{3.7}$$

It is not difficult to verify that η and ξ are two seminorms in $\mathcal{M}_{n,p}(\mathbb{R})$, and thus (3.6) is a d.c. program to which the DCA can be applied.

As indicated above, an important issue in the DCA is a *good* d.c. decomposition of the problem being considered. For this purpose, by using the Lagrangian duality with zero gap relative to the problem of maximization of a finite gauge ψ over the unit ball defined by a finite gauge ϕ such that $\phi^{-1}(0)$ is a subspace contained in $\psi^{-1}(0)$ (see [16], [31]), we will state a problem equivalent to (EDP_1), which is a d.c. program too. Let us first recall the following result (see [16], [31]).

LEMMA 3.1. *The convex maximization program*

$$\omega := \max\{\xi(X) : \quad X \in U(\eta)\}, \tag{3.8}$$

with $U(\eta) := \{X \in \mathcal{M}_{n,p}(\mathbb{R}) : \eta(X) \le 1\}$ *formulated as a d.c. program (*$\chi_{\mathcal{U}(\eta)}$ *is the indicator of* $\mathcal{U}(\eta)$ *defined in section* 2*)*

$$-\omega := \min\{\chi_{\mathcal{U}(\eta)}(X) - \xi(X) : \quad X \in \mathcal{M}_{n,p}(\mathbb{R})\}, \tag{3.9}$$

is equivalent to the d.c. program

$$-\frac{\omega^2}{2} = \min\left\{F_1(X) := \frac{1}{2}\eta^2(X) - \xi(X) : \quad X \in \mathcal{M}_{n,p}(\mathbb{R})\right\} \tag{3.10}$$

in the sense that

(i) *the solutions to problem* (3.9) *are of the form* $X^*/\eta(X^*)$*, with* X^* *being a solution to problem* (3.10)*;*

(ii) *the solutions to problem* (3.10) *are of the form* $\xi(X^*)X^*$*, with* X^* *being a solution to problem* (3.9)*.*

For our distance geometry problem, $\omega = \eta_\delta$, and the next useful result is a consequence of Lemma 3.1.

PROPOSITION 3.2. *The convex maximization program*

$$\eta_\delta^2 = \max\{\xi(X) : \quad \eta(X) \le \eta_\delta\}, \tag{3.11}$$

formulated as a d.c. program

$$-\eta_\delta^2 = \min\{\chi_C(X) - \xi(X) : \quad X \in \mathcal{M}_{n,p}(\mathbb{R})\} \tag{3.12}$$

with $C := \{X \in \mathcal{M}_{n,p}(\mathbb{R}) : \frac{1}{2}\eta^2(X) \leq \frac{1}{2}\eta_\delta^2\} = \eta_\delta U(\eta)$, *is equivalent to the d.c. program*

$$-\frac{\eta_\delta^2}{2} = \min\left\{F_1(X) := \frac{1}{2}\eta^2(X) - \xi(X) : \quad X \in \mathcal{M}_{n,p}(\mathbb{R})\right\} \tag{3.13}$$

in the sense that they have the same solution set. Moreover, X^* *solves these problems if and only if* $\xi(X^*) = \eta^2(X^*) = \eta_\delta^2$.

3.1.1. The dissimilarity geometry problem in Phase 1. As has been said in section 1, phase 1 of our approach is concerned with the complete dissimilarity matrix $\widetilde{\Delta} = (\widetilde{\delta}_{ij}), (i,j) \in \{1,\dots,n\}^2$, with $\widetilde{\delta}_{ij}$ being the length of the shortest path between atoms i and j (section 4) and the resulting d.c. programs in the normal case

$$\min\left\{\frac{1}{2}\sum_{i<j} c(\tilde{\delta}_{ij} - \|X_i^T - X_j^T\|)^2 : X \in \mathcal{M}_{n,p}(\mathbb{R})\right\}, \tag{3.14}$$

$$\min\left\{\widetilde{F}_1(X) := \frac{1}{2}\widetilde{\eta}^2(X) - \widetilde{\xi}(X) : \quad X \in \mathcal{M}_{n,p}(\mathbb{R})\right\}, \tag{3.15}$$

$$\min\left\{\chi_{\widetilde{C}}(X) - \widetilde{\xi}(X) : \quad X \in \mathcal{M}_{n,p}(\mathbb{R})\right\}, \tag{3.16}$$

where

$$\widetilde{\eta}(X) := \left[c\sum_{i<j} d_{ij}^2(X)\right]^{1/2}, \qquad \widetilde{\xi}(X) = c\sum_{i<j}\widetilde{\delta}_{ij}d_{ij}(X),$$
$$\widetilde{C} := \{X \in \mathcal{M}_{n,p}(\mathbb{R}) : \widetilde{\eta}(X) \leq \widetilde{\eta}_{\tilde{\delta}}\}.$$

Here c is a positive number, $\tilde{\eta}_{\tilde{\delta}} := [c\sum_{i<j}\tilde{\delta}_{ij}^2]^{1/2}$, and the summations are taken for all $(i,j) \in \{1,\dots,n\}^2$.

The next subsection is devoted to the description of the DCA applied to problems (3.6) and (3.15) on the one hand, and problems (3.12) and (3.16) on the other hand. Performing this scheme is thus reduced to calculating subdifferentials of the functions ξ, $((1/2)\eta^2)$, $((1/2)\eta^2)^*$, and $\chi_{\mathcal{C}}^*$.

3.2. Solving (3.6), (3.15), and (3.12), (3.16) by the DCA. Under assumptions (a2) and (a3), we can restrict the working matrix space to an appropriate set, which is, as will be seen later, favorable to our calculations. Indeed, let $\mathcal{A}$ denote the set of matrices in $\mathcal{M}_{n,p}(\mathbb{R})$ whose rows are all identical, i.e.,

$$\mathcal{A} := \{X \in \mathcal{M}_{n,p}(\mathbb{R}) : X_1 = \cdots = X_n\},$$

and let $P_{\mathcal{A}}$ be the orthogonal projection on $\mathcal{A}$; we then have the following result.

LEMMA 3.3.

(i) $\mathcal{A} = \{ev^T : v \in \mathbb{R}^p\}$ *is a p-dimensional subspace of* $\mathcal{M}_{n,p}(\mathbb{R})$, *whose orthogonal subspace is given by* $\mathcal{A}^\perp = \{Y \in \mathcal{M}_{n,p}(\mathbb{R}) : \sum_{i=1}^n Y_i = 0\}$.

(ii) $\mathcal{A} \subset \xi^{-1}(0)$; $\mathcal{A} \subset \eta^{-1}(0)$.

(iii) $P_{\mathcal{A}} = (1/n)ee^T; P_{\mathcal{A}^\perp} = I - (1/n)ee^T$ *(e is the vector of ones in* $\mathbb{R}^n$*).*

(iv) *If the weight matrix W is irreducible (resp., W is irreducible and $w_{ij}\delta_{ij} > 0$ whenever $w_{ij} > 0$), then $\mathcal{A} = \eta^{-1}(0)$ (resp., $\mathcal{A} = \xi^{-1}(0)$). If $\mathcal{A} = \eta^{-1}(0) = \xi^{-1}(0)$, then the problems*

$$-\frac{\eta_\delta^2}{2} = \min\left\{\frac{1}{2}\eta^2(X) - \xi(X) : \quad X \in \mathcal{A}^\perp\right\} \tag{3.17}$$

and

$$-\eta_\delta^2 = \min\{\chi_{\mathcal{C}}(X) - \xi(X) : \quad X \in \mathcal{A}^\perp\} \tag{3.18}$$

have the same solution set. Moreover, X^* is an optimal solution of (3.17) *(resp.,* (3.18)*) if and only if $X^* + Z$ is an optimal solution of* (3.13) *(resp.,* (3.12)*) for all $Z \in \mathcal{A}$.*

Proof. (i) and (ii) are straightforward from the definition of $\mathcal{A}$. The proof of (iii) is easy. To prove (iv), let $X \in \mathcal{M}_{n,p}(\mathbb{R})$ such that $\eta(X) = 0$ (or $\xi(X) = 0$) and $(i,j) \in \{1,\dots,n\}^2$ with $i \neq j$. Since the matrix W is irreducible, there is a finite sequence $\{i_1,\dots,i_r\} \subset \{1,\dots,n\}$ verifying $w_{ii_1} > 0, w_{i_k i_{k+1}} > 0$ for $k = 1,\dots,r-1$, and $w_{i_r j} > 0$. It follows that $X_i = X_{i_1} = \cdots = X_{i_r} = X_j$, and then $\eta^{-1}(0) = \mathcal{A} = \xi^{-1}(0)$. The remaining part is a direct consequence of [31]. The proof is thus completed. □

REMARK 3.4. *As a consequence of Lemma* 3.3*, the restrictions of the seminorms η and ξ on the subspace $\mathcal{A}^\perp$ are actually norms under the assumptions* (a1), (a2), *and* (a3)*. It follows that their polars η^0 and ξ^0 defined by* [35]*,*

$$\begin{aligned}\eta^0(Y) &= \sup\{\langle X, Y\rangle : \eta(X) \leq 1\} \qquad \forall Y \in \mathcal{M}_{n,p}(\mathbb{R}),\\ \xi^0(Y) &= \sup\{\langle X, Y\rangle : \eta(X) \leq 1\} \qquad \forall Y \in \mathcal{M}_{n,p}(\mathbb{R})\end{aligned}$$

satisfy the following properties:

(i) $\eta^0(Y) = \xi^0(Y) = +\infty$ *if* $Y \notin \mathcal{A}^\perp$,

(ii) $\eta^0(Y) = \sup\{\langle X, Y\rangle : X \in \mathcal{A}^\perp, \eta(X) \leq 1\}\ \forall Y \in \mathcal{A}^\perp$,
$\xi^0(Y) = \sup\{\langle X, Y\rangle : X \in \mathcal{A}^\perp,\ \eta(X) \leq 1\}\ \forall Y \in \mathcal{A}^\perp$.

We shall now compute subdifferentials of the functions ξ, $((1/2)\eta^2)^*$, and $\partial\chi_{\mathcal{C}}^*$. These calculations will fortunately permit us to state new matrix expressions of these functions and thus to provide the simplest computations of the sequences $\{X^{(k)}\}$ and $\{Y^{(k)}\}$ generated by the DCA applied to problems (3.6) and (3.12). They also point out interesting relations between the trust region subproblem and problem (3.6).

3.2.1. Calculation of $\partial\xi$. By definition, $\xi(X) = \sum_{i<j} w_{ij}\delta_{ij}d_{ij}(X)$. Thus, $\partial\xi(X) = \sum_{i<j} w_{ij}\delta_{ij}\partial d_{ij}(X)$. Further, since d_{ij} can be expressed as (using the row representation of $X \in \mathcal{M}_{n,p}(\mathbb{R})$)

$$d_{ij} = \|\cdot\| \circ L_{ij} : (\mathbb{R}^p)^n \longrightarrow \mathbb{R}^p \longrightarrow \mathbb{R},$$

$$X \longmapsto L_{ij}(X) = X_i^T - X_j^T \longmapsto \|X_i^T - X_j^T\|,$$

it follows [35] that $\partial d_{ij}(X) = L_{ij}^T\partial\left(\|\cdot\|\right)(L_{ij}(X))$. Hence

$$Y(i,j) \in \partial d_{ij}(X) \Leftrightarrow Y(i,j) = L_{ij}^T y, \quad y \in \partial\left(\|\cdot\|\right)(X_i^T - X_j^T),$$

which implies

$$Y(i,j)_k = 0 \text{ if } k \notin \{i,\, j\} \quad \text{and} \quad Y(i,j)_i^T = -Y(i,j)_j^T \in \partial\left(\|\cdot\|\right)(X_i^T - X_j^T). \tag{3.19}$$

Thus, ξ is *not differentiable* on the closed set $\{X \in \mathcal{M}_{n,p}(\mathbb{R}) : X_i = X_j \text{ for } (i,j) \in \mathcal{S}, i < j\}$ but on its complement in $\mathcal{M}_{n,p}(\mathbb{R})$, i.e., the open set Ω defined by

$$\Omega = \{X \in \mathcal{M}_{n,p}(\mathbb{R}) : \|X_i^T - X_j^T\| > 0 \ \forall (i,j) \in \mathcal{S}, i < j\}. \tag{3.20}$$

It is clear that

$$\Omega + \mathcal{A} = \Omega. \tag{3.21}$$

Now for $(i,j) \in \mathcal{S}, i < j$, let us choose the particular subgradient $Y(i,j) \in \partial d_{ij}(X)$ defined by

$$Y(i,j)_i = -Y(i,j)_j = \begin{cases} \frac{X_i - X_j}{\|X_i^T - X_j^T\|} & \text{if } X_i \neq X_j, \\ 0 & \text{if } X_i = X_j. \end{cases} \tag{3.22}$$

In this case, the resulting subgradient $Y \in \partial\xi(X)$ is explicitly given by

$$\begin{aligned} Y_k &= \sum_{i<j} w_{ij}\delta_{ij} Y(i,j)_k = \sum_{i<k} w_{ik}\delta_{ik} Y(i,k)_k + \sum_{j>k} w_{kj}\delta_{kj} Y(k,j)_k \\ &= \sum_{i<k} w_{ki}\delta_{ki} s_{ki}(X)(X_k - X_i) + \sum_{j>k} w_{kj}\delta_{kj} s_{kj}(X)(X_k - X_j) \\ &= \left[\sum_{i=1}^n w_{ki}\delta_{ki} s_{ki}(X)\right] X_k - \sum_{i=1}^n w_{ki}\delta_{ki} s_{ki}(X) X_i, \end{aligned}$$

where

$$s_{ij}(X) = \begin{cases} \frac{1}{\|X_i^T - X_j^T\|} & \text{if } X_i \neq X_j, \\ 0 & \text{otherwise.} \end{cases}$$

Let $B(X) = (b_{ij}(X))$ be the $n \times n$ matrix defined by

$$b_{ij}(X) = \begin{cases} -w_{ij}\delta_{ij} s_{ij}(X) & \text{if } i \neq j, \\ -\sum_{k=1, k\neq i}^n b_{ik}(X) & \text{if } i = j. \end{cases} \tag{3.23}$$

It follows that

$$Y = B(X)X, \ B(X+U) = B(X) \qquad \forall X \in \mathcal{M}_{n,p}(\mathbb{R}), \forall U \in \mathcal{A}. \tag{3.24}$$

In what follows, for $i \neq j$, M_{ij} denotes the $n \times n$ matrix given by $M_{ij} = e_i e_i^T + e_j e_j^T - (e_i e_j^T + e_j e_i^T)$, where $\{e_i : i = 1, \ldots, n\}$ forms the canonical basis of $\mathbb{R}^n$. It is clear that the particular subgradient $Y(i,j) \in \partial d_{ij}(X)$ relates to M_{ij} by

$$Y(i,j) = s_{ij}(X) M_{ij} X. \tag{3.25}$$

We will denote by $\mathcal{N}(M)$ and $\operatorname{Im} M$ the null space and the range of the matrix M, respectively.

PROPOSITION 3.5. *Let $B(X)$ be the matrix defined by* (3.23)*. Then we have the following:*

(i) $\mathcal{N}(B(X)) \supset \mathcal{A}$, $\operatorname{Im}(B(X)) \subset \mathcal{A}^\perp$ *for all* $X \in \mathcal{M}_{n,p}(\mathbb{R})$*. Moreover, the preceding inclusions become inequalities under the assumptions* (a1), (a2), *and* (a3).

(ii) $B(X)$ *depends only on* $X_i - X_j$ *for* $(i,j) \in \mathcal{S}$, $i < j$, *and* $B : \mathcal{M}_{n,p}(\mathbb{R}) \longmapsto \Sigma_n^+$ *(the set of* $n \times n$ *symmetric positive semidefinite matrices) is continuous on* Ω *and* $B(X)X \in \partial\xi(X)$ *for all* $X \in \mathcal{M}_{n,p}(\mathbb{R})$.

(iii) *The seminorm* ξ *is differentiable (and so continuously differentiable) on* Ω, *and* $\xi(X) = \langle X, B(X)X\rangle$ *for all* $X \in \mathcal{M}_{n,p}(\mathbb{R})$.

(iv) $\langle X, B(Y)Y\rangle \le \langle X, B(X)X\rangle$ *for all* $X, Y \in \mathcal{M}_{n,p}(\mathbb{R})$.

Proof. (i) follows immediately from Lemma 3.3 and the facts that $\mathcal{A} = \{ev^T : v \in \mathbb{R}^p\}$ and $B(X)e = 0$ for all $X \in \mathcal{M}_{n,p}(\mathbb{R})$.

(ii) $B(X)$ is symmetric and diagonally dominant, and its diagonal entries are nonnegative. Thus it is positive semidefinite [40]. The continuity of the mapping B on Ω directly follows from (3.23).

(iii) The differentiability of the seminorm ξ is straightforward from (3.22) since the subdifferential $\partial\xi(X)$ is reduced to the singleton $\{B(X)X\}$ for $X \in \Omega$. The remaining equality is in fact the generalized Euler relation for convex nondifferentiable functions which are positively homogeneous of degree 1 [35].

(iv) By the definition of the subdifferential, it follows from assertion (ii) that

$$\xi(X) = \langle X, B(X)X\rangle \ge \xi(Y) + \langle X - Y, B(Y)Y\rangle \quad \forall X, Y \in \mathcal{M}_{n,p}(\mathbb{R}),$$

and thus the proof is completed since $\xi(Y) = \langle Y, B(Y)Y\rangle$. □

REMARK 3.6. *We have (see* [35]*)*

$$\partial\xi(X) = \{Y \in \mathcal{A}^\perp : \xi^0(Y) \le 1,\ \langle X, Y\rangle = \xi(X)\} \qquad \forall X \in \mathcal{M}_{n,p}(\mathbb{R}),$$

and the range of the subdifferential $\partial\xi$ *then is bounded (Remark* 3.4*):*

$$range\ \ \partial\xi = \{Y \in \mathcal{A}^\perp : \xi^0(Y) \le 1\}.$$

3.2.2. Calculation of $\partial((1/2)\eta^2)^*$. First we state some fundamental properties of the function $(1/2)\eta^2$ which will be needed for the calculation of $\partial((1/2)\eta^2)^*$. From the definition of η, say $\eta^2(X) = \sum_{i<j} w_{ij}\|X_i^T - X_j^T\|^2 = \sum_{i<j} w_{ij} d_{ij}^2(X)$, we have $\partial\left(\frac{1}{2}\eta^2\right)(X) = \sum_{i<j} w_{ij} d_{ij}(X)\partial d_{ij}(X)$. Thus

$$Y \in \partial\left(\frac{1}{2}\eta^2\right)(X) \Leftrightarrow Y = \sum_{i<j} w_{ij} d_{ij}(X) Y(i,j),$$

with $Y(i,j)$ being defined by (3.19). It follows that η^2 is differentiable on $\mathcal{M}_{n,p}(\mathbb{R})$, and $Y = \nabla(\frac{1}{2}\eta^2)(X)$ is defined as

$$Y_k = \sum_{i<k} w_{ki}(X_k - X_i) + \sum_{j>k} w_{kj}(X_k - X_j) = \left(\sum_{i=1}^n w_{ki}\right) X_k - \sum_{i=1}^n w_{ki} X_i.$$

Hence $Y = VX$, where $V = (v_{ij})$ given by

$$v_{ij} = \begin{cases} -w_{ij} & \text{if } i \ne j, \\ \sum_{k=1}^n w_{ik} & \text{if } i = j. \end{cases} \tag{3.26}$$

Similarly to Proposition 3.5 for the function ξ, one has the following results.

PROPOSITION 3.7. *Let* V *be the matrix defined by* (3.26). *Then the following hold:*

(i) V *is positive semidefinite,* $\nabla(\frac{1}{2})\eta^2(X) = VX$, *and* $\eta^2(X) = \langle X, VX\rangle$.

(ii) *If the weight matrix W is irreducible (assumption* (a3)*), then* $\mathcal{A} = \eta^{-1}(0) = \{X \in \mathcal{M}_{n,p}(\mathbb{R}) : VX = 0\} = \mathcal{N}(V)$, *rank* $V = n-1$, *and* $\mathcal{A}^\perp = \{Y = VX : X \in \mathcal{M}_{n,p}(\mathbb{R})\} = \mathrm{Im}V$.

(iii) $(\frac{1}{2}\eta^2)^*(Y) = \frac{1}{2}\langle Y, V^+Y\rangle$ *if* $Y \in \mathcal{A}^\perp$, *and* $+\infty$ *otherwise. In other words,*

$$\left(\frac{1}{2}\eta^2\right)^*(Y) = \frac{1}{2}\langle V^+Y, Y\rangle + \chi_{\mathcal{A}^\perp}(Y) \quad \textit{for } Y \in \mathcal{M}_{n,p}(\mathbb{R}).$$

(iv) *dom* $(\frac{1}{2}\eta^2)^* = dom\ \partial(\frac{1}{2}\eta^2)^*$ *and* $\partial(\frac{1}{2}\eta^2)^*(Y) = V^+Y + \mathcal{A}$ *for* $Y \in \mathcal{A}^\perp$.

Proof. (i) The positive semidefiniteness of V follows from [40] as in Proposition 3.5. Since $\nabla(\frac{1}{2}\eta^2)(X) = VX$, the generalized Euler relation [35] yields $\eta^2(X) = \langle X, VX\rangle$.

(ii) It remains to prove that rank $V = n-1$, since the other assertions follow from the fact that V is symmetric positive semidefinite. First, we see that rank $V \le n-1$ because $Ve = 0$. Suppose now rank $V < n-1$. Then there exists $v \notin \mathbb{R}e$ such that $Vv = 0$. Let $X = v(e^p)^T$ (e^p is the vector of ones in $\mathbb{R}^p$). Clearly, $VX = 0$, and therefore $X \in \mathcal{A}$. By the definition of $\mathcal{A}$, all rows of X are identical, which implies that $v \in \mathbb{R}e$. This contradiction proves that rank $V = n-1$.

(iii) By the definition $(\frac{1}{2}\eta^2)^*(Y) := \sup\{\langle X, Y\rangle - (\frac{1}{2}\eta^2)(X) : X \in \mathcal{M}_{n,p}(\mathbb{R})\}$, it follows that $(\frac{1}{2}\eta^2)^*(Y) = +\infty$ if $Y \notin \mathcal{A}^\perp$. For $Y \in \mathcal{A}^\perp$, X solves the above problem if and only if $VX = Y$, i.e., $X \in V^+Y + \mathcal{A}$, where V^+ denotes the pseudoinverse of V. Hence $(\frac{1}{2}\eta^2)^*(Y) = \frac{1}{2}\langle V^+Y, Y\rangle$ if $Y \in \mathcal{A}^\perp$, and thus $(\frac{1}{2}\eta^2)^*(Y) = \frac{1}{2}\langle V^+Y, Y\rangle + \chi_{\mathcal{A}^\perp}(Y)$ for $Y \in \mathcal{M}_{n,p}(\mathbb{R})$. Since V^+ is symmetric positive semidefinite, we have $\partial(\frac{1}{2}\eta^2)^*(Y) = V^+Y + \mathcal{A}$ for $Y \in \mathcal{A}^\perp$. The proof then is completed. □

Hence, determining the gradient of $(\frac{1}{2}\eta^2)^*(Y)$ with $Y \in \mathcal{A}^\perp$ amounts to computing the pseudoinverse of V. The next result permits us to calculate V^+.

PROPOSITION 3.8. *If the weight matrix W is irreducible (assumption* (a3)*), then we have the following:*

(i) $\mathrm{Im}V^+ = \mathrm{Im}V = \mathcal{A}^\perp$ *and* $\langle V^+Y, Y\rangle > 0$ *for* $Y \in \mathcal{A}^\perp \setminus \{0\}$.

(ii) $V^+Y = (V + \frac{1}{n}ee^T)^{-1}Y - \frac{1}{n}ee^TY$ *for all* $Y \in \mathcal{M}_{n,p}(\mathbb{R})$. *That implies, for* $Y \in \mathcal{A}^\perp$,

$$X = V^+Y = \left(V + \frac{1}{n}ee^T\right)^{-1}Y, \qquad \textit{i.e.,}\ \left(V + \frac{1}{n}ee^T\right)X = Y. \tag{3.27}$$

Hence, in the normal case where $V = ncI - cee^T$, *the solution to* (3.27) *is* $X = Y/(nc)$. *In other words,*

$$V^+Y = \frac{Y}{nc} \qquad \textit{for } Y \in \mathcal{A}^\perp. \tag{3.28}$$

Proof. Assertion (i) is a well-known property for pseudoinverses of symmetric positive semidefinite matrices (see also (3.27)), while assertion (ii) is easy to prove and is omitted here. □

3.2.3. Calculation of $\partial\chi^*_{\mathcal{C}}$. Recall that $C = \{X \in \mathcal{M}_{n,p}(\mathbb{R}) : \eta(X) \le \eta_\delta\}$. According to [35], $\chi^*_{\mathcal{C}}$ is η_δ times the polar η^0 of the gauge (seminorm) η:

$$\eta^0(Y) := \sup\{\langle X, Y\rangle : \eta(X) \le 1\} = \sup\left\{\langle X, Y\rangle : \frac{1}{2}\eta^2(X) \le \frac{1}{2}\right\}.$$

We have $\eta^0(Y) = +\infty$ if $Y \notin \mathcal{A}^\perp$. Let now $Y \in \mathcal{A}^\perp$. It is clear that X is an optimal solution to the above problem if and only if there is a positive number λ such that

(1) $\langle X, VX\rangle \leq 1$, (2) $Y = \lambda VX$, and (3) $\lambda(\langle X, VX\rangle - 1) = 0$. First assume that $Y \neq 0$. Then λ must be positive, and (2) implies that $X \in \frac{1}{\lambda}V^+Y + \mathcal{A}$. The value of λ is given according to (3) as $\lambda = \langle Y, V^+Y\rangle^{\frac{1}{2}}$. Hence $\eta^0(Y) = \langle Y, V^+Y\rangle^{\frac{1}{2}}$.

This formulation holds also for $Y = 0$ because $\eta^0(0) = 0$. Finally we get

$$\eta^0(Y) = \eta_\delta\langle Y, V^+Y\rangle^{\frac{1}{2}} + \chi_{\mathcal{A}^\perp}(Y) \qquad \forall Y \in \mathcal{M}_{n,p}(\mathbb{R}). \tag{3.29}$$

It follows that ($U(\eta)$ denotes the unit ball of the seminorm η).

PROPOSITION 3.9. (i) *The support function $\chi^*_{U(\eta)}$ of $U(\eta)$ is the polar η^0 of η, and we have $\chi^*_{U(\eta)}(Y) = \eta^0(Y) = \langle Y, V^+Y\rangle^{\frac{1}{2}} + \chi_{\mathcal{A}^\perp}(Y)$ $\forall Y \in \mathcal{M}_{n,p}(\mathbb{R})$ and $(\frac{1}{2}\eta^2)^* = \frac{1}{2}(\eta^0)^2$. Hence*

$$\partial\chi^*_{U(\eta)}(Y) = \begin{cases} \emptyset & \text{if } Y \notin \mathcal{A}^\perp, \\ V^+Y/\langle Y, V^+Y\rangle^{1/2} + \mathcal{A} & \text{if } Y \in \mathcal{A}^\perp \setminus \{0\}, \\ U(\eta) & \text{if } Y = 0. \end{cases} \tag{3.30}$$

The last expression is very simple in the normal case, since we have $V^+Y = Y/nc$ for $Y \in \mathcal{A}^\perp$. Therefore

$$\partial\chi^*_{U(\eta)}(Y) = \begin{cases} \emptyset & \text{if } Y \notin \mathcal{A}^\perp, \\ Y/(\sqrt{nc}\|Y\|) + \mathcal{A} & \text{if } Y \in \mathcal{A}^\perp \setminus \{0\}, \\ U(\eta) & \text{if } Y = 0. \end{cases} \tag{3.31}$$

(ii) *For $C := \eta_\delta U(\eta)$, we have $\chi^*_C = \eta_\delta\chi^*_{U(\eta)}$.*

Before going further, it is worth noting the following crucial consequences on both theoretical and algorithmic viewpoints of DCA of solving problems (3.6) and (3.12).

REMARK 3.10. (i) *It follows from the very definition of the seminorms η and ξ and the Cauchy–Schwarz inequality $\xi(X) \leq \eta_\delta\eta(X)$ $\forall X \in \mathcal{M}_{n,p}(\mathbb{R})$. Hence $\frac{1}{\eta_\delta}\eta^0 = (\eta_\delta\eta)^0 \leq \xi^0$.*

(ii) *We have $VV^+Y = Y = V^+VY$ for $Y \in \mathrm{Im}V = \mathrm{Im}V^+ = \mathcal{A}^\perp$. Hence*

$$\eta(V^+Y) = \langle Y, V^+Y\rangle^{1/2} = \eta^0(Y) \leq \eta_\delta \quad \text{for } Y \in \mathcal{A}^\perp,\ \xi^0(Y) \leq 1. \tag{3.32}$$

(iii) *Under the assumptions* (a1), (a2), *and* (a3), *X^* is a solution to problem* (EDP_1) *if and only if $X^* \in \Omega$ and*

$$B(X^*) = V \tag{3.33}$$

according to (3.23), (3.26). *Moreover, $\rho(\frac{1}{2}\eta^2, \mathcal{A}^\perp)$ and $\rho((\frac{1}{2}\eta^2)^*, \mathcal{A}^\perp)$ are positive.*

From the above displayed calculations, we can now give the description of the DCA applied to (3.6) and (3.12) (or, equivalently (3.17) and (3.18); i.e., the computations of the sequences $\{X^{(k)}\}$ and $\{Y^{(k)}\}$ in $\mathcal{A}^\perp$ generated by the algorithm.

3.2.4. The description of DCA for solving (3.6) and (3.15). We present below the DCA applied to problems (3.6) and (3.15), which are respectively denoted by DCA1 and DCA1bis. The latter will be used to compute an initial point for the former.

DCA1 (DCA applied to (3.6)). Generate two sequences $\{X^{(k)}\}$ and $\{Y^{(k)}\}$ in $\mathcal{A}^\perp$ as follows:

Let $\tau_1 > 0$, $\tau_2 > 0$, and $0 \neq X^{(0)} \in \mathcal{A}^\perp$ be given.

For $k = 0, 1, \ldots$ until

$$\text{either } \|X^{(k+1)} - X^{(k)}\| \le \tau_1 \|X^{(k+1)}\|$$
$$\text{or } |F_1(X^{(k)}) - F_1(X^{(k+1)})| \le \tau_2(|F_1(X^{(k+1)})| + 1)$$

take

$$Y^{(k)} = B(X^{(k)})X^{(k)},$$

$$X^{(k+1)} = V^+ Y^{(k)}. \tag{3.34}$$

DCA1BIS (DCA applied to (3.15)). Replace (3.34) in DCA1 by

$$X^{(k+1)} = \frac{1}{nc} Y^{(k)}. \tag{3.35}$$

The main results on DCA for general d.c. programs (see section 2) can be refined as follows.

PROPOSITION 3.11. *The sequences $\{X^{(k)}\}$ and $\{Y^{(k)}\}$ generated by DCA1 satisfy the following properties:*

(i) $\eta(X^{(k+1)}) = \eta(V^+ Y^{(k)}) = \langle Y^{(k)}, V^+ Y^{(k)}\rangle^{1/2} = \eta^0(Y^{(k)}) \le \eta_\delta \ \ \forall k$*; i.e., they are bounded.*

(ii)

$$\frac{1}{2}\eta^2(X^{(k+1)}) - \xi(X^{(k+1)}) \le -\frac{1}{2}\langle Y^{(k)}, V^+ Y^{(k)}\rangle \le \frac{1}{2}\eta^2(X^{(k)}) - \xi(X^{(k)}) - \delta_k \qquad \forall k,$$

where $\delta_k := \max\{\frac{1}{2}\rho(\frac{1}{2}\eta^2, \mathcal{A}^\perp)\|X^{(k+1)} - X^{(k)}\|^2, \frac{1}{2}\rho(\frac{1}{2}(\eta^0)^2, \mathcal{A}^\perp)\|Y^{(k)} - Y^{(k-1)}\|^2\}$.

(iii) *The sequences $\{\eta(X^{(k)})\}, \{\xi(X^{(k)})\}$, and $\{\eta^0(Y^{(k)})\}$ are increasing:*

$$\eta^2(X^{(k)}) \le \xi(X^{(k)}) \le \frac{1}{2}[\eta^2(X^{(k)}) + \eta^2(X^{(k+1)})] - \delta_k \le \eta_\delta - \delta_k \qquad \forall k,$$

$$\xi(X^{(k+1)}) \ge \xi(X^{(k)}) + \frac{1}{2}[\eta^2(X^{(k+1)}) - \eta^2(X^{(k)})] + \delta_k \qquad \forall k.$$

(iv) *(Finite convergence of DCA1.)* $\frac{1}{2}\eta^2(X^{(k+1)}) - \xi(X^{(k+1)}) = \frac{1}{2}\eta^2(X^{(k)}) - \xi(X^{(k)})$ *if and only if $X^{(k+1)} = X^{(k)}$ (or equivalently, $Y^{(k)} = Y^{(k-1)}$). In such a case DCA1 stops at the kth iteration with*

$$Y^{(k)} = B(X^{(k)})X^{(k)} = VX^{(k)}, \qquad \xi(X^{(k)}) = \eta^2(X^{(k)}) = (\eta^0)^2(Y^{(k-1)}),$$

and $X^{(k)}$ solves problem (3.17) *if and only if $\xi(X^{(k)}) = \eta_\delta^2$ (i.e., $\eta(X^{(k)}) = \eta_\delta$).*

(v) *(Infinite convergence of DCA1.) If the sequences $\{X^{(k)}\}$ and $\{Y^{(k)}\}$ are infinite, then the two series $\{\|X^{(k+1)} - X^{(k)}\|^2\}$, $\{\|Y^{(k+1)} - Y^{(k)}\|^2\}$ converge, and we have for every limit point (X^*, Y^*) of $\{X^{(k)}, Y^{(k)}\}$*

$$X^* = V^+ Y^*, \quad Y^* = VX^* \in \partial\xi(X^*), \quad \xi(X^*) = \eta^2(X^*) = (\eta^0)^2(Y^*).$$

In such a case, X^* solves problem (3.17) *if and only if $\xi(X^*) = \eta_\delta^2$ (i.e., $\eta(X^*) = \eta_\delta$).*

Proof. (i) follows from Remark 3.10. The remaining assertions are consequences of the main results on DCA (see section 2) after simple calculations (related to the conjugate function, the polar, and the subdifferential of the two seminorms η, ξ), the fact that the moduli of strong convexity $\rho(\frac{1}{2}\eta^2, \mathcal{A}^\perp)$ and $\rho(\frac{1}{2}(\eta^0)^2, \mathcal{A}^\perp)$ are positive, and Proposition 3.2. ∎

3.2.5. The description of DCA for solving (3.12) and (3.16). Let us now describe the DCA applied to Problems (3.12) and (3.16).

DCA2 (DCA applied to (3.12)). Generate two sequences $\{X^{(k)}\}$ and $\{Y^{(k)}\}$ in $\mathcal{A}^\perp$ as follows:

Let $\tau_1 > 0$, $\tau_2 > 0$, and $0 \neq X^{(0)} \in \mathcal{A}^\perp \cap C$ be given.

For $k = 0, 1, \ldots$ until

$$\text{either } \|X^{(k+1)} - X^{(k)}\| \leq \tau_1 \|X^{(k+1)}\|$$
$$\text{or } |\xi(X^{(k)}) - \xi(X^{(k+1)})| \leq \tau_2(\xi(X^{(k+1)}) + 1)$$

take

$$Y^{(k)} = B(X^{(k)})X^{(k)}, \tag{3.36}$$

$$X^{(k+1)} = \frac{\eta_\delta V^+ Y^{(k)}}{\langle Y^{(k)}, V^+ Y^{(k)}\rangle^{1/2}} = \frac{\eta_\delta V^+ Y^{(k)}}{\eta(V^+ Y^{(k)})}. \tag{3.37}$$

DCA2BIS (DCA applied to (3.16)). Replace (3.37) in DCA2 with

$$X^{(k+1)} = \frac{Y^{(k)}}{\sqrt{nc}\|Y^{(k)}\|}. \tag{3.38}$$

Like Proposition 3.11, we have the following convergence result for DCA2.

PROPOSITION 3.12. *The sequences $\{X^{(k)}\}$ and $\{Y^{(k)}\}$ generated by DCA2 satisfy the following properties:*

(i) *$\eta(X^{(k)}) = \eta_\delta$ for all k and $\eta^0(Y^{(k)}) \leq \eta_\delta$; i.e., the sequences $\{X^{(k)}\}$ and $\{Y^{(k)}\}$ are bounded.*

(ii) *The sequences $\{\xi(X^{(k)})\}$ and $\{\eta^0(Y^{(k)})\}$ are increasing:*

$$\xi(X^{(k)}) \leq \eta_\delta \eta^0(Y^{(k)}) \leq \xi(X^{(k+1)}) \qquad \forall k.$$

(iii) *(Finite convergence of DCA2.) $\xi(X^{(k+1)}) = \xi(X^{(k)})$ if and only if $X^{(k+1)} = X^{(k)}$ (or equivalently, $Y^{(k+1)} = Y^{(k)}$). In such a case DCA1 stops at the kth iteration with*

$$X^{(k)} = \frac{\eta_\delta V^+ Y^{(k)}}{\eta(V^+ Y^{(k)})}, \qquad Y^{(k)} = \frac{\xi(X^{(k)})}{\eta_\delta^2} V X^{(k)},$$

and $X^{(k)}$ solves Problem (3.18) if and only if $\xi(X^{(k)}) = \eta_\delta^2$ (i.e., $\eta^0(Y^{(k)}) = \eta_\delta$).

(iv) *(Infinite convergence of DCA2.) If the sequences $\{X^{(k)}\}$ and $\{Y^{(k)}\}$ are infinite, then for every limit point (X^*, Y^*) of $\{X^{(k)}, Y^{(k)}\}$ the following hold:*

$$Y^* = \frac{\xi(X^*)}{\eta_\delta^2} V X^* \in \partial\xi(X^*), \quad \eta(X^*) = \eta_\delta \eta^0(Y^*) \leq \eta_\delta \xi(X^*) = \eta_\delta \eta^0(Y^*).$$

Moreover, such an X^* solves Problem (3.18) if and only if $\xi(X^*) = \eta_\delta^2$ (i.e., $\eta^0(Y^*) = \eta_\delta$).

Proof. Assertions (i) and (ii) follow from the main results on DCA for general d.c. programs (section 2) and Proposition 3.2. The remaining assertions can be proved in the same way. Let us demonstrate (iii). According to the results collected in

section 2, $\xi(X^{(k+1)}) = \xi(X^{(k)})$ implies $Y^{(k)} \in \partial\chi_C\,(X^{(k)})$, i.e., $Y^{(k)} = \lambda_k V X^{(k)}$. But $Y^{(k)} \in \partial\xi(X^{(k)})$, so

$$\xi(X^{(k)}) = \langle Y^{(k)}, X^{(k)}\rangle = \lambda_k \langle X^{(k)}, VX^{(k)}\rangle = \lambda_k \eta^2(X^{(k)}) = \lambda_k \eta_\delta^2.$$

It follows that

$$Y^{(k)} = \frac{\xi(X^{(k)})}{\eta_\delta^2} V X^{(k)} \quad \text{and} \quad X^{(k)} = \frac{\eta_\delta V^+ Y^{(k)}}{\eta(V^+ Y^{(k)})} = X^{(k+1)}.$$

The converse is obvious, and the proof is completed. □

REMARK 3.13. (i) *Computing $V^+Y^{(k)}$ in DCA*1 *and/or DCA*2 *amounts to solving the (symmetric positive definite) linear system*

$$\left(V + \frac{1}{n} e e^T\right) X = Y^{(k)}, \tag{3.39}$$

for which the Cholesky factorization seems to be one of the efficient methods.

(ii) *The calculation of $X^{(k+1)}$ in DCA*1*bis and DCA*2*bis requires only matrix-vector products.*

(iii) *In DCA*1*bis we have $\rho(\frac{1}{2}\eta^2, \mathcal{A}^\perp) = nc$ and $\rho((\frac{1}{2}\eta^2)^*, \mathcal{A}^\perp) = \frac{1}{nc}$ (the normal case).*

3.2.6. DCA for solving the proximal regularized d.c. program of (3.6): DCA1r. As indicated in Remark 2.2, it is worth introducing the proximal regularized d.c. program of (EDP_1) (with ρ being a nonnegative number, called the regularization parameter):

$$\min\left\{F_1(X) := \left[\frac{\rho}{2}\|X\|^2 + \frac{1}{2}\eta^2(X)\right] - \left[\frac{\rho}{2}\|X\|^2 + \xi(X)\right] : \quad X \in \mathcal{M}_{n,p}(\mathbb{R})\right\}, \tag{3.40}$$

where the d.c. decomposition of $F(X)$ yields

$$G(X) := \frac{\rho}{2}\|X\|^2 + \frac{1}{2}\eta^2(X), \qquad H(X) := \frac{\rho}{2}\|X\|^2 + \xi(X). \tag{3.41}$$

The original d.c. program (EDP_1) is a special case of (3.40) with $\rho = 0$.

The DCA applied to (3.40) differs from the DCA applied to (EDP_1) by the following two facts: the symmetric positive semidefinite matrices $B(X)$ and V are replaced by $\rho I + B(X)$ and $\rho I + V$, respectively. More precisely, the regularized version of DCA1 can be described as follows.

DCA1R (the DCA applied to the proximal regularized d.c. program (3.40)).

Let $\tau_1 > 0$, $\tau_2 > 0$, and $0 \neq X^{(0)} \in \mathcal{A}^\perp$ be given.

For $k = 0, 1, \ldots$ until

$$\text{either } \|X^{(k+1)} - X^{(k)}\| \le \tau_1 \|X^{(k+1)}\|$$
$$\text{or } |F_1(X^{(k)}) - F_1(X^{(k+1)})| \le \tau_2(|F_1(X^{(k+1)})| + 1)$$

take $Y^{(k)} = B(X^{(k)})X^{(k)} + \rho X^{(k)}$ and solve

$$(V + \rho I)X = Y^{(k)} \tag{3.42}$$

to obtain $X^{(k+1)}$.

REMARK 3.14. *The other visible advantage of DCA*1*r concerns the computation of the pseudoinverse* V^+ *of* V*: for computing* V^+*, we have to apply the Cholesky factorization to the matrix* $V + \frac{1}{n}ee^T$*, which destroys the sparsity structure of* V*, while the sparse Cholesky factorization can be advantageously applied to the symmetric positive matrix* $\rho I + V$*, which preserves the sparsity structure of* V*. In our experiments DCA*1*r seems to be more robust and efficient than DCA*1 *(see section* 6*).*

3.2.7. DCA for solving the proximal regularized d.c. program of (3.12): DCA2r. As for problem (3.13), we introduce the proximal regularization technique into problem (3.12) in order to obtain robustness and stability in numerical computations. Here we will not use the Hilbertian kernel $\frac{\rho}{2}\|.\|^2$ but the quadratic function $\frac{\rho}{2}\eta^2$ (which is positive definite on $\mathcal{A}^\perp$) because we have explicit calculations for the latter. The regularized d.c. program of problem (3.12) is thus its equivalent d.c. program:

$$\min\left\{\left[\frac{\rho}{2}\eta^2(X) + \chi_{\eta_\delta U(\eta)}(X)\right] - \left[\xi(X) + \frac{\rho}{2}\eta^2(X)\right] : X \in \mathcal{M}_{n,p}(\mathbb{R})\right\}. \tag{3.43}$$

The DCA applied to problem (3.43) computes $Y^{(k)} = B(X^{(k)})X^{(k)} + \rho V X^{(k)}$ and $X^{(k+1)} \in \mathcal{A}^\perp$, which is the unique solution of the convex program

$$\min\left\{\frac{\rho}{2}\eta^2(X) - \langle X, Y^{(k)}\rangle : X \in \mathcal{A}^\perp, \eta(X) \le \eta_\delta\right\}. \tag{3.44}$$

LEMMA 3.15. *Let* $\overline{Y} \in \mathcal{A}^\perp$ *be fixed. The unique solution* $\overline{X}$ *to the convex program*

$$\min\left\{\frac{\rho}{2}\eta^2(X) - \langle X, \overline{Y}\rangle : X \in \mathcal{A}^\perp, \eta(X) \le \eta_\delta\right\} \tag{3.45}$$

is given by

$$\overline{X} := \frac{1}{\rho}V^+Y \quad \textit{if} \quad \eta(V^+\overline{Y}) \le \rho\eta_\delta, \qquad \frac{\eta_\delta}{\eta(V^+\overline{Y})}V^+\overline{Y} \quad \textit{otherwise.} \tag{3.46}$$

Moreover, if $\overline{Y} = B(X)X + \rho V X$ *with* $X \in \mathcal{A}^\perp, \rho^2\eta^2(X) + 2\rho\xi(X) + \eta^2(V^+B(X)X) \ge \rho^2\eta_\delta^2$*, then the unique solution* $\overline{X}$ *to problem* (3.45) *is simply* $\overline{X} = \frac{\eta_\delta}{\eta(V^+\overline{Y})}V^+\overline{Y}$*.*

Proof. Since the quadratic function η^2 is positive definite on $\mathcal{A}^\perp$, problem (3.45) has a unique solution $\overline{X} \in \mathcal{A}^\perp$ defined by ($\overline{\lambda}$ being a nonnegative number)

$$\eta(\overline{X}) \le \eta_\delta, \quad \rho V\overline{X} - \overline{Y} = -\overline{\lambda}V\overline{X}, \quad \overline{\lambda}(\eta(\overline{X}) - \eta_\delta) = 0.$$

It follows that $\overline{X} = \frac{1}{\lambda+\rho}V^+\overline{Y}$ and $\eta^2(\overline{X}) = \frac{1}{(\lambda+\rho)^2}\,\eta^2(V^+\overline{Y})$. According to the first and third conditions, we get

$$\overline{\lambda} = 0 \text{ if } \eta(V^+\overline{Y}) \le \rho\eta_\delta, \qquad \frac{\eta(V^+\overline{Y})}{\eta_\delta} - \rho \text{ otherwise.}$$

The formulation (3.46) then is immediate. It remains to prove that if $\overline{Y} = B(X)X + \rho V X$ with $X \in \mathcal{A}^\perp$ given as above, then $\eta(V^+\overline{Y}) \ge \rho\eta_\delta$. Indeed we have

$$\begin{aligned}\eta^2(V^+\overline{Y}) &= \langle VV^+\overline{Y}, V^+\overline{Y}\rangle = \langle \overline{Y}, V^+\overline{Y}\rangle = \langle B(X)X + \rho VX, V^+B(X)X + \rho X\rangle \\ &= \rho^2\langle VX, X\rangle + 2\rho\langle X, B(X)X\rangle + \langle B(X)X, V^+B(X)X\rangle \\ &= \rho^2\eta^2(X) + 2\rho\xi(X) + \eta^2(V^+B(X)X) \ge \rho^2\eta_\delta^2.\end{aligned}$$

The DCA for solving the proximal regularized d.c. program (3.43) is then given by the following.

DCA2R (the DCA applied to the proximal regularized d.c. program (3.43)).

Let $\tau_1 > 0$, $\tau_2 > 0$ and $X^{(0)} \in \mathcal{A}^\perp$, $\eta(X^{(0)}) = \eta_\delta$ be given.

For $k = 0, 1, \ldots$ until

$$\begin{aligned} &\text{either } \|X^{(k+1)} - X^{(k)}\| \le \tau_1 \|X^{(k+1)}\| \\ &\text{or } |\xi(X^{(k)}) - \xi(X^{(k+1)})| \le \tau_2(\xi(X^{(k+1)}) + 1) \end{aligned}$$

take $Y^{(k)} = B(X^{(k)})X^{(k)} + \rho V X^{(k)}$, solve

$$\left(V + \frac{1}{n} e e^T\right) X = Y^{(k)} \tag{3.47}$$

to obtain $V^+ Y^{(k)}$, and set $X^{(k+1)} = \frac{\eta_\delta}{\eta(V^+ Y^{(k)})} V^+ Y^{(k)}$.

REMARK 3.16. *Regarding Proposition* 3.12, *the above regularization implies the following property for DCA*2*: since the moduli of strong convexity* $\rho(\frac{1}{2}\eta^2, \mathcal{A}^\perp)$ *and* $\rho(\frac{1}{2}(\eta^0)^2, \mathcal{A}^\perp)$ *are positive, the two series* $\{\|X^{(k+1)} - X^{(k)}\|^2\}$ *and* $\{\|Y^{(k+1)} - Y^{(k)}\|^2\}$ *converge.*

4. The two-phase algorithm EDCA. In order to obtain a good starting point for DCA applied to the main problem we investigate two techniques. In the first one we compute the complete *dissimilarity matrix* and then apply DCA to solve the new *dissimilarity geometry problem*. This procedure is called *Phase* 1 in our two-phase algorithm named EDCA which is described below.

ALGORITHM EDCA.

Phase 1. *Find an initial point for Phase* 2.

Step 1. *Determine an approximate distance matrix* $\tilde{\Delta} = (\tilde{\delta}_{ij})$.

For $i = 1, \ldots, n, j = i + 1, \ldots, n$, compute $\tilde{\delta}_{ij}$, the length of the shortest path between i and j, within the connected graph $G(N, \mathcal{S})$.

Step 2. *Solve the problem*

$$\min\left\{\frac{1}{2}\sum_{i<j} c(\tilde{\delta}_{ij} - \|X_i^T - X_j^T\|)^2 : X \in \mathcal{M}_{n,p}(\mathbb{R})\right\} \tag{4.1}$$

by applying either DCA1bis to problem (3.6) or DCA2bis to problem (3.12), where w_{ij} and δ_{ij} are replaced by c and $\tilde{\delta}_{ij}$, respectively, to obtain a point denoted by $\tilde{X}$.

Phase 2. *Solve the original problem* (EDP_1) by applying either DCA1 to problem (3.6) or DCA2 to problem (3.12) from the point $\tilde{X}$.

REMARK 4.1. *An approximate distance matrix can be determined in several ways as in the embed algorithm* [5]. *Indeed, by taking*

$$u_{ij} = l_{ij} = \delta_{ij} \quad \textit{for } (i,j) \in \mathcal{S}$$

and using the relationships

$$u_{ij} = \min(u_{ij}, u_{ik} + u_{kj}), \qquad l_{ij} = \max(l_{ij}, l_{ik} - u_{kj}, l_{jk} - u_{ki}\},$$

one obtains a full set of bounds $[l_{ij}, u_{ij}]$, *and then one can take* $\tilde{\delta}_{ij} \in [l_{ij}, u_{ij}]$. *In our algorithm, we attempt to use a simpler procedure for computing the approximate*

matrix $\tilde{\Delta} = (\tilde{\delta}_{ij})$*: the length of the shortest paths (within the connected graph* $G(N, \mathcal{S})$*) between atom* i *and atom* j*. Its direct calculation does not require computing both the bounds* l_{ij} *and* u_{ij} *and so is less expensive. From our experiments we observe that this choice of* $\tilde{\Delta}$ *for DCA is the most efficient, in comparison with the choice* $\tilde{\delta}_{ij} = 0.5(u_{ij} + l_{ij})$ *for* $(i,j) \notin \mathcal{S}$ *and the conditional choice* $\tilde{\delta}_{ij} = 0.5u_{ij} + l_{ij}$ *for* $(i,j) \notin \mathcal{S}$ *if* $l_{ij} \leq 0.5u_{ij}$*.*

There exist in the literature many techniques for choosing a starting point for the distance geometry problem; among them *Algorithm* STRUCT (Moré and Wu [24]) using the *spanning trees procedure* seems to be noteworthy. Our second technique for finding a good starting point for DCA is a modification of Algorithm STRUCT. It provides a point satisfying the *largest* distance constraint in $\mathcal{S}$ and at least $n-2$ different distance constraints that are the *largest* constraint (k, j_k) among the pairs $(k,j) \in \mathcal{S}$ for a given k.

PROCEDURE SP. Let $(i_0, j_0) \in \mathcal{S}$ such that $\delta^0_{ind(i_0,j_0)} = \max\left\{\delta_{ind(i,j)} : (i,j) \in \mathcal{S}\right\}$. Let $x^{i_0} = (0,0,0)^T$ and generate x^{j_0} such that $\|x^{i_0} - x^{j_0}\| = \delta^0_{ind(i,j)}$.
Set $\mathcal{M}$:=$\{i_0, j_0\}, k := j_0$.

do while $|\mathcal{M}| < n$

choose $(k, j_k) \in \mathcal{S}$ such that $\delta_{ind(k,j_k)^0} = \max_j\{\delta^0_{ind(k,j)} : (k,j) \in \mathcal{S}\}$.

generate x^{j_k} such that $\|x^k - x^{j_k}\| = \delta^0_{ind(k,j_k)}$.

Set $\mathcal{M} := \mathcal{M} \cup \{j_k\}, k := j_k$.

end do

This procedure is much less expensive than Phase 1 of the main algorithm EDCA, and as we will see later in numerical experiments, it may provide good starting points for DCA.

5. New d.c. programs for (EDP$_1$): The nonstandard l_∞ (resp., combined $l_1 - l_\infty$) approach. The preceding two d.c. programs (3.13) and (3.12) for (EDP$_1$) involve the l_1-norm in the definition of their objective functions. At least from the computational point of view, as will be shown in the numerical simulations (section 6), it is important to use the l_∞-norm to formulate the following nonstandard d.c. programs for (EDP$_1$). The first reformulation is

$$0 = \min\left\{\Phi(X) := \max_{(i,j)\in\mathcal{S}, i<j}\left\{\Phi_{ij}(X) := \frac{1}{2}w_{ij}[d_{ij}(X) - \delta_{ij}]^2\right\} : X \in \mathcal{M}_{n,p}(\mathbb{R})\right\}.$$

In fact we will tackle the equivalent constrained problem (Lemma 3.3 and Proposition 3.8):

$$0 = \min\{\Phi(X) : X \in \mathcal{C}\}, \tag{5.1}$$

$$\mathcal{C} := \left\{X \in \mathcal{A}^\perp : \sum_{i<j}\|X_i^T - X_j^T\|^2 = n\|X\|^2 \leq r^2\right\} \quad \text{and} \quad r^2 := \sum_{i<j}\widetilde{\delta}_{ij}^2,$$

where $\tilde{\delta}_{ij}$ denotes the length of the shortest paths between nodes i and j (section 4). Note that $\mathcal{C}$ is a compact convex set. Let us now prove that problem (5.1) is actually

a d.c. program. It is clear that

$$\begin{aligned}\Phi_{ij}(X) &= \frac{1}{2}w_{ij}d_{ij}^2(X) + \frac{1}{2}w_{ij}\delta_{ij}^2 - w_{ij}\delta_{ij}d_{ij}(X)\\ &= \frac{1}{2}w_{ij}d_{ij}^2(X) + \frac{1}{2}w_{ij}\delta_{ij}^2 + \sum_{\substack{(k,l)\in\mathcal{S},\,k<l\\(k,l)\neq(i,j)}} w_{kl}\delta_{kl}d_{kl}(X) - \xi(X)\end{aligned}$$

is a d.c. function on $\mathcal{M}_{n,p}(\mathbb{R})$. Hence its finite pointwise supremum Φ is d.c. too, with the d.c. decomposition (see [33])

(5.2)

$$\begin{aligned}\Phi(X) &= \max_{(i,j)\in\mathcal{S},i<j}\left\{\zeta_{ij}(X) := \frac{1}{2}w_{ij}d_{ij}^2(X) + \frac{1}{2}w_{ij}\delta_{ij}^2 + \sum_{\substack{(k,l)\in\mathcal{S},\,k<l\\(k,l)\neq(i,j)}} w_{kl}\delta_{kl}d_{kl}(X) - \xi(X)\right\}\\ &= \zeta(X) - \xi(X).\end{aligned}$$

Hence problem (5.1) can be recast into the following d.c. program:

$$0 = \min\{\zeta(X) - \xi(X) : X \in \mathcal{C}\}. \tag{5.3}$$

In the combined $l_1 - l_\infty$-approach, the distance geometry problem is equivalently stated as the d.c. program

$$\begin{aligned}-\frac{\rho}{2}\eta_\delta^2 &= \min\left\{F_2(X) := [\zeta(X) - \xi(X)] + \rho\left[\frac{1}{2}\eta^2(X) - \xi(X)\right] : X \in \mathcal{C}\right\}\\ &= \min\left\{F_2(X) = \left[\zeta(X) + \frac{\rho}{2}\eta^2(X)\right] - (1+\rho)\xi(X) : X \in \mathcal{C}\right\}. \qquad (5.4)\end{aligned}$$

The positive constant ρ is to be chosen according to the problems under consideration. It is clear that problems (5.3) and (5.4) are actually nonsmooth d.c. programs; i.e., they cannot be transformed into equivalently smooth nonconvex programs. The practical usefulness of these reformulations resides in the fact that DCA applied to problem (5.4) may better approximate global solutions than DCA applied to the standard problems (3.12), (3.13).

REMARK 5.1. *Problems* (3.12), (3.13), (5.3), *and* (5.4) *have the same (global) solution set. But the local optimality condition* (2.5) *used for constructing DCA is not the same for the first two problems as for the last two. This fact is crucial for DCA since we could restart DCA applied to problem* (5.4) *from an initial point computed by DCA applied to problem* (3.13). *In this way, local solutions computed by DCA could be improved (see the computational results in section* 6*).*

To perform the DCA applied to the $l_1 - l_\infty$ d.c. program (5.4), we have only to calculate the subdifferential of the convex functions ζ and $(\zeta + \frac{\rho}{2}\eta^2 + \chi_{\mathcal{C}})^*$.

5.1. Calculation of $\partial\zeta$. We recall that

$$\zeta(X) := \max_{(i,j)\in\mathcal{S},i<j}\left\{\zeta_{ij}(X) := \frac{1}{2}w_{ij}d_{ij}^2(X) + \frac{1}{2}w_{ij}\delta_{ij}^2 + \sum_{\substack{(k,l)\in\mathcal{S},\,k<l\\(k,l)\neq(i,j)}} w_{kl}\delta_{kl}d_{kl}(X)\right\}.$$

Let $\mathcal{S}_\zeta(X) := \{(i,j) \in \mathcal{S}, i < j : \zeta_{ij}(X) = \zeta(X)\}$. According to subsection 3.2, it is simpler to compute $\mathcal{S}_\zeta(X)$ with the following formulation:

$$\mathcal{S}_\zeta(X) = \{(i,j) \in \mathcal{S},\ i < j : \Phi_{ij}(X) = \Phi(X)\}. \tag{5.5}$$

By using usual rules for subdifferential calculus, we have

$$\partial\zeta(X) = co\{\cup\partial\zeta_{ij}(X) : (i,j) \in \mathcal{S}_\zeta(X)\},$$

where *co* stands for the convex hull.

According to the computation of $\partial\xi$ in section 3.2.1 and Remark 3.6,

(1) range $\partial\zeta \subset \mathcal{A}^\perp$;

(2) we can choose the particular subgradient of ζ:

$$B(X)X + w_{ij}[1 - \delta_{ij}s_{ij}(X)]M_{ij}X = B_{ij}(X)X \in \partial\zeta(X) \quad \text{for } (i,j) \in \mathcal{S}_\zeta(X). \tag{5.6}$$

REMARK 5.2. *It follows from the definition of the matrix M_{ij} in section* 3.2.1 *that*

(i) *the symmetric matrices $B(X)$ and $B_{ij}(X)$, serving to calculate subgradients of the convex functions ξ and ζ, respectively ($B(X)X \in \partial\xi(X)$ and $B_{ij}(X)X \in \partial\zeta(X)$ for $(i,j) \in \mathcal{S}_\zeta(X)$), differ from each other at four entries:*

$$\begin{aligned}
[B_{ij}(X)]_{ii} &= [B(X)]_{ii} + w_{ij}[1 - \delta_{ij}s_{ij}(X)],\\
[B_{ij}(X)]_{jj} &= [B(X)]_{jj} + w_{ij}[1 - \delta_{ij}s_{ij}(X)],\\
[B_{ij}(X)]_{ij} &= [B_{ij}(X)]_{ji} = [B(X)]_{ij} - w_{ij}[1 - \delta_{ij}s_{ij}(X)];
\end{aligned}$$

(ii) *X^* is an optimal solution to (EDP_1) if and only if $B(X^*) = B_{ij}(X^*)$ for all $(i,j) \in \mathcal{S}$.*

5.2. Calculation of $\partial(\zeta + \frac{\rho}{2}\eta^2 + \chi_{\mathcal{C}})^*$. Since the convex function η^2 is strongly convex on $\mathcal{A}^\perp$ (Remark 3.10), the function $(\zeta + \frac{\rho}{2}\eta^2 + \chi_{\mathcal{C}})^*$ is differentiable on $\mathcal{M}_{n,p}(\mathbb{R})$. But unlike the preceding convex functions, it seems that the gradient $\nabla(\zeta + \frac{\rho}{2}\eta^2 + \chi_{\mathcal{C}})^*(Y)$, which is the unique solution of the convex program

$$\min\left\{\zeta(X) + \frac{\rho}{2}\eta^2(X) - \langle X, Y\rangle : X \in \mathcal{A}^\perp,\ n\|X\|^2 \le r^2\right\}, \tag{5.7}$$

cannot be explicitly calculated. On the other hand, since the projection on the convex set $\mathcal{C} := \{X \in \mathcal{A}^\perp : \sum_{i<j} \|X_i^T - X_j^T\|^2 = n\|X\|^2 \le r^2\}$ is explicit, for solving problem (5.7) we suggest the use of the subgradient projection method [6], [34], which we succinctly describe below.

Subgradient projection algorithm for solving problem (5.7): *SGPA.* From a given initial point $Z^0 \in \mathcal{C}$, SGPA generates a sequence $\{Z^k\}$ in $\mathcal{C}$ as follows. Assume that Z^k has already been calculated. Calculate the particular subgradient G^k of the objective function at $Z^k : G^k := \rho V Z^k + B_{ij}(Z^k)Z^k - Y$ with $(i,j) \in \mathcal{S}_\zeta(Z^k)$. If $G^k = 0$, then Z^k is an optimal solution to problem (5.7). Otherwise, calculate the next iteration

$$Z^{k+1} := P_{\mathcal{C}}\left(Z^k - \lambda_k \frac{G^k}{\|G^k\|}\right),$$

where the sequence of positive numbers$\{\lambda_k\}$ is chosen such that

$$\lambda_k \to 0 \ \text{ as } k \to +\infty \quad \text{and} \quad \sum_{k=1}^{+\infty} \lambda_k = +\infty,$$

and $P_{\mathcal{C}}$ stands for the orthogonal projection onto $\mathcal{C}$. As said above, the projection $P_{\mathcal{C}}$ is explicit, since we have for $A \in \mathcal{A}^{\perp}$: $P_{\mathcal{C}}(A) = A$ if $A \in \mathcal{C}$, and $\frac{r}{\sqrt{n}}\frac{A}{\|A\|}$ otherwise. Z^k, G^k are in $\mathcal{A}^{\perp}$. It has been shown [6], [34] that the sequence $\{Z^k\}$ converges to the unique optimal solution to the convex program (5.7).

5.3. DCA for solving the $l_1 - l_\infty$ d.c. program (5.4): DCA3. The results displayed in section 4 enable us to outline the DCA for solving the $l_1 - l_\infty$ d.c. program (5.4)

DCA3 (DCA applied to (5.4)). Generate two sequences $\{X^{(k)}\} \subset \mathcal{C}$ and $\{Y^{(k)}\} \subset \mathcal{A}^{\perp}$ as follows:

Let $\tau_1 > 0, \tau_2 > 0$, and $0 \neq X^{(0)} \in \mathcal{C}$ be given.

For $k = 0, 1, \ldots$ until

$$\text{either } \|X^{(k+1)} - X^{(k)}\| \leq \tau_1 \|X^{(k+1)}\|$$
$$\text{or } |F_2(X^{(k)}) - F_2(X^{(k+1)})| \leq \tau_2(|F_2(X^{(k+1)})| + 1)$$

take

$$Y^{(k)} = (1+\rho)B(X^{(k)})X^{(k)} \in (1+\rho)\partial\xi(X^{(k)})$$

and compute the sequence $\{X^{(k,l)} : l \geq 0\}$ generated by SGPA for solving the convex program (starting with $(X^{(k,0)} := X^{(k)})$

$$\min\left\{\zeta(X) + \frac{\rho}{2}\eta^2(X) - \langle X, Y^{(k)}\rangle : X \in \mathcal{C}\right\} \tag{5.8}$$

to obtain $X^{(k+1)}$ as the unique optimal solution to (5.8): $X^{(k+1)} := \lim_{l\to+\infty} X^{(k,l)}$.

REMARK 5.3. *Like DCA*1*, the two sequences* $\{X^{(k)}\}$ *and* $\{Y^{(k)}\}$ *generated by DCA*3 *are bounded, and the general convergence result for the DCA (section* 2.1*) is strengthened by the strong convexity of* $\zeta + \frac{\rho}{2}\eta^2$ *on* $\mathcal{C}$*.*

6. Computational experiments. Our algorithms are coded in FORTRAN 77 with double precision and run on an SGI Origin 2000 multiprocessor with an IRIX system. We have tested our code on three sets of data: the first one is the artificial data from Moré and Wu [22], the second is derived from proteins in the PDB [2], and the third is generated by Hendrickson [12], [13].

The purpose of the experiments is threefold. The first is to show that the DCA can efficiently solve large-scale distance geometry problems (EDP_1). We consider molecules containing at most 4096 atoms (12288 variables) in the artificial data and at most 4189 atoms in the PDB data.

The second is to study the effect of starting points for the DCA applied to the main problem (EDP_1). We compare the efficiency of the two-phase algorithm EDCA and Algorithm SDCA (below), i.e., DCA applied to (EDP_1) with and/or without Phase 1.

The third goal is to exploit the effect of the d.c. decomposition on the solution of (EDP_1) by the DCA via the regularization technique and the Lagrangian duality.

For these purposes, we have tested the following variants of our methods:

- EDCA1: the two-phase algorithm EDCA, which uses DCA1bis and DCA1 in Phase 1 and Phase 2, respectively,
- EDCA2: the two-phase algorithm EDCA, which uses DCA2bis and DCA2 in Phase 1 and Phase 2, respectively,
- SDCA: DCA1 with Procedure SP for computing a starting point,

- RSDCA: DCA1r with Procedure SP for computing a starting point (the regularized version of SDCA),
- REDCA1: the two-phase algorithm EDCA, which uses DCA1bis and DCA1r in Phase 1 and Phase 2, respectively (the regularized version of EDCA1),
- EDCA1-3: a variant of EDCA1, which uses a combination of DCA1 and DCA3 in Phase 2: in EDCA1, we perform only enough iterations of DCA1 to get a sufficiently good guess $X^{(k)}$ (we stop DCA1 with a quite large tolerance τ_2) and then apply DCA3 to terminate Phase 2.

We consider $w_{ij} = 1$ for all $i \neq j$ in Phase 1, and $w_{ij} = 1$ for $(i,j) \in \mathcal{S}$, $i \neq j$, in Phase 2.

For starting DCA1bis, we use Procedure SP to compute $X^{(0)}$, and then set $X^{(0)} := P_{\mathcal{A}^\perp}(X^{(0)})$. The initial point of DCA2bis is then set to $\frac{\tilde{\eta}_\delta X^{(0)}}{(\sqrt{n}\|X^{(0)}\|)}$. We take $\tau_1 = 10^{-8}$ and $\tau_2 = 10^{-9}$ in all algorithms (except for DCA1 in the combined EDCA1-3, where we choose $\tau_2 = 10^{-3}$).

For solving the linear system (3.39) (resp., (3.42)) in Phase 2, we first decompose the matrix $V + \frac{1}{n}ee^T = R^T R$ (resp., $V + \rho I = R^T R$) by the Cholesky factorization, and then at each iteration we solve two systems $R^T U = Y^{(k)}$ and $RX = U$.

In the tables presented below we indicate the following values:

- data: the number of given distances, i.e., $(1/2)|\mathcal{S}|$, where $|\mathcal{S}|$ is the cardinality of $\mathcal{S}$.
- t0: CPU time of Procedure SP and the completion of the matrix $\tilde{\Delta}$ in the two-phase algorithm EDCA, and/or CPU time of Procedure SP in Algorithm SDCA.
- it1 and time1: the number of iterations and CPU time of DCA1bis and/or DCA2bis, respectively.
- it2 and time2: the number of iteration and CPU time of DCA1 and/or DCA2, respectively.
- ttotal: the total CPU time of the algorithm.
- aver: the average relative error defined by $\frac{1}{|\mathcal{S}|}\sum_{(i,j)\in\mathcal{S}} \frac{|\delta_{ij} - \|X_i^{*T} - X_j^{*T}\||}{\delta_{ij}}$.
- maxer: the maximum relative error defined by

$$\max\left\{\frac{|\delta_{ij} - \|X_i^{*T} - X_j^{*T}\||}{\delta_{ij}} : (i,j) \in \mathcal{S}\right\}.$$

Note that all CPU times are computed in seconds.

6.1. The data.

6.1.1. The artificial data. We consider two models of problems given in Moré and Wu [22], where the molecule has $n = s^3$ atoms located in the three-dimensional lattice

$$\{(i_1, i_2, i_3) : 0 \leq i_1 < s, 0 \leq i_2 < s, 0 \leq i_3 < s\}$$

for some integer $s \geq 1$.

In the first model problem the ordering for the atoms is specified by letting i be the atom at the position (i_1, i_2, i_3),

$$i = 1 + i_1 + si_2 + s^2 i_3,$$

TABLE 6.1
Summarized information about test problems from the PDB.

ID code	Exp. method	Classification	Atoms (n)	Residues
1A1D	NMR (MAS)	Nucleotidyltransferase	146	146
304D	X-ray diffraction	Deoxyribonucleic acid	237	52
8DRH	NMR (MAS)	Deoxyribonucleic acid/Ribonucleic acid	329	16
1AMD	NMR (MAS)	Deoxyribonucleic acid	380	12
2MSJ	X-ray diffraction	Antifreeze protein	480	66
124D	NMR	Deoxyribonucleic/Ribonucleic acid	509	16
141D	NMR	Deoxyribonucleic acid	527	26
132D	NMR	Deoxyribonucleic acid	750	24
1A84	NMR	Deoxyribonucleic acid	758	24
104D	NMR	DNA/RNA chimeric hybrid duplex	766	24
103D	NMR (MAS)	Deoxyribonucleic acid	772	24
2EQL	X-ray diffraction	Hydrolase (O-Glycosyl)	1023	129
1QS5	X-ray diffraction	Hydrolase	1429	162
1QSB	X-ray diffraction	Hydrolase	1431	162
1ITH	X-ray diffraction	Oxygen transport	2366	282
2CLJ	Theoretical model	Hydrolase	4189	543

and distance data are generated for all pairs of atoms in

$$\mathcal{S} = \{(i,j) : |i-j| \le r\}, \tag{6.1}$$

where r is an integer between 1 and n.

In the second model problem the set $\mathcal{S}$ is specified by $(X_i^T = (i_1, i_2, i_3))$

$$\mathcal{S} = \{(i,j) : \|X_i^T - X_j^T\| \le \sqrt{r}\}. \tag{6.2}$$

As indicated in Moré and Wu [22], a difference between these definitions of $\mathcal{S}$ is that (6.2) includes *all nearby atoms*, while (6.1) includes *some of nearby atoms and some relatively distant atoms*. Thus these model problems may capture various features in distance data from applications.

6.1.2. The PDB data. We consider 16 problems whose data are derived from 16 structures of proteins given in the PDB. Table 6.1 gives the summarized information about these structures (in this table, "Exp." is the abbreviation of "exploitation," and "MAS" is that of "minimized average structure").

For each structure we generate a set of distances by using all distances between the atoms in the same residue as well as those in the neighboring residues. More precisely, if $\mathcal{R}_k$ is the kth residue, then

$$\mathcal{S} = \{(i,j) : x_i \in \mathcal{R}_k, : x_j \in \mathcal{R}_k \cup \mathcal{R}_{k+1}\}$$

specifies the set of distances.

6.1.3. Hendrickson's benchmark problems. This set of data is composed of significantly more difficult test problems. We consider the twelve problems generated by Hendrickson [12], [13] from the bovine pancreatic ribonuclease by using fragments consisting of the first 20, 40, 60, 80, and 100 amino acids as well as the full protein (124 amino acids), with two sets of distance constraints for each size corresponding to the largest unique subgraphs and the reduced graphs. These problems have from 63 up to 777 atoms. The protein actually has 1849 atoms, but some simple structure exploitation allowed the author to start the numerical method with only 777 atoms.

TABLE 6.2
*The performance of REDCA*1 *for the second model problem,* $r = 1$, $r = 2$, *and* $r = s^2$.

n	r	data	t0	iter1	time1	iter2	time2	ttotal	aver	maxer
64	1	144	0.00	55	0.07	23	0.02	0.09	1.24E-8	8.66E-5
	2	360	0.01	209	0.26	20	0.02	0.29	2.23E-8	9.90E-5
	s^2	1880	0.00	68	0.09	21	0.05	0.14	2.24E-8	8.84E-5
125	1	300	0.03	74	0.37	59	0.13	0.52	8.12E-8	9.44E-5
	2	780	0.05	79	0.38	24	0.07	0.50	2.21E-7	9.14E-5
	s^2	7192	0.03	71	0.34	26	0.24	0.61	2.13E-8	8.57E-6
216	1	540	0.17	75	1.07	80	0.47	1.71	3.02E-8	9.60E-5
	2	1440	0.22	77	1.10	25	0.19	1.52	1.22E-8	9.70E-5
	s^2	21672	0.15	64	0.92	27	0.77	1.84	1.02E-8	9.01E-5
343	1	882	0.66	69	2.61	109	1.56	4.83	2.36E-8	9.96E-5
	2	2394	0.80	71	2.69	32	0.60	4.09	2.23E-8	9.11E-5
	s^2	53 799	0.64	90	3.38	27	2.00	6.02	1.02E-8	8.95E-5
512	1	1344	2.75	65	6.63	145	6.43	15.81	2.12E-8	9.98E-5
	2	3696	3.10	68	6.94	32	1.79	11.84	1.23E-8	9.97E-5
	s^2	119692	2.14	81	8.10	33	6.22	16.46	4.52E-8	9.07E-5
729	1	1944	11.40	78	20.26	184	28.10	59.76	2.23E-8	9.90E-5
	2	5400	12.22	73	18.96	35	6.35	37.54	1.45E-8	9.02E-5
	s^2	243858	6.59	88	22.38	33	15.69	44.66	1.23E-7	9.14E-5
1000	1	2700	35.33	72	40.55	230	89.05	164.93	3.24E-7	9.90E-5
	2	7560	37.00	79	44.53	37	16.83	98.36	1.23E-8	9.19E-5
	s^2	456872	17.55	89	48.38	1261	1197.99	1263.92	2.35E-6	2.21E-4
1331	1	3630	111.08	78	95.62	272	405.56	611.91	1.23E-7	8.92E-5
	2	10230	112.95	81	99.14	40	61.48	271.52	1.24E-7	9.98E-5
	s^2	809763	58.52	83	115.89	28	72.19	246.60	1.28E-6	8.90E-4
1728	1	4752	368.25	75	215.89	326	860.11	1444.25	1.89E-7	7.78E-5
	2	13464	349.02	76	209.58	50	135.26	693.85	2.45E-7	9.23E-5
	s^2		133.80	96	228.90	29	130.76	493.46	1.25E-6	8.52E-4
2197	1	6084	719.79	78	351.22	406	1642.10	2713.11	3.22E-7	1.05E-5
	2	17316	726.37	77	348.01	44	208.87	1283.25	1.08E-7	6.72E-5
	s^2	2014666	315.59	85	383.87	60	480.28	1179.74	1.0E-6	1.0E-3
2744	1	7644	1620.07	88	741.54	237	1979.76	4341.37	8.52E-5	1.00E-3
	2	21840	1629.80	77	648.85	25	280.51	2559.16	1.22E-6	1.00E-3
	s^2	3436528	595.18	86	729.34	75	1215.57	2600.75	1.24E-6	1.00E-3
3375	1	9450	3552.69	92	1274.72	270	3738.10	8565.50	5.23E-5	1.00E-3
	2	27090	3570.90	85	1175.36	28	529.58	5275.86	2.34E-5	1.00E-3
	s^2	5196129	1201.28	111	1526.32	5	266.82	2994.42	1.0E-6	4.92E-2
4096	s^2	7640952	5434.57	83	6668.17	3	1282.16	13384.90	1.00E-4	4.85E-2

6.2. Experimental results.

6.2.1. The performance of the two-phase algorithm REDCA1. In this experiment we have tested Algorithm REDCA1 (the regularized version of the two-phase algorithm EDCA1) on the first two sets of data (the second model of the artificial data and the PDB data). To observe the behavior of our method when the number of given distances varies, we consider three different values of r in the artificial data: $r = 1$, $r = 2$, and $r = s^2$ (Table 6.2). The results on the PDB data are reported in Table 6.3. The regularization parameter ρ is set to $\rho = 0.01$.

6.2.2. The performance of Algorithm RSDCA. In the second experiment we study the efficiency of DCA applied to (EDP_1) without Phase 1. In Tables 6.4 and 6.5, respectively, we report our experimental results with Algorithm RSDCA (the regularized version of Algorithm SDCA) on the second model problem of the artificial data with $r = 1$ and $r = 2$ and on the PDB data.

6.2.3. Comparison of SDCA and RSDCA. In the third experiment we are interested in the effect of the regularization technique for DCA. We present in Table 6.6 the results of DCA with and/or without regularization, say RSDCA and SDCA for the second set of data, say PDB data. We consider different values for the regularization parameter: $\rho = 0.0001$, $\rho = 0.001$, $\rho = 0.01$, $\rho = 0.1$, and $\rho = 100$.

TABLE 6.3
*The performance of Algorithm REDCA*1 *on the PDB data.*

ID code	n	data	t0	iter1	time1	iter2	time2	ttotal	aver	maxer
1A1D	146	145	0.0	0	0.00	0	0	0.00	0.00E+0	0.00E+0
304D	237	2319	0.38	251	4.85	175	1.52	6.76	7.00E-4	1.00E-3
8DRH	329	3549	0.91	481	17.36	660	10.65	28.92	1.80E-4	1.00E-2
1AMD	380	6186	1.91	912	40.82	844	19.31	65.51	1.00E-6	1.00E-4
2MSJ	480	715	2.04	110	9.81	918	22.96	34.81	1.50E-4	1.00E-2
124D	509	8307	4.16	401	36.06	492	18.03	58.26	2.00E-6	1.00E-3
141D	527	5615	3.44	544	52.01	498	26.95	88.53	1.00E-5	1.00E-3
132D	750	12094	14.21	262	65.94	182	29.81	106.84	1.00E-4	1.00E-2
1A84	758	12345	19.55	452	122.00	886	118.51	260.07	1.02E-6	1.00E-3
104D	766	12609	15.56	524	135.50	13281	1704.65	1855.71	1.00E-5	1.50E-3
103D	772	12777	18.08	1179	343.39	414	58.12	419.59	2.36E-5	2.02E-3
2EQL	1023	4888	54.60	169	133.42	1665	939.40	1127.41	3.00E-4	3.05E-2
1QS5	1429	6355	154.88	72	104.66	3572	4439.67	4698.99	1.10E-3	5.02E-2
1QSB	1431	6344	165.67	112	166.93	2555	3215.16	3547.75	1.20E-03	7.53E-2
1ITH	2366	6239	957.38	87	458.62	1441	6944.09	8360.08	1.00E-4	8.58E-2
2CLJ	4189	19833	8238.57	119	3056.88	1230	31949.27	43244.71	1.00E-3	1.00E-1

TABLE 6.4
The performance of Algorithm RSDCA in the second model with $r = 1$ *and* $r = 2$.

n	r	iter2	ttotal	aver	maxer
64	1	1658	1.06	1.00E-4	1.00E-3
	2	447	0.42	5.81E-2	1.35E-1
125	1	3500	7.26	2.00E-4	1.50E-3
	2	817	2.29	1.37E-1	7.45E-1
216	1	3500	19.81	3.00E-4	4.90E-3
	2	2149	14.93	1.27E-1	5.94E-1
343	1	3500	47.49	3.00E-4	3.70E-3
	2	1615	25.79	1.38E-1	6.92E-1
512	1	3500	150.00	4.00E-4	6.60E-3
	2	2630	123.19	1.41E-1	7.79E-1
729	1	3499	517.59	5.00E-4	8.00E-3
	2	1875	307.39	1.50E-1	8.18E-1
1000	1	1699	637.01	7.00E-4	9.20E-3
	2	1518	639.93	1.50E-1	8.64E-1
1331	1	1450	1404.90	8.10E-5	7.00E-4
	2	1551	1515.11	1.49E-1	8.56E-1
1728	1	1005	3018.26	7.40E-5	4.99E-4
	2	1781	5489.43	1.47E-1	8.93E-1
2197	1	1105	5241.95	5.00E-4	5.70E-5
	2	2429	12280.41	1.49E-1	9.14E-1
2744	1	945	6613.19	5.90E-5	4.98E-4
	2	3634	29415.09	1.01E-1	1.49E+0

6.2.4. Comparison of two variants EDCA1 and EDCA2. In this experiment we consider two versions of EDCA which correspond to different d.c. decompositions to solve the first model problem, where the parameter r in (6.1) is set to $r = s^2$. This data set is also considered in [41]. We note that when $r = s^2$, the set defined by (6.1) is included in the set defined by (6.2). In Table 6.7 we present the performance of two algorithms EDCA1 and EDCA2.

6.2.5. The performance of EDCA1 and EDCA1-3 for Hendrickson's benchmark problems. In this experiment we are interested in the efficiency of

TABLE 6.5
The performance of Algorithm RSDCA in the PDB data.

ID code	n	iter2	ttotal	aver	maxer
1A1D	146	0	0.00	0.00E+0	0.00E+0
304D	237	701	6.30	1.22E-7	1.00E-4
8DRH	329	560	9.08	2.10E-4	2.00E-1
1AMD	380	631	16.06	1.02E-8	9.89E-5
2MSJ	480	666	16.88	1.24E-7	8.55E-5
124D	509	715	32.53	1.24E-7	5.75E-4
141D	527	637	26.61	3.72E-3	2.28E-1
132D	750	337	46.31	2.50E-3	4.53E-1
1A84	758	4438	565.91	0.00E+0	1.70E-3
104D	766	875	151.46	2.46E-06	1.38E-04
103D	772	743	116.18	2.22E-3	4.73E-1
2EQL	1023	1219	669.25	5.76E-4	1.30E-1
1QS5	1429	1109	1402.66	8.00E-4	1.50E-1
1QSB	1431	1282	1445.91	7.56E-4	1.44E-1
1ITH	2366	2101	9799.02	4.82E-4	8.21E-2
2CLJ	4189	1002	27001.29	6.72E-4	2.03E-1

TABLE 6.6
Comparison between SDCA and RSDCA with different choice of ρ.

Pb	Algorithm	ρ	iter2	ttotal	aver	maxer
304D	SDCA		792	9.06	1.37E-6	2.66E-5
	RDCA	1.0	1404	15.88	2.31E-6	7.89E-5
		0.1	797	9.25	2.19E-6	7.73E-5
		0.01	738	8.61	2.14E-6	4.41E-05
		0.001	798	9.25	2.38E-6	5.50E-5
2MSJ	SDCA		1001	39.87	1.23E-5	7.15E-5
	RSDCA	1.0	902	36.80	1.12E-6	4.97E-5
		0.1	919	42.37	1.29E-6	3.06E-5
		0.01	735	33.49	1.96E-6	4.39E-5
		0.001	688	27.53	3.30E-6	5.60E-5
141D	SDCA		688	37.97	4.01E-3	2.28E-1
	RSDCA	1.0	1176	63.35	4.74E-3	2.27E-1
		0.1	703	51.25	4.03E-3	2.27E-1
		0.01	637	26.62	4.01E-03	2.28E-1
		0.001	691	37.95	4.01E-03	2.28E-1
104D	SDCA		849	147.73	1.23E-6	1.18E-3
	RSDCA	1.0	1540	270.46	1.02E-6	5.94E-5
		0.1	987	175.39	2.06E-6	9.81E-5
		0.01	2614	352.51	9.75E-6	6.62E-4
		0.001	875	151.46	2.46E-06	1.38E-4
1A84	SDCA		780	137.23	8.18E-04	2.90E-1
	RSDCA	1.0	1931	354.95	9.02E-7	6.22E-5
		0.1	1042	173.27	3.38E-06	2.92E-4
		0.01	527	90.42	8.21E-04	2.9E-1
		0.001	590	118.49	8.18E-04	2.90E-1

EDCA1 and the combined EDCA1-3 with the last set of data, Hendrickson's benchmark problems. The results are summarized in Table 6.8.

6.3. Comments. Our main concerns in this paper are both the ability to treat large-scale problems and the cost of algorithms. The numerical results show that our algorithms are quite efficient for all sets of data. Our experiments suggest the following comments.

TABLE 6.7
*The performance of EDCA*1 *and EDCA*2 *for the first model problem,* $r = s^2$.

n	data	Algorithm	t0	iter1	time1	iter2	time2	ttotal	aver	maxer
64	888	EDCA1	0.00	70	0.09	70	0.09	28	1.00E-3	4.84E-2
		EDCA2	0.00	66	0.01	28	0.04	0.05	1.00E-3	4.85E-2
125	2800	EDCA1	0.03	117	0.58	52	0.24	0.85	6.00E-4	4.89E-2
		EDCA2	0.03	112	0.52	52	0.24	0.79	6.00E-4	4.87E-2
216	7110	EDCA1	0.23	90	1.51	87	1.75	3.50	4.00E-4	4.94E-2
		EDCA2	0.23	84	1.32	87	1.75	3.26	4.00E-4	4.96E-2
343	15582	EDCA1	0.54	86	3.66	137	6.56	10.77	3.00E-4	4.97E-2
		EDCA2	0.54	83	3.25	136	6.55	10.34	3.00E-4	4.97E-2
512	30688	EDCA1	3.20	81	7.68	205	20.68	31.55	0.0002	4.98E-2
		EDCA2	3.20	78	7.18	206	20.55	30.93	0.0002	4.97E-2
729	55728	EDCA1	15.48	92	28.61	288	77.18	121.33	1.00E-4	4.98E-2
		EDCA2	15.48	90	28.24	288	77.19	120.91	1.00E-4	4.98E-2
1000	94950	EDCA1	76.20	95	61.05	369	271.33	347.53	1.00E-4	4.99E-2
		EDCA2	76.20	92	59.01	369	271.33	345.49	1.00E-4	4.99E-2
1331	153670	EDCA1	178.85	85	103.68	471	537.68	820.21	1.00E-4	4.99E-2
		EDCA2	178.85	83	100.64	470	537.62	817.11	1.00E-4	4.99E-2
1728	238392	EDCA1	404.15	87	285.34	581	1930.06	2619.55	1.00E-4	4.99E-2
		EDCA2	404.15	84	277.92	432	1929.01	2611.14	1.00E-4	4.99E-2
2197	356928	EDCA1	1073.46	103	563.41	702	4009.28	5646.15	1.00E-4	4.99E-2
		EDCA2	1073.46	99	542.84	703	4012.22	5628.52	1.00E-4	4.99E-2
2744	518518	EDCA1	2745.87	130	1132.52	848	7593.98	11472.37	1.00E-4	4.99E-2
		EDCA2	2745.87	124	1073.21	850	7601.22	11420.30	1.00E-4	4.99E-2

TABLE 6.8
*The performance of EDCA*1 *and EDCA*1-3 *for Hendrickson's problems.*

n	data	EDCA1			EDCA1-3		
		ttotal	aver	maxer	ttotal	aver	maxer
63	236	12.23	9.36E-5	7.85E-4	5.74	1.81E-4	1.58E-3
102	336	17.62	9.46E-5	1.12E-3	11.48	1.34E-4	1.35E-3
174	786	28.19	5.76E-4	2.49E-2	31.62	5.59E-4	2.49E-2
236	957	100.85	2.73E-5	6.25E-4	52.88	8.56E-5	9.05E-4
287	1319	74.59	3.40E-3	1.43E-1	310.24	2.43E-4	3.60E-2
362	1526	316.24	3.24E-4	2.48E-2	111.38	3.0E-4	2.38E-2
377	1719	325.68	5.75E-4	4.32E-2	165.25	7.0E-4	2.92E-2
472	2169	359.95	3.63E-3	1.66E-1	258.22	3.67E-4	5.44E-2
480	2006	747.49	6.50E-4	4.97E-2	287.14	7.02E-4	3.89E-2
599	2532	331.56	3.24E-3	1.77E-1	1746.20	2.32E-3	7.20E-2
695	3283	1067.15	1.35E-2	2.49E-1	9899.00	1.50E-3	8.50E-2
777	3504	619.57	6.48E-4	2.50E-2	620.80	6.48E-4	2.50E-2

6.3.1. About the two-phase algorithm EDCA and its variants. The most important fact is that *in all experiments Algorithm EDCA gives an* ε*-global solution of* (EDP_1). Moreover, since the basic DCA is efficient, EDCA can solve large-scale problems in a short time when $n \leq 1000$ (3000 variables), and in a reasonable time when $1331 \leq n \leq 4189$ (to 12567 variables).

We observe from Tables 6.3 and 6.7 that the rate of convergence of the DCA in Phase 1 does not depend much on the distance data (i.e., the number and the length of given distances between *nearby* or *far away* atoms). In other words, the DCA is quite stable in the normal case (in the artificial data).

On the contrary, the DCA in Phase 2 (DCA1 and DCA2) is quite sensitive to the data. In the first model (where given distances are determined between both nearby and distant atoms), the number of iterations of DCA1 and/or DCA2 is much higher than that in the second model. A simple explanation is that for the given distances *between relatively far away atoms* the approximate distance matrix $\tilde{\Delta}$ does not seem to be "good," and the resulting initial point $\tilde{X}$ (given by Phase 1) is not relatively *near* a solution of (EDP_1). Then DCA1 and DCA2 need more iterations to yield a

solution. In general, the cost of Algorithm EDCA in the first model problem is more important than in the second one (see Tables 6.3 and 6.7).

Consider now the influence of the number of given distances, *data*, in the first experiment. We observe that, with $n \geq 1000$, when the number of given distances is small ($r = 1$ and $r = 2$), the cost for determining $\tilde{\Delta}$ is very important in the two-phase algorithm (21% to 68% of the total cost). However when the number of given distances is large ($r = 2$ and $r = s^2$ in the artificial data), this step is necessary because Phase 1 is important for EDCA to obtain a global solution of (EDP_1) in such a case. The results given in Table 6.2 show that when $n \geq 512$, the more the number of given distances increases, the more t0 decreases, and thus the faster EDCA becomes.

On the other hand, although the sequences $\{X^{(k)}\}$ in DCA1 and DCA2 are not in an explicit form, one iteration of these algorithms (which comprises computing the matrix $B(X^{(k)})$, the product $B(X^{(k)})X^{(k)}$, and the solution of two triangular systems) is not more expensive than that of DCA1bis and DCA2bis, which need only matrix-vector products. This shows that DCA1 and DCA2 exploit well the sparsity of $\mathcal{S}$ (in the determination of matrix $B(X^{(k)})$ and the product $B(X^{(k)})X^{(k)}$).

6.3.2. About the Algorithm SDCA and its regularized version. Algorithm RSDCA is very efficient when the number of constraints is not large. In the artificial data with $r = 1$, RSDCA provided an optimal solution for all test problems with the maximal error $maxer \leq 0.009$ (Table 6.4). In the PDB data it successfully solved 7 of 16 problems with $maxer \leq 0.001$, and the average errors in all test problems are small (smaller than 0.003). Hence, Phase 1 in the two-phase algorithm can be replaced efficiently by Procedure SP when a small subset of distances is known. In any case we see that the objective function decreases very fast during some first iterations of DCA1.

6.3.3. The effect of Phase 1 in the two-phase algorithm EDCA. From experimental results we see that Phase 1 is *important* for EDCA when the number of constraints is large. In other words, for our d.c. approach, the technique using the shortest path in Phase 1 of EDCA seems to be more advantageous than Procedure SP when the number of distance constraints is large. Nevertheless when a small subset of constraints is known, Phase 1 does not seem to be efficient because it is expensive to complete the "distance" matrix, and the resulting complete dissimilarity matrix may not be a good approximation to the complete exact distance matrix.

6.3.4. The effect of the regularization technique. As indicated in section 3.2.6, the regularization technique has a visible advantage. In all test problems (most of them have not been presented here), with an appropriate choice of the regularization parameter ρ, DCA1r is better than DCA1. (In our experiments the best choice of ρ is $\rho \in [0.01, 0.001]$.)

6.3.5. About two variants EDCA1 and EDCA2: The effect of d.c. decomposition. The sequence $\{\|X^{(k+1)} - X^{(k)}\|\}$ in DCA2 (resp., DCA2bis) decreases faster than in DCA1 (resp., DCA1bis). In all problems, the number of iterations of DCA2bis is smaller than that of DCA1bis. In several test problems, EDCA2 is less expensive than EDCA1.

6.3.6. More about the results on PDB data and on Hendrickson's benchmark problems. Not surprisingly, the problems derived from PDB data are more difficult to solve than the artificial problems. For these real-life problems Phase

2 needs many more iterations than for artificial problems, while Phase 1 has the same behavior. On the other hand, in contrast to the artificial data, here the smaller the number of distance constraints, the more efficient EDCA1 is. We note that the average error of the obtained solution is very small in both algorithms REDCA1 and RSDCA.

The twelve problems generated by Hendrickson [13] are the most difficult: the cost of the algorithm is higher than that for the first two sets of test problems. However, our algorithm is still efficient on these problems: the maximal error on distance constraints of the solution given by EDCA1-3 is below, respectively, 0.025, 0.055, and 0.085 in problems 6, 4, and 2. We observe that the solutions obtained by DCA3 are better than those provided by DCA1, but that DCA3 is more expensive than DCA1. Then it is interesting to use the combined algorithm EDCA1-3. We note that this set of test problems has been considered in [41], with exact distances for the first seven problems and with inexact distances, say the general problem (1.2) with $0.01 \leq \varepsilon \leq 0.04$, for the remaining five problems ($n \geq 472$). Here we consider the exact distances for all test problems, and these results indicate that our approach has the potential to locate exact (or nearly exact) solutions.

7. Conclusions and future work. We have presented a nonstandard approach, based on d.c. optimization and DCA, for solving large-scale molecular optimization problems from distance matrices. The main points in this approach are

(1) mathematical modeling of the exact distance geometry problem (1.1) as a d.c. program,
(2) a strategy of choosing an acceptable starting point for applying DCA to the resulting d.c. programs,
(3) exploiting the nice effect of d.c. decompositions for DCA.

Suitable d.c. programs of problem (1.1) are indicated in the matrix framework. These nondifferentiable nonconvex optimization problems paradoxically make it possible to express DCA in its simplest way: except for the predictor-corrector EDCA1-3, which additionally needs to solve a convex program, it actually requires matrix-vector products and only one Cholesky factorization and allows the exploitation of sparsity in the large-scale setting.

To find an acceptable initial point for applying the DCA to the d.c. programs under consideration (Phase 2), we have proposed two strategies. The first consists of completing the missing data and applying DCA to the new problem with complete data (Phase 1), and the second uses a spanning tree procedure to compute acceptable starting points.

Computational experiments say that our method is successful in locating large configurations satisfying given distance constraints: the DCA globally solved distance geometry problems with up to $4189 \times 3 = 12567$ variables.

Several interesting problems for future research arise from our results. First, although the strategy of Phase 1 is quite suitable for DCA to reach global solutions to the distance geometry problem, the running time of Phase 2 could be reduced by improving Phase 1 and by exploiting the sparsity better still. On the other hand, since DCA is a fast decreasing method, some methods such as a smoothing technique may be investigated to replace Phase 1.

Second, it is possible to develop a new d.c. approach to the general distance geometry problem (1.3). The standard optimization problem (EDP), (1.6) is a d.c. program, but the objective d.c. function σ is too complex and not convenient for DCA. Other d.c. formulations of problem (1.3) must be found in order to provide

efficient DCA. Preliminary computational results concerning the new d.c. approach to problem (1.3) seem to be quite promising. We plan to address these issues in future work.

Acknowledgments. The authors would like to express their gratitude to the referee, Professor Bruce Hendrickson, and to the associate editor, Professor Robert B. Schnabel, for their valuable remarks and proposals which led to an improvement of the paper. Thanks also to Zhijun Wu for providing the twelve Hendrickson's benchmark problems.

REFERENCES

[1] A.Y. Alfakih, A. Khandani, and H. Wolkowicz, *Solving Euclidean distance matrix completion problems via semidefinite programming. Computational optimization—A tribute to Olvi Mangasarian, Part* I, Comput. Optim. Appl., 12 (1999), pp. 13–30.

[2] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, W. Weissig, I.N. Shindyalov, and P.E. Bourne, *The Protein Data Bank*, Nucleic Acids Res., 28 (2000), pp. 235–242; available online at http://www.resb.org/pdb/.

[3] L.M. Blumenthal, *Theory and Applications of Distance Geometry*, Oxford University Press, London, 1953.

[4] G.M. Crippen, *Rapid calculation of coordinates from distance measures*, J. Comput. Phys., 26 (1978), pp. 449–452.

[5] G.M. Crippen and T.F. Havel, *Distance Geometry and Molecular Conformation*, John Wiley & Sons, New York, 1988.

[6] V.F. Demyanov and L.V. Vasilev, *Nondifferentiable Optimization*, Optimization Software Publications, New York, 1985.

[7] J. De Leeuw, *Applications of convex analysis to multidimensional scaling*, in Recent Developments in Statistics, J.R. Barra, F. Brodeau, G. Romier, and B. van Cutsem, eds., North–Holland, Amsterdam, 1977, pp. 133–145.

[8] J. De Leeuw, *Convergence of the majorization method for multidimensional scaling*, J. Classification, 5 (1988), pp. 163–180.

[9] W. Glunt, T.L. Hayden, and M. Raydan, *Molecular conformation from distance matrices,* J. Comput. Chem., 14 (1993), pp. 114–120.

[10] L. Guttman, *A general nonmetric technique for finding the smallest coordinate space for a configuration of point*, Psychometrika, 33 (1968), pp. 469–506.

[11] T.F. Havel, *An evaluation of computational strategies for use in the determination of protein structure from distance geometry constraints obtained by nuclear magnetic resonance*, Prog. Biophys. Mol. Biol., 56 (1991), pp. 43–78.

[12] B.A. Hendrickson, *The Molecule Problem: Determining Conformation from Pairwise Distances*, Ph.D. thesis, Cornell University, Ithaca, NY, 1991.

[13] B. Hendrickson, *The molecule problem: Exploiting structure in global optimization*, SIAM J. Optim., 5 (1995), pp. 835–857.

[14] P.J. Laurent, *Approximation et Optimisation*, Hermann, Paris, 1972.

[15] M. Laurent, *Cuts, matrix completions and a graph rigidity*, Math. Programming, 79 (1997), pp. 255–283.

[16] H.A. Le Thi, *Contribution à l'optimisation non convexe et l'optimisation globale: Théorie, Algorithmes et Applications*, Habilitation à Diriger des Recherches, Université de Rouen, Rouen, France, 1997.

[17] H.A. Le Thi and T. Pham Dinh, *Solving a class of linearly constrained indefinite quadratic problems by d.c. algorithms,* J. Global Optim., 11 (1997), pp. 253–285.

[18] H.A. Le Thi, T. Pham Dinh, and L.D. Muu, *Numerical solution for optimization over the efficient set by d.c. optimization algorithm*, Oper. Res. Lett., 19 (1996), pp. 117–128.

[19] H.A. Le Thi and T. Pham Dinh, *The DC programming and DCA revisited with DC models of real world nonconvex optimization problems*, Ann. Oper. Res., to appear.

[20] P. Mahey and T. Pham Dinh, *Partial regularization of the sum of two maximal monotone operators,* M2AN Math. Model. Numer. Anal., 27 (1993), pp. 375–395.

[21] P. Mahey, S. Oualibouch, and T. Pham Dinh, *Proximal decomposition on the graph of a maximal monotone operator*, SIAM J. Optim., 5 (1995), pp. 454–466.

[22] J.J. Moré and Z. Wu, *Global continuation for distance geometry problems*, SIAM J. Optim., 7 (1997), pp. 814–836.

[23] J.J. Moré and Z. Wu, *Issues in Large-Scale Molecular Optimization*, Preprint MCS-P539-1095, Argonne National Laboratory, Argonne, IL, 1996.
[24] J.J. Moré and Z. Wu, *Distance Geometry Optimization for Protein Structures*, Preprint MCS-P628-1296, Argonne National Laboratory, Argonne, IL, 1996.
[25] A.M. Ostrowski, *Solutions of Equations and Systems of Equations*, Academic Press, New York, 1966.
[26] T. Pham Dinh, *Contribution à la théorie de normes et ses applications à l'analyse numérique*, Thèse de Doctorat d'Etat Es Science, Université Joseph Fourier-Grenoble, Grenoble, France, 1981.
[27] T. Pham Dinh, *Convergence of subgradient method for computing the bound norm of matrices*, Linear Algebra Appl., 62 (1984), pp. 163–182.
[28] T. Pham Dinh, *Algorithmes de calcul d'une forme quadratique sur la boule unité de la norme maximum*, Numer. Math., 45 (1985), pp. 377–440.
[29] T. Pham Dinh and S. El Bernoussi, *Algorithms for solving a class of non convex optimization problems. Methods of subgradients,* in Fermat Days 85. Mathematics for Optimization (Tolouse, France, 1985), North-Holland Math. Stud. 129, Elsevier, North-Holland, Amsterdam, 1986, pp. 249–271.
[30] T. Pham Dinh and S. El Bernoussi, *Duality in d.c. (difference of convex functions) optimization. Subgradient methods*, in Trends in Mathematical Optimization, Internat. Schriftenreihe Numer. Math. 84, Birkhäuser, Basel, 1988, pp. 277–293.
[31] T. Pham Dinh and H.A. Le Thi, *Stabilité de la dualité lagrangienne en optimisation d.c. (différence de deux fonctions convexes)*, C.R. Acad. Paris Sér. I Math., 318 (1994), pp. 379–384.
[32] T. Pham Dinh and H.A. Le Thi, *A d.c. optimization algorithm for solving the trust-region subproblem*, SIAM J. Optim., 8 (1998), pp. 476–505.
[33] T. Pham Dinh and H.A. Le Thi, *Convex analysis approach to d.c. programming: Theory, algorithms and applications (dedicated to Professor Hoang Tuy on the occasion of his* 70*th birthday*), Acta Math. Vietnam., 22 (1997), pp. 289–355.
[34] B. Polyak, *Introduction to Optimization*, Optimization Software Publications, New York, 1987.
[35] R.T. Rockafellar, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.
[36] R.T. Rockafellar, *Monotone operators and the proximal point algorithm,* SIAM J. Control Optim., 14 (1976), pp. 877–898.
[37] J.B. Saxe, *Embeddability of weighted Graphs in k-space is strongly NP-hard,* in Proceedings of the 17th Allerton Conference in Communication, Control, and Computing, Monticello, IL, 1979, pp. 480–489.
[38] N.Z. Shor, *Minimization Methods for Nondifferentiable Functions*, Springer-Verlag, New York, Berlin, 1979.
[39] J.F. Toland, *On subdifferential calculus and duality in nonconvex optimization*, Bull. Soc. Math. France, 60 (1979), pp. 177–183.
[40] R. Varga, *Matrix Iterative Analysis*, Prentice–Hall, Englewood Cliffs, NJ, 1962.
[41] Z. Zou, R.H. Bird, and R.B. Schnabel, *A stochastic/perturbation global optimization algorithm for distance geometry problems*, J. Global Optim., 11 (1997), pp. 91–105.

SIAM J. OPTIM.
Vol. 14, No. 1, pp. 115–138

RISK AVERSION VIA EXCESS PROBABILITIES IN STOCHASTIC PROGRAMS WITH MIXED-INTEGER RECOURSE*

RÜDIGER SCHULTZ† AND STEPHAN TIEDEMANN†

Abstract. We consider linear two-stage stochastic programs with mixed-integer recourse. Instead of basing the selection of an optimal first-stage solution on expected costs alone, we include into the objective a risk term reflecting the probability that a preselected cost threshold is exceeded. After we have put the resulting mean-risk model into perspective with stochastic dominance, we study further structural properties of the model and derive some basic stability results. In the algorithmic part of the paper, we propose a scenario decomposition method and report initial computational experience.

Key words. stochastic programming, mean-risk models, mixed-integer optimization

AMS subject classifications. 90C15, 90C11, 90C06

DOI. 10.1137/S1052623402410855

1. Introduction.

Stochastic programming with recourse deals with two-stage or multistage sequential decision processes under uncertainty and aims, in its traditional setting, at the optimization of the expected value of some random objective reflecting costs or revenues, for instance. In the present paper, we are heading for an extension towards risk aversion with a risk measure based on excess probabilities of random costs.

Criteria for the selection of risk measures are a topic of extensive discussion in the literature; cf., e.g., [2, 31, 32] and the references therein. The discussion covers a wide range of issues, such as compatibility with axiomatic settings or stochastic ordering principles, smoothness and convexity properties, and, last but not least, algorithmic possibilities for the resulting optimization problems. It goes without saying that, given the variety of criteria, there is no universally recommendable risk measure.

In recourse stochastic programming, the random variables whose risk shall be controlled are implicit entities closely related to value functions that become discontinuous and nonconvex in the presence of integer decision variables. When imposing a risk measure in this situation, care has to be taken to arrive at stochastic integer programs that are structurally sound and amenable to algorithmic treatment. In what follows we will confirm that excess probabilities lead to a risk measure that serves these purposes and is consistent with first-order stochastic dominance [31, 32].

Our paper is organized as follows. In section 2 we extend the traditional modeling in two-stage stochastic integer programming towards risk aversion. We formulate a risk measure based on excess probabilities, put it into perspective with stochastic dominance, and discuss relations of the resulting mean-risk model with robust optimization. Section 3 analyzes structure and stability of the added model components. Algorithmic issues including remarks on the efficient frontier and some first numerical experiments are presented in section 4.

*Received by the editors July 8, 2002; accepted for publication (in revised form) February 25, 2003; published electronically July 18, 2003.
http://www.siam.org/journals/siopt/14-1/41085.html
†Institute of Mathematics, University Duisburg-Essen, Lotharstr. 65, D-47048 Duisburg, Germany (schultz@math.uni-duisburg.de, tiedemann@math.uni-duisburg.de).

2. Two-stage stochastic integer programs with excess probabilities. Throughout the paper, we impose a cost minimization framework. We are given the following random mixed-integer linear program:

$$(1)\quad \min_{x,y,y'}\{c^Tx + q^Ty + q'^Ty' \ : \ Tx + Wy + W'y' = h(\omega), x \in X, y \in \mathbb{Z}_+^{\bar{m}}, y' \in \mathbb{R}_+^{m'}\}.$$

It is assumed that all ingredients in (1) have conformal dimensions, that W, W' are rational matrices, and that $X \subseteq \mathbb{R}^m$ is a nonempty closed set, possibly involving integer requirements to components of x. The right-hand side $h(\omega) \in \mathbb{R}^s$ is a random vector on some probability space $(\Omega, \mathcal{A}, \mathbb{P})$. Along with (1) we have the constraint that the variables x are to be fixed before observing $h(\omega)$ and that the variables (y, y') may be fixed afterwards. Accordingly, x and (y, y') are called first- and second-stage variables, respectively. The mentioned information constraint is usually referred to as nonanticipativity.

The mixed-integer value function

$$(2)\qquad \Phi(t) := \min\{q^Ty + q'^Ty' \ : \ Wy + W'y' = t, \ y \in \mathbb{Z}_+^{\bar{m}}, \ y' \in \mathbb{R}_+^{m'}\}$$

is an essential object in our subsequent stochastic programming models. According to integer programming theory [30], this function is real-valued on $\mathbb{R}^s$, provided that $W(\mathbb{Z}_+^{\bar{m}}) + W'(\mathbb{R}_+^{m'}) = \mathbb{R}^s$ and $\{u \in \mathbb{R}^s \ : \ W^Tu \le q, \ W'^Tu \le q'\} \ne \emptyset$, which, therefore, will be assumed throughout.

With

$$(3)\qquad Q_{\mathbb{E}}(x) := \int_\Omega (c^Tx + \Phi(h(\omega) - Tx))\, \mathbb{P}(d\omega)$$

the traditional expectation-based stochastic program with recourse is the optimization problem

$$(4)\qquad \min\{Q_{\mathbb{E}}(x) \ : \ x \in X\}.$$

Introducing the excess probability functional

$$(5)\qquad Q_{\mathbb{P}}(x) := \mathbb{P}\big(\{\omega \in \Omega \ : \ c^Tx + \Phi(h(\omega) - Tx) > \varphi_o\}\big),$$

problem (4) is extended into the mean-risk model

$$(6)\qquad \min\{Q_{\mathbb{E}}(x) + \rho Q_{\mathbb{P}}(x) \ : \ x \in X\}.$$

Here $\varphi_o \in \mathbb{R}$ denotes some preselected threshold, and $\rho \in \mathbb{R}_+$ is a suitable weight factor. The proposal to include a probability term like (5) into the objective of a stochastic program with recourse seemingly dates back to Bereanu [6] and, hitherto, has not been elaborated on in much detail.

The modeling background behind the above construction is the following: The central issue is optimizing the first-stage decisions x that have to be taken without anticipation of future realizations of $h(\omega)$. After having decided for x and observed $h(\omega)$, the remaining decisions (y, y'), of course, shall be taken in an optimal way. This results in the mixed-integer linear program defining the function Φ in (2). The costs of the sequential process of decision and observation then are expressed by the random variable $c^Tx + \Phi(h(\omega) - Tx)$. Finding an optimal first-stage decision $x \in X$ can be understood as selecting a "best" random variable from the indexed family

$\big(c^T x + \Phi(h(\omega) - Tx)\big)_{x \in X}$. The models (4) and (6) are based on scalar criteria for making this selection.

The mean-risk model (6) aims at controlling variability of second-stage solutions and, thus, is closely related to the robust optimization approach proposed by Mulvey, Vanderbei, and Zenios in [29]. To discuss similarities and differences between (6) and [29] we denote $(y_{opt}(x,\omega), y'_{opt}(x,\omega))$ an optimal solution to the optimization problem defining $\Phi(h(\omega) - Tx)$; cf. (2). The random variable $c^T x + \Phi(h(\omega) - Tx)$ then coincides with

$$f(x,\omega) \ := \ c^T x + q^T y_{opt}(x,\omega) + q'^T y'_{opt}(x,\omega),$$

and the mean-risk model (6) can be written as

$$\min \big\{ \mathbb{E}_\omega[f(x,\omega)] \ + \ \rho Risk_\omega[f(x,\omega)] \ : \ x \in X \big\}. \tag{7}$$

Here $\mathbb{E}_\omega$ denotes the expectation, and $Risk_\omega$ is a symbol for an abstract risk measure, specified in our model by (5). The ROBUST model of [29] incorporates both variability of second-stage costs and penalization of second-stage infeasibility. In our terminology, the variability term of [29] is based on the random variable

$$g(x,\omega) \ := \ c^T x + q^T y(\omega) + q'^T y'(\omega),$$

and the counterpart model of [29] to our model (6) would read

$$\min \big\{ \mathbb{E}_\omega[g(x,\omega)] \ + \ \rho Risk_\omega[g(x,\omega)] \ : \tag{8}$$

$$Tx + Wy(\omega) + W'y'(\omega) = h(\omega), \ x \in X, \ y(\omega) \in \mathbb{Z}^{\bar{m}}_+, \ y'(\omega) \in \mathbb{R}^{m'}_+ \ \forall \omega \in \Omega \big\},$$

where $Risk_\omega$ is specified by the variance.

In (7) the statistical parameter $\mathbb{E}_\omega[.] + \rho Risk_\omega[.]$ is optimized over all feasible first-stage solutions $x \in X$ and all optimal second-stage decisions $(y_{opt}(x,\omega), y'_{opt}(x,\omega))$. In (8) the statistical parameter is optimized jointly over all feasible first- and second-stage decisions. The essential structural difference between (7) and (8), hence, is in the order of integration and second-stage minimization. If the statistical parameter in (7), (8) is just $\mathbb{E}_\omega[.]$, then (7) and (8) are equivalent, which is a basic fact of stochastic linear programming. For statistical parameters involving risk terms this equivalence is no longer valid in general. In particular, second-stage portions of optimal solutions to (8) no longer need to be optimal for the second-stage, i.e., for the optimization problem behind $\Phi(h(\omega) - Tx)$. For problems without integer variables this issue is addressed in [21, 48].

In [48] it is shown that (7) and (8) are equivalent if the risk term depends on the second-stage costs only and fulfills a monotonicity condition. If, in addition, the risk term is convex, then numerical treatment of the problem is possible by a direct transfer of L-shaped decomposition techniques. Another issue discussed in [48] is ranking the random variables $\big(c^T x + \Phi(h(\omega) - Tx)\big)_{x \in X}$ by means of a convex disutility function. Again L-shaped techniques can readily be employed to solve the resulting optimization problem.

Apart from the integer decision variables in both stages of our model, the major distinction between (6) and the ranking model of [48] is in the nonconvex discontinuous disutility function we employ, in fact a sum of indicator functions of level sets. In section 4 we will show how numerical treatment of (6) becomes possible by a reformulation using additional Boolean variables.

Among the fundamental concepts in decision theory the relation of stochastic dominance introduces a partial order in the space of real random variables. This provides some basis for selecting "best" members from families of random variables. Recently, Ogryczak and Ruszczyński have studied scalar criteria and their consistency with the multiobjective criteria induced by stochastic dominance; see [31, 32]. Let us quickly outline how the scalar criterion proposed in (6) can be put into the perspective of [31, 32].

Let $f(x,\omega)$, $x \in X \subseteq \mathbb{R}^m$, be real random variables on some probability space $(\Omega, \mathcal{A}, \mathbb{P})$. Consider the distribution function

$$F_x(\eta) \;:=\; \mathbb{P}\big(\{\omega \in \Omega \,:\, f(x,\omega) \le \eta\}\big).$$

Since we prefer smaller outcomes to larger ones we define that x' dominates x'' to first degree ($x' \succeq x''$) if

$$F_{x'}(\eta) \;\ge\; F_{x''}(\eta) \quad \forall \eta \in \mathbb{R}. \tag{9}$$

Let m_x be the expectation $\mathbb{E}\big(f(x,\omega)\big)$, and let r_x denote some functional measuring the risk of the outcome $f(x,\omega)$. We adapt the setting of [31, 32] to our preference for smaller outcomes and say that the mean-risk model (m_x, r_x) is α-consistent with first degree stochastic dominance, where $\alpha > 0$, if $x' \succeq x''$ implies that $m_{x'} + \alpha r_{x'} \le m_{x''} + \alpha r_{x''}$.

LEMMA 2.1. *The mean-risk model (m_x, r_x) with $m_x := \mathbb{E}\big(f(x,\omega)\big)$ and $r_x := \mathbb{P}\big(\{\omega \in \Omega \,:\, f(x,\omega) > \varphi_o\}\big)$, with fixed $\varphi_o \in \mathbb{R}$, is α-consistent with first degree stochastic dominance for all $\alpha > 0$.*

Proof. The fact that (9) implies that $m_{x'} \le m_{x''}$ is well known in probability theory. Moreover, it holds that

$$r_{x'} \;=\; 1 - F_{x'}(\varphi_o) \;\le\; 1 - F_{x''}(\varphi_o) \;=\; r_{x''},$$

and the proof is complete. □

In conclusion, the excess probability in (5) entails a risk measure fulfilling a weak requirement of consistency with stochastic dominance. For further risk measures fulfilling stronger consistencies with stochastic dominance we refer the reader to [31, 32].

In subsequent sections we will show that (6) is well-posed from the formal viewpoint. We will establish structural properties of the functional $Q_{\mathbb{P}}$, and we will demonstrate that solution methodology from mixed-integer linear programming (the class our initial random optimization problem (1) belongs to) can be employed for solving (6).

The expectation-based optimization problem (4) meanwhile belongs to the well-studied objects in stochastic programming. Therefore, the main focus in our further investigations will be on the functional $Q_{\mathbb{P}}$. Without going into detail, we mention that, under mild conditions, $Q_{\mathbb{E}}$ is real-valued and lower semicontinuous and that $Q_{\mathbb{E}}$ is continuous if the distribution of $h(\omega)$ has a density. Optimal values and optimal solutions to (4) behave in a stable manner under perturbations of the probability distribution of $h(\omega)$. This gives rise to discrete approximations of the probability distribution for which (4) can be rewritten equivalently as a block-structured mixed-integer linear program. The latter is amenable to decomposition methods splitting (4) into smaller mixed-integer linear programs that are often tractable by standard solvers like CPLEX [14]. Detailed expositions of the mentioned results can be found in [1, 13, 18, 45, 46]. Without integer requirements, (4) becomes a convex optimization problem allowing for application of various analytical and algorithmic techniques from convex analysis; see [10, 19, 35] and the references therein.

3. Structure and stability. The mixed-integer value function Φ from (2) is crucial for the structural understanding of $Q_{\mathbb{P}}$. From parametric optimization [4, 11] the following is known about Φ.

PROPOSITION 3.1. *Assume that* $W(\mathbb{Z}_+^{\bar{m}})+W'(\mathbb{R}_+^{m'}) = \mathbb{R}^s$ *and* $\{u \in \mathbb{R}^s \ : \ W^T u \leq q,\ {W'}^T u \leq q'\} \neq \emptyset$. *Then it holds that*

(i) Φ *is real-valued and lower semicontinuous on* $\mathbb{R}^s$;

(ii) *there exists a countable partition* $\mathbb{R}^s = \cup_{i=1}^{\infty} \mathcal{T}_i$ *such that the restrictions of* Φ *to* $\mathcal{T}_i$ *are piecewise linear and Lipschitz continuous with a uniform constant not depending on* i; *more specifically, on each* $\mathcal{T}_i$, *the function* Φ *admits a representation*

$$\Phi(t) = \min_{y \in Y(t)} \big\{q^T y + \max_{k=1,\ldots,K} d_k^T (t - Wy)\big\},$$

where $Y(t) := \{y \in \mathbb{Z}_+^{\bar{m}} \ : \ t \in Wy + W'(\mathbb{R}_+^{m'})\}$ *and* $d_k, k = 1, \ldots, K$, *are the vertices of the polyhedron* $\{u \in \mathbb{R}^s \ : \ {W'}^T u \leq q'\}$;

(iii) *each of the sets* $\mathcal{T}_i$ *has a representation* $\mathcal{T}_i = \{t_i + \mathcal{K}\} \ \setminus \ \cup_{j=1}^{N} \{t_{ij} + \mathcal{K}\}$, *where* $\mathcal{K}$ *denotes the polyhedral cone* $W'(\mathbb{R}_+^{m'})$ *and* t_i, t_{ij} *are suitable points from* $\mathbb{R}^s$; *moreover,* N *does not depend on* i;

(iv) *there exist positive constants* β, γ *such that* $|\Phi(t_1) - \Phi(t_2)| \leq \beta \|t_1 - t_2\| + \gamma$ *whenever* $t_1, t_2 \in \mathbb{R}^s$.

To facilitate notation we introduce for all $x \in \mathbb{R}^m$

$$\begin{aligned} M(x) &:= \{h \in \mathbb{R}^s \ : \ c^T x + \Phi(h - Tx) > \varphi_o\}, \\ M_e(x) &:= \{h \in \mathbb{R}^s \ : \ c^T x + \Phi(h - Tx) = \varphi_o\}, \\ M_d(x) &:= \{h \in \mathbb{R}^s \ : \ \Phi \text{ is discontinuous at } h - Tx\}. \end{aligned}$$

By $\liminf_{x_n \to x} M(x_n)$ and $\limsup_{x_n \to x} M(x_n)$ we denote the (set-theoretic) limes inferior and limes superior, i.e., the sets of all points belonging to all but a finite number of the sets $M(x_n)$, $n \in \mathbb{N}$, and to infinitely many of the sets $M(x_n)$, respectively.

LEMMA 3.2. *Assume that* $W(\mathbb{Z}_+^{\bar{m}}) + W'(\mathbb{R}_+^{m'}) = \mathbb{R}^s$ *and* $\{u \in \mathbb{R}^s \ : \ W^T u \leq q,\ {W'}^T u \leq q'\} \neq \emptyset$. *Then it holds for all* $x \in \mathbb{R}^m$ *that*

$$M(x) \ \subseteq \ \liminf_{x_n \to x} M(x_n) \ \subseteq \ \limsup_{x_n \to x} M(x_n) \ \subseteq \ M(x) \cup M_e(x) \cup M_d(x).$$

Proof. To show the first inclusion let $h \in M(x)$. By the lower semicontinuity of Φ (Proposition 3.1(i)) we have

$$\liminf_{x_n \to x} (c^T x_n + \Phi(h - Tx_n)) \ \geq \ c^T x + \Phi(h - Tx) \ > \ \varphi_o.$$

Therefore, there exists an $n_o \in \mathbb{N}$ such that $c^T x_n + \Phi(h - Tx_n) > \varphi_o$ for all $n \geq n_o$, implying that $h \in M(x_n)$ for all $n \geq n_o$, and we obtain that $M(x) \subseteq \liminf_{x_n \to x} M(x_n)$. The second inclusion being valid by definition, we turn to the third.

Let $h \in \limsup_{x_n \to x} M(x_n) \setminus M(x)$. Then there exists an infinite subset $\tilde{\mathbb{N}}$ of $\mathbb{N}$ such that

$$c^T x_n + \Phi(h - Tx_n) > \varphi_o \qquad \forall n \in \tilde{\mathbb{N}} \quad \text{and} \quad c^T x + \Phi(h - Tx) \leq \varphi_o.$$

Now two cases are possible. First, Φ is continuous at $h - Tx$. Passing to the limit in the first inequality then yields that $c^T x + \Phi(h - Tx) \geq \varphi_o$, and $h \in M_e(x)$. Second, Φ is discontinuous at $h - Tx$. In other words, $h \in M_d(x)$. ∎

For convenience, we denote μ the image measure $\mathbb{P} \circ h^{-1}$ on $\mathbb{R}^s$. By the lower semicontinuity of Φ, the sets $M(x), M_e(x)$, and $M_d(x)$ are μ-measurable for all $x \in \mathbb{R}^m$.

PROPOSITION 3.3. *Assume that* $W(\mathbb{Z}_+^{\bar{m}}) + W'(\mathbb{R}_+^{m'}) = \mathbb{R}^s$ *and* $\{u \in \mathbb{R}^s \,:\, W^T u \leq q,\ W'^T u \leq q'\} \neq \emptyset$. *Then* $Q_{\mathbb{P}} : \mathbb{R}^m \longrightarrow \mathbb{R}$ *is a real-valued lower semicontinuous function. If, in addition, it holds that* $\mu\big(M_e(x) \cup M_d(x)\big) = 0$, *then* $Q_{\mathbb{P}}$ *is continuous at* x. *If* μ *has a density, then* $Q_{\mathbb{P}}$ *is continuous on* $\mathbb{R}^m$.

Proof. The function $Q_{\mathbb{P}}$ is real-valued on $\mathbb{R}^m$ due to the μ-measurability of $M(x)$. By Lemma 3.2 and the (semi-) continuity of the probability measure on sequences of sets we have for all $x \in \mathbb{R}^m$

$$Q_{\mathbb{P}}(x) \;=\; \mu\big(M(x)\big) \;\leq\; \mu\big(\liminf_{x_n \to x} M(x_n)\big) \;\leq\; \liminf_{x_n \to x} \mu\big(M(x_n)\big) \;=\; \liminf_{x_n \to x} Q_{\mathbb{P}}(x_n),$$

establishing the asserted lower semicontinuity. In case $\mu\big(M_e(x) \cup M_d(x)\big) = 0$ this argument extends as follows:

$$\begin{aligned} Q_{\mathbb{P}}(x) \;&=\; \mu\big(M(x)\big) \;=\; \mu\big(M(x) \cup M_e(x) \cup M_d(x)\big) \;\geq\; \mu\big(\limsup_{x_n \to x} M(x_n)\big) \\ &\geq\; \limsup_{x_n \to x} \mu\big(M(x_n)\big) \;=\; \limsup_{x_n \to x} Q_{\mathbb{P}}(x_n), \end{aligned}$$

and $Q_{\mathbb{P}}$ is continuous at x. In view of Proposition 3.1(ii), (iii), for given $x \in \mathbb{R}^m$, both $M_e(x)$ and $M_d(x)$ are contained in a countable union of hyperplanes, i.e., in a set of Lebesgue measure zero. Since μ has a density, it is absolutely continuous with respect to the Lebesgue measure; hence $\mu\big(M_e(x) \cup M_d(x)\big) = 0$, and the proof is complete. □

Remark 3.4. The above lower semicontinuity of $Q_{\mathbb{P}}$ implies in particular that problem (6) is well-posed in the sense that, provided that X is compact, the infimum in (6) is finite and is attained. Given the discontinuity of Φ, the well-posedness of (6) may become critical with other risk measures; cf. [27]. To see this let us consider the variance, leading to the functional

$$Q_{\mathbb{V}}(x) := \int_\Omega \left[c^T x + \Phi(h(\omega) - Tx) - \int_\Omega (c^T x + \Phi(h(\omega) - Tx))\, \mathbb{P}(d\omega)\right]^2 \mathbb{P}(d\omega).$$

We consider the counterpart

$$\min\{Q_{\mathbb{E}}(x) + \rho Q_{\mathbb{V}}(x) \;:\; x \in X\} \tag{10}$$

to problem (6) with the specifications $m = s = 1$, $c = 1$, $T = -1$, $\rho = 4$, $X = \{x \in \mathbb{R} : x \geq 0\}$, $\Phi(t) = \min\{y : y \geq t, y \in \mathbb{Z}\}$, and $h(\omega)$ attaining the values 0 and $\frac{1}{2}$ each with probability $\frac{1}{2}$. One computes that $Q_{\mathbb{E}}(x) = x + \frac{1}{2}\lceil x \rceil + \frac{1}{2}\lceil x + \frac{1}{2} \rceil$ and $Q_{\mathbb{V}}(x) = \frac{1}{4}(\lceil x \rceil - \lceil x + \frac{1}{2} \rceil)^2$.

Then (10) has the infimum 1, and any sequence $(x_n)_{n \in \mathbb{N}}$ with $x_n \downarrow 0, x_n \neq 0$ is a minimizing sequence. However, the infimum is not attained since the objective value for $x = 0$ is $\frac{3}{2}$.

Before we turn to the Lipschitz continuity of $Q_{\mathbb{P}}$ we study the boundary $bd\, M(x)$ of the set $M(x) = \{h \in \mathbb{R}^s : c^T x + \Phi(h - Tx) > \varphi_o\}$.

LEMMA 3.5. *Adopt the setting of Proposition* 3.1 *and assume in addition that* q, q' *are rational vectors. Then there exist a* $K_o \in \mathbb{N}$, *affine hyperplanes* $H_{\kappa y} \subseteq \mathbb{R}^s, \kappa = 1, \ldots, K_o, y \in \mathbb{Z}_+^{\bar{m}}$, *and matrices* $T_\kappa, \kappa = 1, \ldots, K_o$, *such that*

(i)

$$bd\, M(x) \;\subseteq\; \bigcup_{\kappa=1}^{K_o} \bigcup_{y\in\mathbb{Z}_+^{\bar m}} \{T_\kappa x \,+\, H_{\kappa y}\} \qquad \forall x\in\mathbb{R}^m; \tag{11}$$

(ii) *there exists* $\mathtt{r} > 0$ *such that for any* $\kappa \in \{1,\dots,K_o\}$ *and any* $y', y'' \in \mathbb{Z}_+^{\bar m}$, *the hyperplanes* $H_{\kappa y'}, H_{\kappa y''}$ *are either identical or have a Hausdorff distance of at least* $\mathtt{r}$.

Proof. Let us assume that $W'(\mathbb{R}_+^{m'})$ has a representation $\{t\in\mathbb{R}^s : \tilde d_l^T t \le 0, l = 1,\dots,L\}$ with suitable vectors $\tilde d_l, l = 1,\dots,L$. According to Proposition 3.1(ii) we obtain that h belongs to the complement of $M(x)$ if and only if there exists a $y\in\mathbb{Z}_+^{\bar m}$ such that the following system of inequalities in h is fulfilled:

$$q^T y + d_k^T(h - Tx - Wy) \le \varphi_o - c^T x, \quad k = 1,\dots,K, \tag{12}$$

$$\tilde d_l^T(h - Tx - Wy) \le 0, \quad l = 1,\dots,L. \tag{13}$$

If $d_k = 0$ for some $k\in\{1,\dots,K\}$, then the corresponding inequality in (12) turns into

$$q^T y \;\le\; \varphi_o - c^T x,$$

leading to a restriction on y but not on h. It then would actually suffice to form the union in (11) over $\mathbb{Z}_+^{\bar m}\cap\{y \,:\, q^T y \,\le\, \varphi_o - c^T x\}$ instead of $\mathbb{Z}_+^{\bar m}$. So let us assume that $d_k \ne 0$ for all $k\in\{1,\dots,K\}$. For some fixed k we assume that the first component $d_{k(1)}$ of d_k is nonzero. We put

$$\tilde T_k \;:=\; T - (d_{k(1)})^{-1}\begin{pmatrix} c^T \\ 0 \end{pmatrix}$$

and obtain that

$$d_k^T\tilde T_k \;=\; d_k^T T - (d_{k(1)})^{-1} d_k^T \begin{pmatrix} c^T \\ 0 \end{pmatrix} \;=\; d_k^T T - c^T.$$

This allows us to rewrite (12) as

$$q^T y + d_k^T(h - \tilde T_k x - Wy) \;\le\; \varphi_o, \quad k = 1,\dots,K. \tag{14}$$

If h belongs to the boundary of $M(x)$, then at least one of the inequalities in (13) and (14) has to be fulfilled as an equation. Defining the affine hyperplanes

$$H_{ky} \;:=\; \{t\in\mathbb{R}^s \,:\, d_k^T t \;=\; \varphi_o + d_k^T W y - q^T y\}, \qquad y\in\mathbb{Z}_+^{\bar m},\ k\in\{1,\dots,K\} \tag{15}$$

and

$$H_{ly} \;:=\; \{t\in\mathbb{R}^s \,:\, \tilde d_l^T t \;=\; \tilde d_l^T W y\}, \qquad y\in\mathbb{Z}_+^{\bar m},\ l\in\{1,\dots,L\} \tag{16}$$

we obtain that $h\in bd\, M(x)$ implies that

$$h \;\in\; \bigcup_{k=1}^{K}\bigcup_{y\in\mathbb{Z}_+^{\bar m}} \{\tilde T_k x + H_{ky}\} \cup \bigcup_{l=1}^{L}\bigcup_{y\in\mathbb{Z}_+^{\bar m}} \{Tx + H_{ly}\} \;=:\; \bigcup_{\kappa=1}^{K_o}\bigcup_{y\in\mathbb{Z}_+^{\bar m}} \{T_\kappa x \,+\, H_{\kappa y}\},$$

proving (i).

Since q', W' are rational, $d_k, k = 1, \dots, K$, and $\tilde{d}_l, l = 1, \dots, L$, can be selected as rational vectors. According to the definitions in (15), (16) and the rationality of q, W this yields that there exists $\mathtt{r} > 0$ such that, for arbitrary $\kappa \in \{1, \dots, K_o\}$ and $y', y'' \in \mathbb{Z}_+^{\bar{m}}$, the hyperplanes $H_{\kappa y'}, H_{\kappa y''}$ are either identical or have a Hausdorff distance that is bounded below by $\mathtt{r}$. The latter is seen as follows. For $H_{\kappa y}$ we have a representation

$$H_{\kappa y} \;=\; \{t \in \mathbb{R}^s \;:\; \delta_\kappa^T t \;=\; b_{\kappa o} + b_\kappa^T y\} \tag{17}$$

with rational vectors δ_κ, b_κ. So if $H_{\kappa y'} \neq H_{\kappa y''}$, then $b_\kappa^T(y' - y'') \neq 0$. Let $b_* > 0$ denote the least common multiple of the (absolute values of) the denominators of the components of b_κ. By $y' - y'' \in \mathbb{Z}$ the number $b_\kappa^T(y' - y'') \neq 0$ then has to be a multiple of $\frac{1}{b_*}$. In the representations for $H_{\kappa y'}$ and $H_{\kappa y''}$ the right-hand sides therefore differ at least by $\frac{1}{b_*}$, proving the claim. □

PROPOSITION 3.6. *Assume that q, q' are rational vectors, $W(\mathbb{Z}_+^{\bar{m}}) + W'(\mathbb{R}_+^{m'}) = \mathbb{R}^s$, $\{u \in \mathbb{R}^s \;:\; W^T u \leq q,\; W'^T u \leq q'\} \neq \emptyset$, and that for any nonsingular linear transformation $B \in L(\mathbb{R}^s, \mathbb{R}^s)$ all one-dimensional marginal distributions of $\mu \circ B$ have bounded densities which, outside some bounded interval, are monotonically decreasing with the growing absolute value of the argument. Then $Q_{\mathbb{P}}$ is Lipschitz continuous on any bounded subset of $\mathbb{R}^m$.*

Proof. Let S be a bounded subset of $\mathbb{R}^m$ and $x', x'' \in S$. It holds that

$$|Q_{\mathbb{P}}(x') - Q_{\mathbb{P}}(x'')| = |\mu(M(x')) - \mu(M(x''))| \leq \mu(M(x') \setminus M(x'')) + \mu(M(x'') \setminus M(x')).$$

For symmetry reasons it is sufficient to establish the assertion for the first member of the sum on the right.

Recall the representation (17) for the hyperplanes $H_{\kappa y}$ arising in Lemma 3.5 and put $\bar{\delta}_{\kappa y} := b_{\kappa o} + b_\kappa^T y$ such that we have $H_{\kappa y} = \{t \in \mathbb{R}^s : \delta_\kappa^T t = \bar{\delta}_{\kappa y}\}$ for $\kappa = 1, \dots K_o, y \in \mathbb{Z}_+^{\bar{m}}$. Consider the halfspaces

$$H_{\kappa y}^- := \{t \in \mathbb{R}^s : \delta_\kappa^T t \leq \bar{\delta}_{\kappa y}\}. \tag{18}$$

Then it holds that

$$M(x') \setminus M(x'') \;\subseteq\; \bigcup_{\kappa=1}^{K_o} \bigcup_{y \in \mathbb{Z}_+^{\bar{m}}} \Big\{ \mathrm{cl}\Big[\big\{T_\kappa x' + H_{\kappa y}^-\big\} \setminus \big\{T_\kappa x'' + H_{\kappa y}^-\big\}\Big] \cup \mathrm{cl}\Big[\big\{T_\kappa x'' + H_{\kappa y}^-\big\} \setminus \big\{T_\kappa x' + H_{\kappa y}^-\big\}\Big] \Big\}.$$

As usual, the symbol "cl" denotes the closure. To estimate the μ-measure of the set on the right we fix some $\kappa \in \{1, \dots, K_o\}$. Without loss of generality we may assume that $\delta_\kappa^T T_\kappa x'' \leq \delta_\kappa^T T_\kappa x'$ such that $\{T_\kappa x'' + H_{\kappa y}^-\} \setminus \{T_\kappa x' + H_{\kappa y}^-\} = \emptyset$ for all $y \in \mathbb{Z}_+^{\bar{m}}$. It remains to estimate

$$\mu\Big(\bigcup_{y \in \mathbb{Z}_+^{\bar{m}}} \mathrm{cl}\Big[\big\{T_\kappa x' + H_{\kappa y}^-\big\} \setminus \big\{T_\kappa x'' + H_{\kappa y}^-\big\}\Big]\Big)$$

$$= \mu\Big(\bigcup_{y \in \mathbb{Z}_+^{\bar{m}}} \big\{h \in \mathbb{R}^s \;:\; \bar{\delta}_{\kappa y} + \delta_\kappa^T T_\kappa x'' \;\leq\; \delta_\kappa^T h \;\leq\; \bar{\delta}_{\kappa y} + \delta_\kappa^T T_\kappa x'\big\}\Big).$$

Let B_κ be a nonsingular matrix whose first row coincides with δ_κ^T. Let $\zeta := B_\kappa h$ be the corresponding linear transformation, and let $\zeta_{(1)}$ denote the first component of ζ. Then it holds that

$$\begin{aligned}
&\mu\big(\{h \in \mathbb{R}^s \,:\, \bar\delta_{\kappa y} + \delta_\kappa^T T_\kappa x'' \,\le\, \delta_\kappa^T h \,\le\, \bar\delta_{\kappa y} + \delta_\kappa^T T_\kappa x'\}\big) \\
&\qquad = (\mu \circ B_\kappa^{-1})\big(B_\kappa(\{h \in \mathbb{R}^s \,:\, \bar\delta_{\kappa y} + \delta_\kappa^T T_\kappa x'' \,\le\, \delta_\kappa^T h \,\le\, \bar\delta_{\kappa y} + \delta_\kappa^T T_\kappa x'\})\big) \\
&\qquad = (\mu \circ B_\kappa^{-1})\big(\{\zeta \in \mathbb{R}^s \,:\, \bar\delta_{\kappa y} + \delta_\kappa^T T_\kappa x'' \,\le\, \zeta_{(1)} \,\le\, \bar\delta_{\kappa y} + \delta_\kappa^T T_\kappa x'\}\big) \\
&\qquad = \int_{\bar\delta_{\kappa y} + \delta_\kappa^T T_\kappa x''}^{\bar\delta_{\kappa y} + \delta_\kappa^T T_\kappa x'} \theta_\kappa(\tau)\, d\tau.
\end{aligned}$$

In the last row above, θ_κ denotes a marginal density of the first component with respect to the image measure $\mu \circ B_\kappa^{-1}$. The density is selected such that it fulfills the requirements made in the assumptions of our proposition.

Let $(\bar\delta_{\kappa i})_{i\in\mathbb{N}}$ be an enumeration of the distinct values attained by the numbers $\bar\delta_{\kappa y}$, $y \in \mathbb{Z}_+^{\bar m}$. By the argument from the proof of Lemma 3.5(ii), the sequence $(\bar\delta_{\kappa i})_{i\in\mathbb{N}}$ has no accumulation points.

Since x', x'' belong to the bounded set S and the $\bar\delta_{\kappa i}$ do not accumulate, there exists an index $\bar\imath = \bar\imath(S)$, independent of x', x'', such that the intervals $[\bar\delta_{\kappa i} + \delta_\kappa^T T_\kappa x'', \bar\delta_{\kappa i} + \delta_\kappa^T T_\kappa x']$, up to renumbering, meet the bounded interval arising in the assumptions at most for $i \le \bar\imath$. By assumption, we have an upper bound $\bar\theta_\kappa$ for $\theta_\kappa(.)$. For $i > \bar\imath$, we denote $\tilde\tau_{\kappa i}$ the left or right endpoint of $[\bar\delta_{\kappa i} + \delta_\kappa^T T_\kappa x'', \bar\delta_{\kappa i} + \delta_\kappa^T T_\kappa x']$ depending on whether θ_κ is decreasing or increasing on that interval. This allows the estimate

$$\begin{aligned}
\sum_{i\in\mathbb{N}} \int_{\bar\delta_{\kappa i} + \delta_\kappa^T T_\kappa x''}^{\bar\delta_{\kappa i} + \delta_\kappa^T T_\kappa x'} \theta_\kappa(\tau)\, d\tau \le{}& \sum_{i \le \bar\imath} \bar\theta_\kappa \cdot \|\delta_\kappa^T T_\kappa\| \cdot \|x' - x''\| \\
&+ \sum_{i > \bar\imath} \theta_\kappa(\tilde\tau_{\kappa i}) \cdot \|\delta_\kappa^T T_\kappa\| \cdot \|x' - x''\|.
\end{aligned}$$

Next we show that $\sum_{i>\bar\imath} \theta_\kappa(\tilde\tau_{\kappa i})$ is finite. It is sufficient to do that for the sum over all $i > \bar\imath$ belonging to those $\tilde\tau_{\kappa i}$ around which θ_k is decreasing. For the remaining $i > \bar\imath$ a similar argument applies. Since the $\bar\delta_{\kappa i}$ do not accumulate, there exists an $\varepsilon > 0$ such that

$$1 \;\ge\; \sum_i \int_{\tilde\tau_{\kappa i} - \varepsilon}^{\tilde\tau_{\kappa i}} \theta_\kappa(\tau)\, d\tau \;\ge\; \sum_i \int_{\tilde\tau_{\kappa i} - \varepsilon}^{\tilde\tau_{\kappa i}} \theta_\kappa(\tilde\tau_{\kappa i})\, d\tau \;=\; \varepsilon \cdot \sum_i \theta_\kappa(\tilde\tau_{\kappa i}).$$

This provides the desired finiteness. Repeating the above arguments for all $\kappa = 1, \ldots, K_o$ we obtain a constant $c_o > 0$, not depending on x', x'', such that

$$\mu(M(x') \setminus M(x'')) \;\le\; c_o \cdot \|x' - x''\|,$$

and the proof is complete. □

Remark 3.7. Among the probability measures fulfilling the requirements of Proposition 3.6 there are the so-called r-convex measures, in particular the (non-degenerate) multivariate normal distribution and the t-distribution. For details see [44].

Remark 3.8. Without integer requirements ($\bar m = 0$) the function Φ is less complicated. Imposing the assumptions $W'(\mathbb{R}_+^{m'}) = \mathbb{R}^s$, $\{u \in \mathbb{R}^s : W'^T u \le q'\} \ne \emptyset$, one obtains due to linear programming duality

$$\Phi(t) \;=\; \max\{t^T u \,:\, W'^T u \le q'\} \;=\; \max_{k=1,\ldots,K} d_k^T t,$$

where $d_k, k = 1, \ldots, K$, are the vertices of $\{u \in \mathbb{R}^s : W'^T u \leq q'\}$. This implies that, for all $x \in \mathbb{R}^m$, the set $M_d(x)$ is empty and the complement $M(x)^c$ of $M(x)$ is a single polyhedron. This provides a link with linear chance constraints which belong to the well-studied objects in stochastic programming [10, 19, 35]. Lower semicontinuity of $Q_{\mathbb{P}}$, for instance, then already follows from Proposition 3.1 in [42]. Further material about $Q_{\mathbb{P}}$ in the absence of integer requirements can be found in [39].

Remark 3.9. For some early work on continuity properties of general probability functionals we refer to Raik [37, 38]; see also [20, 35].

Remark 3.10. With the additional assumption that $\int_{\mathbb{R}^s} \|h\| \, \mu(dh) < \infty$ the statements of Propositions 3.3 and 3.6 are valid for $Q_{\mathbb{E}}$ as well. The set $M_e(x)$ turns out to be irrelevant for the continuity of $Q_{\mathbb{E}}$ such that the corresponding assumption turns into $\mu(M_d(x)) = 0$. For details we refer the reader to [44, 45].

In many practical modeling situations knowledge about the underlying probability measure $\mu = \mathbb{P} \circ h^{-1}$ in (6) is subjective. Furthermore, the multivariate integration required in (3) and (5) often has to rely on approximations, in particular if $h(\omega)$ is multidimensional and follows a continuous probability distribution. These issues motivate the stability analysis of (6) under perturbations of μ. The aim is to identify sufficient conditions such that "small" perturbations in μ result in only "small" perturbations of optimal values and optimal solutions to (6). Qualitative and quantitative continuity of $Q_{\mathbb{E}}$ and $Q_{\mathbb{P}}$ jointly in x and μ then become a key issue. For $Q_{\mathbb{E}}$ this has been settled in [1, 45, 46] such that we will now focus on $Q_{\mathbb{P}}$.

Let $\mathcal{P}(\mathbb{R}^s)$ denote the set of all Borel probability measures on $\mathbb{R}^s$. We consider $Q_{\mathbb{P}}$ as a function mapping from $\mathbb{R}^m \times \mathcal{P}(\mathbb{R}^s)$ to $\mathbb{R}$, where $\mathbb{R}^m$ is equipped with the usual topology. On $\mathcal{P}(\mathbb{R}^s)$ a notion of convergence is desirable that is both sufficiently general to cover relevant applications and sufficiently specific to enable substantial statements. This is met by weak convergence of probability measures for which [9] is a basic reference. We say that a sequence $\{\mu_n\}$ in $\mathcal{P}(\mathbb{R}^s)$ converges weakly to $\mu \in \mathcal{P}(\mathbb{R}^s)$, written $\mu_n \xrightarrow{w} \mu$, if for any bounded continuous function $g : \mathbb{R}^s \to \mathbb{R}$ it holds that

$$\int_{\mathbb{R}^s} g(h)\mu_n(dh) \to \int_{\mathbb{R}^s} g(h)\mu(dh) \quad \text{as} \quad n \to \infty. \tag{19}$$

PROPOSITION 3.11. *Assume that* $W(\mathbb{Z}_+^{\bar{m}}) + W'(\mathbb{R}_+^{m'}) = \mathbb{R}^s$ *and* $\{u \in \mathbb{R}^s : W^T u \leq q, \ W'^T u \leq q'\} \neq \emptyset$. *Let* $\mu \in \mathcal{P}(\mathbb{R}^s)$ *be such that* $\mu\big(M_e(x) \cup M_d(x)\big) = 0$. *Then* $Q_{\mathbb{P}} : \mathbb{R}^m \times \mathcal{P}(\mathbb{R}^s) \longrightarrow \mathbb{R}$ *is continuous at* (x, μ).

Proof. Let $x_n \longrightarrow x$ and $\mu_n \xrightarrow{w} \mu$ be arbitrary sequences. By $\chi_n, \chi : \mathbb{R}^s \longrightarrow \{0, 1\}$ we denote the indicator functions of the sets $M(x_n), M(x), n \in \mathbb{N}$. In addition, we introduce the exceptional set

$$E := \{h \in \mathbb{R}^s : \exists h_n \to h \text{ such that } \chi_n(h_n) \not\to \chi(h)\}.$$

Now we have $E \subseteq M_e(x) \cup M_d(x)$. To see this, consider $h \in \big(M_e(x) \cup M_d(x)\big)^c = M_e(x)^c \cap M_d(x)^c$, where the superscript c denotes the set-theoretic complement. Then Φ is continuous at $h - Tx$, and either $c^T x + \Phi(h - Tx) > \varphi_o$ or $c^T x + \Phi(h - Tx) < \varphi_o$. Thus, for any sequence $h_n \to h$ there exists an $n_o \in \mathbb{N}$ such that for all $n \geq n_o$ either $c^T x_n + \Phi(h_n - Tx_n) > \varphi_o$ or $c^T x_n + \Phi(h_n - Tx_n) < \varphi_o$. Hence, $\chi_n(h_n) \to \chi(h)$ as $h_n \to h$, implying that $h \in E^c$.

In view of $E \subseteq M_e(x) \cup M_d(x)$ and $\mu\big(M_e(x) \cup M_d(x)\big) = 0$ we obtain that $\mu(E) = 0$. A theorem on weak convergence of image measures attributed to Rubin in

[9, p. 34] now yields that the weak convergence $\mu_n \stackrel{w}{\longrightarrow} \mu$ implies the weak convergence $\mu_n \circ \chi_n^{-1} \stackrel{w}{\longrightarrow} \mu \circ \chi^{-1}$.

Note that $\mu_n \circ \chi_n^{-1}, \mu \circ \chi^{-1}, n \in \mathbb{N}$, are probability measures on $\{0,1\}$. Their weak convergence then particularly implies that

$$\mu_n \circ \chi_n^{-1}\big(\{1\}\big) \;\longrightarrow\; \mu \circ \chi^{-1}\big(\{1\}\big).$$

In other words, $\mu_n\big(M(x_n)\big) \longrightarrow \mu\big(M(x)\big)$ or $Q_{\mathbb{P}}(x_n, \mu_n) \longrightarrow Q_{\mathbb{P}}(x, \mu)$. □

To analyze the quantitative continuity of $Q_{\mathbb{P}}$ as a function of the underlying probability measure let us again consider the hyperplane arrangement

$$\bigcup_{\kappa=1}^{K_o} \bigcup_{y \in \mathbb{Z}_+^{\bar{m}}} \big\{T_\kappa x \,+\, H_{\kappa y}\big\} \tag{20}$$

arising in Lemma 3.5. Associated with that arrangement there are the affine halfspaces

$$T_\kappa x \,+\, H^-_{\kappa y} \quad \text{and} \quad T_\kappa x \,+\, H^+_{\kappa y}, \;\; \kappa = 1, \ldots, K_o, \; y \in \mathbb{Z}_+^{\bar{m}}, \tag{21}$$

where $H^-_{\kappa y}$ is defined as in (18) and, accordingly, $H^+_{\kappa y} := \{t \in \mathbb{R}^s : \delta_\kappa^T t \geq \bar{\delta}_{\kappa y}\}$.

Let $\Pi(x)$ denote the family of all, not necessarily full-dimensional, polyhedra in $\mathbb{R}^s$ arising as intersections of halfspaces from (21). In the proof of Lemma 3.5 we have seen that the complement of $M(x)$ is a countable union of polyhedra, each arising as an intersection of halfspaces from (21). Thus, the set $M(x)$ admits a representation

$$M(x) \;=\; \bigcup_{\iota=1}^{\infty} P_\iota(x) \tag{22}$$

such that $P_{\iota_1}(x) \cap P_{\iota_2}(x) = \emptyset$ whenever $\iota_1 \neq \iota_2$, and, for all $\iota \in \mathbb{N}$, the closure $\operatorname{cl} P_\iota(x)$ belongs to $\Pi(x)$.

Consider the outer normals δ_κ and $-\delta_\kappa$ of the affine halfspaces $H^-_{\kappa y}$ and $H^+_{\kappa y}$, respectively. By $\mathcal{B}_o$ we denote the family of all subsets in $\mathbb{R}^s$ which are given as intersections of affine halfspaces with outer normals in $\{\pm\delta_\kappa \,:\, \kappa = 1, \ldots, K_o\}$. Clearly, $\Pi(x) \subseteq \mathcal{B}_o$ for all $x \in \mathbb{R}^m$, provided that the setting of Lemma 3.5 is adopted.

The representation (22) now gives rise to the following variational distance of probability measures in $\mathcal{P}(\mathbb{R}^s)$:

$$\alpha_{\mathcal{B}_o}(\mu, \nu) \;:=\; \sup\big\{|\mu(B) - \nu(B)| \;:\; B \in \mathcal{B}_o\big\}. \tag{23}$$

We further introduce

$$\Delta_{a,C}(\mathbb{R}^s) \;:=\; \left\{\nu \in \mathcal{P}(\mathbb{R}^s) \;:\; \int_{\mathbb{R}^s} \|h\|^a \, \nu(dh) \;\leq\; C\right\},$$

where $a > 0$ and $C > 0$ are fixed constants.

PROPOSITION 3.12. *Assume that $W(\mathbb{Z}_+^{\bar{m}}) + W'(\mathbb{R}_+^{m'}) = \mathbb{R}^s$ and $\{u \in \mathbb{R}^s \;:\; W^T u \leq q, \; W'^T u \leq q'\} \neq \emptyset$, and that q, q' are rational vectors. Then there exists a constant $\mathtt{L}_o > 0$ such that*

$$|Q_{\mathbb{P}}(x, \mu) \;-\; Q_{\mathbb{P}}(x, \nu)| \;\leq\; \mathtt{L}_o \cdot \alpha_{\mathcal{B}_o}(\mu, \nu)^{\frac{a}{s+a}}$$

for all $x \in \mathbb{R}^m$ and all $\mu, \nu \in \Delta_{a,C}(\mathbb{R}^s)$ with $\alpha_{\mathcal{B}_o}(\mu, \nu) \neq 0$.

Proof. Let $x \in \mathbb{R}^m$ and $\mu, \nu \in \Delta_{a,C}(\mathbb{R}^s)$ such that $\alpha_{\mathcal{B}_o}(\mu,\nu) \neq 0$. With $R := \alpha_{\mathcal{B}_o}(\mu,\nu)^{-\frac{1}{s+a}}$ we consider the ball $B_R := \{h \in \mathbb{R}^s : \|h\| \leq R\}$. Recall the representation (22) for $M(x)$ and denote $\mathbb{N}_o := \{\iota \in \mathbb{N} : P_\iota(x) \cap B_R \neq \emptyset\}$. Then it holds that

$$\begin{aligned} |Q_{\mathbb{P}}(x,\mu) - Q_{\mathbb{P}}(x,\nu)| &= |\mu(M(x)) - \nu(M(x))| \\ &\leq \sum_{\iota\in\mathbb{N}} |\mu(P_\iota(x)) - \nu(P_\iota(x))| \\ &\leq \sum_{\iota\in\mathbb{N}_o} |\mu(P_\iota(x)) - \nu(P_\iota(x))| \ + \ (\mu+\nu)\big(\{h \in \mathbb{R}^s \ : \ \|h\| \geq R\}\big) \\ &\leq \sum_{\iota\in\mathbb{N}_o} |\mu(P_\iota(x)) - \nu(P_\iota(x))| \ + \ \frac{2C}{R^a}, \end{aligned} \tag{24}$$

where Markov's inequality has been used in the last estimate. We continue by estimating the cardinality of $\mathbb{N}_o$. Due to Lemma 3.5(ii) there exists a constant $c_1 > 0$, which does not depend on x, such that at most $c_1 \cdot R$ hyperplanes from the arrangement in (20) intersect the ball B_R. From the theory of hyperplane arrangements it is known that the complement of an arrangement of N hyperplanes in $\mathbb{R}^s$ consists of at most $\sum_{i=0}^{s} \binom{N}{i} = O(N^s)$ connected cells of dimension s; see [12, 33]. Hence there exists a constant $c_2 > 0$ such that at most $c_2 \cdot R^s$ full-dimensional sets $P_\iota(x)$ intersect the ball B_R. Since there are only finitely many normals $\delta_\kappa, \kappa = 1, \ldots, K_o$, in the arrangement (20), there exists a constant $c_3 > 0$, again not depending on x, such that the number of all (not necessarily full-dimensional) sets $P_\iota(x)$ intersecting the ball B_R, i.e., the cardinality of $\mathbb{N}_o$, is bounded above by $c_3 \cdot R^s$.

For any $\iota \in \mathbb{N}_o$ we have the estimate

$$|\mu(P_\iota(x)) - \nu(P_\iota(x))| \ \leq \ \alpha_{\mathcal{B}_o}(\mu,\nu),$$

where, in case $P_\iota(x)$ is not closed, $P_\iota(x)$ is approximated by a sequence that is monotonically increasing with respect to set inclusion and that consists of closed polyhedra from $\mathcal{B}_o$ which are contained in the relative interior of $P_\iota(x)$. Altogether, this allows us to continue the estimate (24) as follows:

$$\begin{aligned} &\leq c_3 \cdot R^s \cdot \alpha_{\mathcal{B}_o}(\mu,\nu) \ + \ \frac{2C}{R^a} \\ &\leq c_3 \cdot \alpha_{\mathcal{B}_o}(\mu,\nu)^{-\frac{s}{s+a}+1} \ + \ 2C \cdot \alpha_{\mathcal{B}_o}(\mu,\nu)^{\frac{a}{s+a}} \ = \ (c_3 + 2C) \cdot \alpha_{\mathcal{B}_o}(\mu,\nu)^{\frac{a}{s+a}}, \end{aligned}$$

and the proof is complete. □

Remark 3.13. In general, $\alpha_{\mathcal{B}_o}$ need not define a metric on $\mathcal{P}(\mathbb{R}^s)$ since $\alpha_{\mathcal{B}_o}(\mu,\nu) = 0$ is possible with $\mu \neq \nu$. This can be overcome by enriching $\mathcal{B}_o$, for instance by adding the canonical basis vectors in $\mathbb{R}^s$ to the set of relevant outer normals in the definition of $\mathcal{B}_o$. Then $\alpha_{\mathcal{B}_o}$ majorizes the uniform distance of distribution functions which is known to be a metric on $\mathcal{P}(\mathbb{R}^s)$, and Proposition 3.12 holds without the restriction that $\alpha_{\mathcal{B}_o}(\mu,\nu) \neq 0$.

Remark 3.14. Under suitable assumptions, weak convergence of probability measures implies convergence in $\alpha_{\mathcal{B}_o}$, and Proposition 3.12 can be seen as a quantification of Proposition 3.11: A class $\mathcal{B}$ of Borel sets in $\mathbb{R}^s$ is called a μ-uniformity class if $\alpha_{\mathcal{B}}(\mu_n,\mu) \to 0$ holds for every sequence μ_n in $\mathcal{P}(\mathbb{R}^s)$ converging weakly to $\mu \in \mathcal{P}(\mathbb{R}^s)$. By Theorem 2.11 in [8] the family of all convex Borel sets in $\mathbb{R}^s$ is a μ-uniformity

class, provided that μ has a density. Since all members of $\mathcal{B}_o$ are convex and Borel, weak convergence of probability measures implies convergence in $\alpha_{\mathcal{B}_o}$ if the limiting measure has a density.

Remark 3.15. The fact that the number of affine halfspaces defining the members of $\mathcal{B}_o$ is uniformly bounded implies speed-of-convergence estimates for $\alpha_{\mathcal{B}_o}$ in the context of estimation by empirical measures. Given a sequence $\xi_1, \xi_2, \ldots, \xi_n, \ldots$ of independent $\mathbb{R}^s$-valued random variables on some probability space $(\Omega, \mathcal{A}, \mathsf{P})$ with joint distribution μ, the empirical measures $\mu_n(\omega)$ ($\omega \in \Omega$, $n \in \mathbb{N}$) are defined by

$$\mu_n(\omega) = \frac{1}{n} \sum_{i=1}^{n} \delta_{\xi_i(\omega)},$$

where $\delta_{\xi_i(\omega)}$ denotes the measure with unit mass at $\xi_i(\omega)$ (cf. [15, 34, 47]). For the variational distance $\alpha_{\mathcal{B}}(\mu, \mu_n(\omega))$ then the following law of iterated logarithm established in [26] holds, provided that $\mathcal{B}$ is a so-called Vapnik–Červonenkis class:

$$\limsup_{n\to\infty} \Big(\frac{n}{2 \log\log n}\Big)^{1/2} \cdot \alpha_{\mathcal{B}}(\mu, \mu_n(\omega)) \le \frac{1}{2} \qquad \text{for } \mathsf{P}\text{-almost all } \omega \in \Omega. \tag{25}$$

A family $\mathcal{B}$ of Borel sets in $\mathbb{R}^s$ is called a Vapnik–Červonenkis class if there exists an $m_o \in \mathbb{N}$ such that for any finite set $E \subset \mathbb{R}^s$ with m_o elements not every subset E_o of E arises as an intersection $E_o = E \cap B$ for some $B \in \mathcal{B}$. The catch is now that, thanks to the uniform bound on the number of defining halfspaces, the family $\mathcal{B}_o$ is a Vapnik–Červonenkis class; for details see, e.g., [34, 47]. Proposition 3.12 and (25) then provide a speed-of-convergence estimate for $|Q_{\mathbb{P}}(x, \mu) - Q_{\mathbb{P}}(x, \mu_n(\omega))|$.

Propositions 3.11 and 3.12 are the essential ingredients for studying the stability of optimal solutions to optimization problems whose objective function involves the excess probability functional $Q_{\mathbb{P}}$. Stability of the traditional expectation-based stochastic program (4) was studied in [1, 36, 45, 46]. We will close this section with some stability results for the risk minimization problem

$$P(\mu) \qquad \min\{Q_{\mathbb{P}}(x, \mu) \ : \ x \in X\}.$$

This specific problem has been chosen to display the direct impact of Propositions 3.11 and 3.12 on stability. If one is interested in the stability of the mean-risk model (6) one has to combine the assumptions in the statements below with assumptions in [1, 36, 45, 46].

In general, the function $Q_{\mathbb{P}}(., \mu)$ is nonconvex such that an analysis of local optimal solutions is appropriate. To this end we follow [40, 24] and consider localized optimal values and solution sets. With some subset $V \subset \mathbb{R}^m$ we define

$$\begin{aligned} \varphi_V(\mu) &:= \inf\{Q_{\mathbb{P}}(x, \mu) \ : \ x \in X \cap \operatorname{cl} V\}, \\ \Psi_V(\mu) &:= \{x \in X \cap \operatorname{cl} V \ : \ Q_{\mathbb{P}}(x, \mu) = \varphi_V(\mu)\}. \end{aligned}$$

Given $\mu \in \mathcal{P}(\mathbb{R}^s)$, a nonempty set $Z \subset \mathbb{R}^m$ is called a complete local minimizing (CLM) set of $P(\mu)$ with respect to V if V is open and $Z = \Psi_V(\mu) \subset V$. Examples for CLM sets are the set of global minimizers and isolated local minimizers. The basic feature of CLM sets is that they contain all local minimizers "nearby." Without such a completeness property, pathologies may occur under perturbations; see [40, 24] for details.

PROPOSITION 3.16. *Assume that* $W(\mathbb{Z}_+^{\bar{m}}) + W'(\mathbb{R}_+^{m'}) = \mathbb{R}^s$ *and* $\{u \in \mathbb{R}^s : W^T u \leq q,\ W'^T u \leq q'\} \neq \emptyset$, *that* q, q' *are rational vectors, and that* $\mu \in \mathcal{P}(\mathbb{R}^s)$ *has a density. Suppose further that there exists a subset* $Z \subset \mathbb{R}^m$ *which is a CLM set for* $P(\mu)$ *with respect to some bounded open set* $V \subset \mathbb{R}^m$. *Then it holds that*

(i) *the function* $\varphi_V : \mathcal{P}(\mathbb{R}^s) \longrightarrow \mathbb{R}$ *is continuous at* μ, *where* $\mathcal{P}(\mathbb{R}^s)$ *is equipped with weak convergence of probability measures;*

(ii) *the multifunction* $\Psi_V : \mathcal{P}(\mathbb{R}^s) \longrightarrow 2^{\mathbb{R}^m}$ *is Berge upper semicontinuous at* μ; *i.e., for any open set* $\mathcal{O}$ *in* $\mathbb{R}^m$ *with* $\mathcal{O} \supseteq \Psi_V(\mu)$ *there exists a neighborhood* $\mathcal{N}$ *of* μ *in* $\mathcal{P}(\mathbb{R}^s)$, *again equipped with the topology of weak convergence of probability measures, such that* $\Psi_V(\nu) \subseteq \mathcal{O}$ *for all* $\nu \in \mathcal{N}$;

(iii) *there exists a neighborhood* $\mathcal{N}'$ *of* μ *in* $\mathcal{P}(\mathbb{R}^s)$ *such that for all* $\nu \in \mathcal{N}'$ *the set* $\Psi_V(\nu)$ *is a CLM set for* $P(\nu)$ *with respect to* V;

(iv) *there exists a constant* $\mathtt{L}_o > 0$ *such that*

$$|\varphi_V(\mu) - \varphi_V(\nu)| \ \leq \ \mathtt{L}_o \cdot \alpha_{\mathcal{B}_o}(\mu, \nu)^{\frac{a}{s+a}}$$

for all $\mu, \nu \in \Delta_{a,C}(\mathbb{R}^s)$ *with* $\alpha_{\mathcal{B}_1}(\mu, \nu) \neq 0$.

Before proving the above proposition let us add a few comments. The above assertions are paradigmatic statements in the stability analysis of nonconvex optimization problems. Their proofs rely on well-established arguments that date back (at least) to Berge [7] and that were adapted and extended by many authors; cf. [3, 41], for instance. The main ingredients to make these arguments work are qualitative and quantitative continuity properties as established in Propositions 3.11 and 3.12 together with nonemptiness and compactness of the unperturbed solution set that, in Proposition 3.16, is hidden in the boundedness assumption on V. Therefore, we will refrain from presenting all details of the proof and merely outline its main ideas.

Proof. Using the joint continuity established in Proposition 3.11 the proof of (i) and (ii) follows the lines of Berge's theory as displayed, for instance, in the proof of Theorem 4.2.2 in [3].

To prove (iii), one first confirms the nonemptiness of $\Psi_V(\nu)$, which is a consequence of the lower semicontinuity of $Q_{\mathbb{P}}(., \nu)$; see Proposition 3.3, together with the nonemptiness and compactness of $X \cap \mathrm{cl}\, V$. The CLM property then follows from (ii).

For proving (iv) we, as in the proof of (iii), confirm that $\Psi_V(\mu) \neq \emptyset$ and $\Psi_V(\nu) \neq \emptyset$. Let $\mu, \nu \in \Delta_{a,C}(\mathbb{R}^s)$ and $x_\nu \in \Psi_V(\nu)$, $x_\mu \in \Psi_V(\mu)$. Then it holds that

$$\varphi_V(\mu) \ \leq \ Q_{\mathbb{P}}(x_\nu, \mu) \ \leq \ \varphi_V(\nu) + |Q_{\mathbb{P}}(x_\nu, \mu) - Q_{\mathbb{P}}(x_\nu, \nu)|$$

and

$$\varphi_V(\nu) \ \leq \ Q_{\mathbb{P}}(x_\mu, \nu) \ \leq \ \varphi_V(\mu) + |Q_{\mathbb{P}}(x_\mu, \nu) - Q_{\mathbb{P}}(x_\mu, \mu)|.$$

Together with Proposition 3.12 this implies that

$$|\varphi_V(\nu) - \varphi_V(\mu)| \ \leq \ \mathtt{L}_o \cdot \alpha_{\mathcal{B}_o}(\nu, \mu)^{\frac{a}{s+a}},$$

and the proof is complete. □

Remark 3.17. Due to the lower semicontinuity of $Q_{\mathbb{P}}(., \nu)$ and the fact that $X \cap \mathrm{cl} V$ is nonempty and compact, nonemptiness of $\Psi_V(\nu)$ is immediate. Not immediate, however, is that $\Psi_V(\nu)$ consists of local minimizers to $P(\nu)$, i.e., when minimizing over X. The latter is confirmed by assertion (iii) above, which says that for all $\nu \in \mathcal{N}'$ the set $\Psi_V(\nu)$ is a CLM set, and hence a set of local minimizers to $P(\nu)$.

4. Algorithm and computational experiments. The following result establishes a useful link between two-stage stochastic programs with excess probabilities and traditional expectation-based two-stage models. Before stating the proposition we recall that the support $\operatorname{supp}\mu$ of $\mu \in \mathcal{P}(\mathbb{R}^s)$ is the smallest closed subset of $\mathbb{R}^s$ with μ-measure 1.

PROPOSITION 4.1. *Assume that* $W(\mathbb{Z}_+^{\bar{m}}) + W'(\mathbb{R}_+^{m'}) = \mathbb{R}^s$, $\{u \in \mathbb{R}^s \ : \ W^T u \leq q,\ W'^T u \leq q'\} \neq \emptyset$, *and that* μ *has bounded support. Then the following holds.*

(i) *For any* $x \in \mathbb{R}^m$ *there exists a constant* $\mathtt{M}_x > 0$ *such that*

$$Q_{\mathbb{P}}(x) \ = \ \tilde{Q}_{\mathbb{E}}(x) \ := \ \int_{\mathbb{R}^s} \tilde{\Phi}(h - Tx, c^T x - \varphi_o)\, \mu(dh), \tag{26}$$

where

$$\begin{aligned}\tilde{\Phi}(t_1, t_2) \ := \ \min\{\theta : Wy + W'y' \ = \ t_1, \ -q^T y - q'^T y' + (\mathtt{M}_x - \varphi_o)\theta \ \geq \ t_2,\\ y \in \mathbb{Z}_+^{\bar{m}},\ y' \in \mathbb{R}_+^{m'},\ \theta \in \{0,1\}\}.\end{aligned} \tag{27}$$

(ii) *If, in addition,* X *is bounded, then* $\mathtt{M}_x$ *in* (i) *can be chosen as a uniform constant* $\mathtt{M}$ *for all* $x \in X$, *and the stochastic programs*

$$\min\{Q_{\mathbb{P}}(x) \ : \ x \in X\} \quad \textit{and} \quad \min\{\tilde{Q}_{\mathbb{E}}(x) \ : \ x \in X\}$$

are equivalent.

Proof. To prove (i) we define

$$\mathtt{M}_x \ := \ \sup\{c^T x + \Phi(h - Tx) \ : \ h \in \operatorname{supp}\mu\}.$$

This supremum is finite since, by Proposition 3.1(iv) and $\Phi(0) = 0$, it holds that

$$|\Phi(h - Tx)| \ = \ |\Phi(h - Tx) - \Phi(0)| \ \leq \ \beta\|h\| + \beta\|Tx\| + \gamma, \tag{28}$$

and since $\operatorname{supp}\mu$ is bounded.

Next we verify that the integral in (26) is taken over a real-valued function, i.e., that $\tilde{\Phi}(h - Tx, c^T x - \varphi_o) \in \mathbb{R}$ for all $h \in \operatorname{supp}\mu$. By Proposition 3.1(i) it holds that $\Phi(t) \in \mathbb{R}$ for all $t \in \mathbb{R}^s$. Hence, the optimization problem associated with $\Phi(h - Tx)$ (cf. (2)) is solvable for all $h \in \operatorname{supp}\mu$. Let $h \in \operatorname{supp}\mu$ and $(y, y') \in \mathbb{Z}_+^{\bar{m}} \times \mathbb{R}_+^{m'}$ be an optimal solution associated with $\Phi(h - Tx)$. Then we have

$$Wy + W'y' = h - Tx \quad \text{and} \quad \mathtt{M}_x - \varphi_o \ \geq \ c^T x + \Phi(h - Tx) - \varphi_o \ = \ c^T x + q^T y + q'^T y' - \varphi_o.$$

Hence, the tuple $(y, y', 1)$ is feasible for the optimization problem associated with $\tilde{\Phi}(h - Tx, c^T x - \varphi_o)$; cf. (27). Since this optimization problem has an objective with values in $\{0, 1\}$ only, it is solvable, and $\tilde{\Phi}(h - Tx, c^T x - \varphi_o) \in \mathbb{R}$.

The integral in (26) makes sense for measurable functions only. Therefore, we have to show that $\tilde{\Phi}$ is measurable. Since $\tilde{\Phi}$ takes values in the finite set $\{0, 1\}$, it is sufficient to show measurability of the preimages $\tilde{\Phi}^{-1}(\{0\})$ and $\tilde{\Phi}^{-1}(\{1\})$. For these sets we have the following representations:

$$\begin{aligned}&\tilde{\Phi}^{-1}(\{0\})\\ &= \big\{(t_1, t_2) \in \mathbb{R}^{s+1} \ : \ \exists (y, y') \in \mathbb{Z}_+^{\bar{m}} \times \mathbb{R}_+^{m'} \quad Wy + W'y' = t_1,\ q^T y + q'^T y' \leq -t_2\big\}\\ &= \big\{(t_1, t_2) \in \mathbb{R}^{s+1} \ : \ \Phi(t_1) \leq -t_2\big\}\end{aligned}$$

and

$$
\begin{aligned}
&\tilde{\Phi}^{-1}(\{1\}) \\
&= \tilde{\Phi}^{-1}(\{0,1\}) \setminus \tilde{\Phi}^{-1}(\{0\}) \\
&= \Big\{(t_1,t_2) \in \mathbb{R}^{s+1} \,:\, \exists (y,y',\theta) \in \mathbb{Z}_+^{\bar{m}} \times \mathbb{R}_+^{m'} \times \{0,1\} \quad Wy + W'y' = t_1, \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad q^T y + q'^T y' \le -t_2 + (\mathtt{M}_x - \varphi_o)\theta \Big\} \\
&\quad \cap \Big\{(t_1,t_2) \in \mathbb{R}^{s+1} \,:\, \Phi(t_1) > -t_2 \Big\} \\
&= \Big[\big\{(t_1,t_2) \in \mathbb{R}^{s+1} \,:\, \exists (y,y') \in \mathbb{Z}_+^{\bar{m}} \times \mathbb{R}_+^{m'} \quad Wy + W'y' = t_1,\ q^T y + q'^T y' \le -t_2 \big\} \\
&\quad \cup \big\{(t_1,t_2) \in \mathbb{R}^{s+1} \,:\, \exists (y,y') \in \mathbb{Z}_+^{\bar{m}} \times \mathbb{R}_+^{m'} \quad Wy + W'y' = t_1, \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad q^T y + q'^T y' \le -t_2 + \mathtt{M}_x - \varphi_o \big\}\Big] \\
&\quad \cap \Big\{(t_1,t_2) \in \mathbb{R}^{s+1} \,:\, \Phi(t_1) > -t_2 \Big\} \\
&= \{(t_1,t_2) \in \mathbb{R}^{s+1} \,:\, \Phi(t_1) \le -t_2 + \mathtt{M}_x - \varphi_o\} \cap \{(t_1,t_2) \in \mathbb{R}^{s+1} \,:\, \Phi(t_1) > -t_2\} \\
&= \{(t_1,t_2) \in \mathbb{R}^{s+1} \,:\, -t_2 < \Phi(t_1) \le -t_2 + \mathtt{M}_x - \varphi_o\}.
\end{aligned}
$$

In view of Proposition 3.1(i) the function Φ is lower semicontinuous, and hence measurable. The above representations then yield measurability of the sets $\tilde{\Phi}^{-1}(\{0\})$ and $\tilde{\Phi}^{-1}(\{1\})$. Note that in case $\mathtt{M}_x - \varphi_o \le 0$ we have $\tilde{\Phi}^{-1}(\{1\}) = \emptyset$. Since the integrand in (26) is globally bounded on its domain of finiteness, now measurability of $\tilde{\Phi}$ implies that the integral in (26) is well-defined.

To check the asserted equality in (26) we denote $\chi_{M(x)}(h)$ the indicator function of $M(x)$, and we show that $\tilde{\Phi}(h - Tx, c^T x - \varphi_o) = \chi_{M(x)}(h)$ for all $h \in \operatorname{supp}\mu$.

If $\tilde{\Phi}(h - Tx, c^T x - \varphi_o) = 0$, then there exists $(y,y') \in \mathbb{Z}_+^{\bar{m}} \times \mathbb{R}_+^{m'}$ fulfilling $Wy + W'y' = h - Tx$ and $c^T x + q^T y + q'^T y' \le \varphi_o$. Hence $c^T x + \Phi(h - Tx) \le \varphi_o$, implying that $h \notin M(x)$, and we have $\chi_{M(x)}(h) = 0$.

If $\tilde{\Phi}(h - Tx, c^T x - \varphi_o) = 1$, then $c^T x + q^T y + q'^T y' > \varphi_o$ for all $(y,y') \in \mathbb{Z}_+^{\bar{m}} \times \mathbb{R}_+^{m'}$ fulfilling $Wy + W'y' = h - Tx$. Since the optimization problem associated with $\Phi(h - Tx)$ is solvable, it follows that $c^T x + \Phi(h - Tx) > \varphi_o$. Therefore $h \in M(x)$, and we obtain $\chi_{M(x)}(h) = 1$. This completes the proof of (i).

To verify (ii) we observe that the estimate (28) yields a uniform upper bound $\mathtt{M}$ for $\sup\{c^T x + \Phi(h - Tx) \;:\; h \in \operatorname{supp}\mu\}, x \in X$, provided that $\operatorname{supp}\mu$ and X are bounded. Equivalence of the listed stochastic programs then is a direct consequence of (i). □

Remark 4.2. As a particular consequence of Proposition 4.1 we obtain that the stochastic program $\min\{\tilde{Q}_{\mathbb{E}}(x) : x \in X\}$ has relative complete mixed-integer recourse, meaning that for any $h \in \operatorname{supp}\mu$ and any $x \in X$ there exists a feasible tuple (y,y',θ) to the optimization problem associated with $\tilde{\Phi}(h - Tx, c^T x - \varphi_o)$. On the other hand, the stochastic program fails to have complete mixed-integer recourse. Namely, if we fix $t_1 \in \mathbb{R}^s$, consider a sequence $(t_2^n)_{n\in\mathbb{N}}$ tending to $+\infty$, and assume that there exist $(y_n, y'_n, \theta_n) \in \mathbb{Z}_+^{\bar{m}} \times \mathbb{R}_+^{m'} \times \{0,1\}$ such that

$$
Wy_n + W'y'_n = t_1 \quad \text{and} \quad -q^T y_n - q'^T y'_n + (\mathtt{M} - \varphi_o)\theta_n \ge t_2^n;
$$

then we have found $(y_n, y'_n) \in \mathbb{Z}_+^{\bar{m}} \times \mathbb{R}_+^{m'}$ for which $Wy_n + W'y'_n = t_1$ and $q^T y_n + q'^T y'_n \to -\infty$ as $n \to \infty$. This contradicts the fact that $\min\{q^T y + q'^T y' : Wy + W'y' = t_1, (y, y') \in \mathbb{Z}_+^{\bar{m}} \times \mathbb{R}_+^{m'}\}$ is solvable, or, in other words, that $\Phi(t_1) \in \mathbb{R}$; cf. Proposition 3.1(i).

This lack of complete mixed-integer recourse prevented the application of existing results on structure and stability of expectation-based two-stage stochastic integer programs (see [45, 46]) in our analysis of section 3. In addition, of course, Proposition 4.1 requires the underlying probability measures to have bounded support while the analysis of section 3 does not.

In the remainder of this section we assume that the set X is bounded and closed, and arises as a solution set to a system of linear inequalities, possibly involving integer requirements to components of x. Moreover, we assume that the underlying probability measure μ is discrete with finitely many realizations (or scenarios) h_j and probabilities $\pi_j, j = 1, \ldots, J$. Clearly, the support of μ is bounded then, and we obtain the following corollary to Proposition 4.1.

COROLLARY 4.3. *Adopt the setting of Proposition* 4.1 *and let* X, μ *be as above. Then there exists a constant* $\mathtt{M} > 0$ *such that the stochastic program*

$$\min\{Q_{\mathbb{P}}(x) \ : \ x \in X\} \tag{29}$$

can be equivalently restated as

$$\min_{x,y,y',\theta} \left\{ \sum_{j=1}^{J} \pi_j \theta_j \ : \begin{array}{l} Wy_j + W'y'_j \ = \ h_j - Tx, \\ -q^T y_j - q'^T y'_j + (\mathtt{M} - \varphi_o)\theta_j \ \geq \ c^T x - \varphi_o, \\ x \in X, \ \ y_j \in \mathbb{Z}_+^{\bar{m}}, \ y'_j \in \mathbb{R}_+^{m'}, \ \theta_j \in \{0,1\}, \ \ j = 1, \ldots, J \end{array} \right\}. \tag{30}$$

Problem (30) quickly becomes large scale such that general-purpose mixed-integer linear programming algorithms and software fail. On the other hand, the constraint matrix of (30) obeys the same block-angular structure as with traditional expectation-based linear two-stage stochastic programs. Second-stage variables (y_j, y'_j, θ_j) for different scenarios are not linked in explicit constraints but only through the scenario-independent first stage variable x.

In analogy to the traditional expectation-based model (cf. [13]), this suggests the following algorithmic approach to (30) via scenario decomposition, i.e., Lagrangian relaxation of nonanticipativity.

Introduce in (30) copies $x_j, j = 1, \ldots, J$, referring to the number of scenarios, and add the nonanticipativity constraints $x_1 = \cdots = x_J$ (or an equivalent system), for which we use the notation $\sum_{j=1}^{J} H_j x_j = 0$ with proper (l, m)-matrices $H_j, j = 1, \ldots, J$. Problem (30) then becomes

$$\min_{x,y,y',\theta} \left\{ \sum_{j=1}^{J} \pi_j \theta_j \ : \begin{array}{l} Wy_j + W'y'_j \ = \ h_j - Tx_j, \quad \sum_{j=1}^{J} H_j x_j = 0, \\ -q^T y_j - q'^T y'_j + (\mathtt{M} - \varphi_o)\theta_j \ \geq \ c^T x_j - \varphi_o, \\ x_j \in X, \ \ y_j \in \mathbb{Z}_+^{\bar{m}}, \ y'_j \in \mathbb{R}_+^{m'}, \ \theta_j \in \{0,1\}, \ \ j = 1, \ldots, J \end{array} \right\}. \tag{31}$$

The constraint system of (31) can be decoupled by Lagrangian relaxation of the constraints $\sum_{j=1}^{J} H_j x_j = 0$. To this end, we consider for $\lambda \in \mathbb{R}^l$ the functions

$$L_j(x_j, y_j, y_j', \theta_j, \lambda) \ := \ \pi_j \theta_j \ + \ \lambda^T H_j x_j, \ \ j = 1, \ldots, J,$$

and form the Lagrangian

$$L(x, y, y', \theta, \lambda) \ := \ \sum_{j=1}^{J} L_j(x_j, y_j, y_j', \theta_j, \lambda).$$

The Lagrangian dual of (31) reads

$$\max\{D(\lambda) \ : \ \lambda \in \mathbb{R}^l\}, \tag{32}$$

where

$$\begin{aligned} D(\lambda) \ = \ \min\Bigg\{ & \sum_{j=1}^{J} L_j(x_j, y_j, y_j', \theta_j, \lambda) : W y_j + W' y_j' \ = \ h_j - T x_j, \\ & -q^T y_j - q'^T y_j' + (\mathtt{M} - \varphi_o)\theta_j \ \geq \ c^T x_j - \varphi_o, \\ & x_j \in X, \ \ y_j \in \mathbb{Z}_+^{\bar{m}}, \ y_j' \in \mathbb{R}_+^{m'}, \ \theta_j \in \{0,1\}, \ \ j = 1, \ldots, J \Bigg\}. \end{aligned}$$

Separability yields

$$D(\lambda) \ = \ \sum_{j=1}^{J} D_j(\lambda), \tag{33}$$

where

$$\begin{aligned} D_j(\lambda) \ = \ \min\{ & L_j(x_j, y_j, y_j', \theta_j, \lambda) : W y_j + W' y_j' \ = \ h_j - T x_j, \\ & -q^T y_j - q'^T y_j' + (\mathtt{M} - \varphi_o)\theta_j \ \geq \ c^T x_j - \varphi_o, \\ & x_j \in X, \ \ y_j \in \mathbb{Z}_+^{\bar{m}}, \ y_j' \in \mathbb{R}_+^{m'}, \ \theta_j \in \{0,1\}\}. \end{aligned} \tag{34}$$

$D(\lambda)$ is the pointwise minimum of affine functions in λ. Therefore it is piecewise affine and concave. Thus, (32) is a nonsmooth concave maximization (or convex minimization) problem that can be solved by bundle methods from nondifferentiable optimization, for instance by the conic bundle method of [17] or the proximal bundle method of [22, 23]. At each iteration, these methods require the objective value and one subgradient of D. The structure of D (cf. (33)) enables substantial decomposition, since the single-scenario problems (34) can be tackled separately. Their moderate size often allows application of general-purpose mixed-integer linear programming codes.

Altogether, the optimal value z_{LD} of (32) provides a lower bound to the optimal value z of problem (30). From integer programming [30] it is well known that in general one has to live with a positive duality gap. On the other hand, it holds that $z_{LD} \geq z_{LP}$, where z_{LP} denotes the optimal value to the LP relaxation of (30). The lower bound obtained by the above procedure, hence, is never worse than the bound obtained by eliminating the integer requirements.

In Lagrangian relaxation, the results of the dual optimization often provide starting points for heuristics to find promising feasible points. Our relaxed constraints

being very simple ($x_1 = \cdots = x_J$), ideas for such heuristics come up straightforwardly. For example, examine the x_j-components, $j = 1, \ldots, J$, of solutions to (34) for optimal or nearly optimal λ, and decide for the most frequent value arising, or average and round if necessary.

If the heuristic yields a feasible solution to (30), then the objective value of the latter provides an upper bound $\bar{z}$ for z. Together with the lower bound z_{LD} this gives the quality certificate (gap) $\bar{z} - z_{LD}$.

The full algorithm improves this certificate by embedding the procedure described so far into a branch-and-bound scheme for (29) seen as a nonconvex global optimization problem. Let $\mathbf{P}$ denote the list of current problems, and let $z_{LD} = z_{LD}(P)$ denote the Lagrangian lower bound for $P \in \mathbf{P}$. The algorithm then proceeds as follows.

ALGORITHM 4.4.

Step 1 *(Initialization): Set $\bar{z} = +\infty$ and let $\mathbf{P}$ consist of problem* (31).

Step 2 *(Termination): If $\mathbf{P} = \emptyset$, then the solution $\hat{x}$ that yielded $\bar{z} = Q_{\mathbb{P}}(\hat{x})$ is optimal.*

Step 3 *(Node selection): Select and delete a problem P from $\mathbf{P}$ and solve its Lagrangian dual. If the optimal value $z_{LD}(P)$ hereof equals $+\infty$ (infeasibility of a subproblem), then go to Step* 2.

Step 4 *(Bounding): If $z_{LD}(P) \geq \bar{z}$ go to Step* 2 *(this step can be carried out as soon as the value of the Lagrangian dual rises above $\bar{z}$). Consider the following situations:*

1. *The scenario solutions x_j, $j = 1, \ldots, J$, are identical: If $Q_{\mathbb{P}}(x_j) < \bar{z}$, then let $\bar{z} = Q_{\mathbb{P}}(x_j)$ and delete from $\mathbf{P}$ all problems P' with $z_{LD}(P') \geq \bar{z}$. Go to Step* 2.

2. *The scenario solutions x_j, $j = 1, \ldots, J$ differ: Compute the average $\bar{x} = \sum_{j=1}^{J} \pi_j x_j$ and round it by some heuristic to obtain $\bar{x}^R$. If $Q_{\mathbb{P}}(\bar{x}^R) < \bar{z}$, then let $\bar{z} = Q_{\mathbb{P}}(\bar{x}^R)$ and delete from $\mathbf{P}$ all problems P' with $z_{LD}(P') \geq \bar{z}$. Go to Step* 5.

Step 5 *(Branching): Select a component $x_{(k)}$ of x and add two new problems to $\mathbf{P}$ obtained from P by adding the constraints $x_{(k)} \leq \lfloor \bar{x}_{(k)} \rfloor$ and $x_{(k)} \geq \lfloor \bar{x}_{(k)} \rfloor + 1$, respectively (if $x_{(k)}$ is an integer component), or $x_{(k)} \leq \bar{x}_{(k)} - \varepsilon$ and $x_{(k)} \geq \bar{x}_{(k)} + \varepsilon$, respectively, where $\varepsilon > 0$ is a tolerance parameter to have disjoint subdomains. Go to Step* 3.

The algorithm is obviously finite if all x-components have to be integers. (Recall that X is bounded!) If x is a mixed-integer variable some stopping criterion to avoid endless branching on the continuous components has to be employed.

As already mentioned, the algorithm follows the same lines as the algorithm for $\min\{Q_{\mathbb{E}}(x) \ : \ x \in X\}$ developed in [13]. In a straightforward manner this leads to a scenario decomposition algorithm for the mean-risk model $\min\{Q_{\mathbb{E}}(x) + \rho Q_{\mathbb{P}}(x) \ : \ x \in X\}$. At the end of the present section we will report some initial computational experience with this algorithm.

Relations with efficient points in multiobjective optimization. The mean-risk model (6) can be seen as a scalarization of the multiobjective optimization problem

$$\min \big\{ \big(Q_{\mathbb{E}}(x), Q_{\mathbb{P}}(x) \big) \ : \ x \in X \big\}. \tag{35}$$

A common notion of optimality in multiobjective optimization is efficiency (or nondominance). In terms of (35) a point $x^* \in X$ is called efficient if there is no $x \in X$ fulfilling $Q_{\mathbb{E}}(x) \leq Q_{\mathbb{E}}(x^*)$ and $Q_{\mathbb{P}}(x) \leq Q_{\mathbb{P}}(x^*)$, with at least one strict inequality. For basic facts of multiobjective optimization we refer the reader to [5, 25] and the references therein. Given $\rho_1, \rho_2 \in \mathbb{R}_+$, every optimal solution to

$$\min\{\rho_1 Q_{\mathbb{E}}(x) + \rho_2 Q_{\mathbb{P}}(x) \ : \ x \in X\} \tag{36}$$

is efficient. This result enables computation of efficient points by solving scalar optimization problems. In general, only a subset of the efficient points of (35) can be computed via (36). For computing the full efficiency set via (36), additional assumptions, e.g., convexity of the individual objectives and the feasible set, are needed. Although we did not elaborate on this in section 3, it is quite easy to confirm that neither $Q_{\mathbb{E}}$ nor $Q_{\mathbb{P}}$ is convex in general. The following example demonstrates that, indeed, there exist efficient points for (35) not computable as solutions to (36) for any $\rho_1, \rho_2 \in \mathbb{R}_+$.

Example 4.5. We specify (1) and (2) as follows. Let $m = 1, \bar{m} = 1, m' = 2, s = 1$, and $c = 0, T = -1, X = \{\frac{4}{12}, \frac{6}{12}, \frac{7}{12}\}$. The random variable $h(\omega)$ is given by the realizations $0, \frac{4}{12}, \frac{6}{12}$ with the probabilities $\frac{2}{5}, \frac{2}{5}, \frac{1}{5}$. The second stage is defined by

$$\begin{aligned}\Phi(t) &= \min\left\{\frac{1}{2}y_1 + y_1' + y_2' \;:\; y_1 + y_1' - y_2' = t,\, y_1 \in \mathbb{Z}_+,\, (y_1', y_2') \in \mathbb{R}^2_+\right\} \\ &= \min\left\{\frac{1}{2}y_1 + |t - y_1| \;:\; y_1 \in \mathbb{Z}_+\right\}.\end{aligned}$$

Finally, the probability threshold is selected as $\varphi_o = \frac{7}{12}$. We have just three points in X such that the image set $(Q_{\mathbb{E}}, Q_{\mathbb{P}})(X)$ can be computed explicitly. It holds that

$$(Q_{\mathbb{E}}, Q_{\mathbb{P}})(X) = \left\{\left(\frac{35}{60}, 0\right),\ \left(\frac{34}{60}, \frac{2}{5}\right),\ \left(\frac{32}{60}, \frac{3}{5}\right)\right\}.$$

Clearly, all three members of the image set are efficient. One confirms that the point $(\frac{33}{60}, \frac{2}{5})$ is located at the straight line passing through $(\frac{35}{60}, 0)$ and $(\frac{32}{60}, \frac{3}{5})$. Hence there is no straight line supporting (from below) $(Q_{\mathbb{E}}, Q_{\mathbb{P}})(X)$ and passing through $(\frac{34}{60}, \frac{2}{5})$. In other words, the efficient point $(\frac{34}{60}, \frac{2}{5})$ is not computable as a solution to (36) for any $\rho_1, \rho_2 \in \mathbb{R}_+$.

A prominent example of efficiency in the context of mean-risk models is induced by Markowitz's mean-variance model for portfolio selection [28, 43]. The model aims at finding an optimal asset allocation where the quality of the allocation is judged by both expectation and variance of the return. The total return being the sum of individual returns multiplied by the allocation proportions, both expectation and variance of the return are convex functions of the allocation. Hence the full set of efficient points, also called the efficient frontier, can be traced by solving scalarizations as in (36).

Due to lacking convexity and the above example we cannot hope to be able to trace the full efficient set, or efficient frontier, of (35) by solving scalarizations (36). However, Algorithm 4.4 bears the potential of tracing the supported part of the efficient frontier, i.e., those efficient points that arise as optimal solutions to scalarizations (36). To this end, it is sufficient to vary the parameter ρ in (6) within the nonnegative reals. For every individual ρ, Algorithm 4.4 then provides a global solution to the nonconvex optimization problem (6). The numerics of tracing nonsupported parts of efficient frontiers to nonconvex multiobjective optimization problems still is a widely open field; cf. [25] for an account of existing methods.

Modeling background for computational tests. To illustrate our initial computational experience we will report tracing of supported efficient points at an example from chemical engineering. The modeling background is given by a real-life multiproduct batch plant producing expandable polystyrene (EPS). A detailed description of the EPS production process can be found in [16].

The process consists of preparation, polymerization, and finishing. During preparation different kinds of intermediates are produced. In certain mixtures depending on a finite number of recipes the intermediates are fed batchwise into the polymerization reactors. After termination of each polymerization its product is transferred immediately into a mixing tank of a finishing line, leading to a discontinuous inflow into these tanks. Each finishing line further consists of a separation stage where different grain sizes of EPS are separated from each other. These grain sizes are the final product of the process and have to match customer demand. The separation stages are driven continuously. Shut-down and start-up procedures for separation stages are time consuming, expensive, and have to adhere to minimum up- and down-times of the stages.

The EPS process is controlled by fixing starting times and choices of recipes for the polymerizations and by selecting start-up and shut-down times as well as feed rates for the separation stages. A typical scheduling horizon is given by two weeks, with a time discretization into five equidistant intervals. The major source of uncertainty is customer demand. Optimization aims at minimizing a weighted sum of costs caused by running the polymerizations, switching the separation stages, and compensating the deficit between production and customer demand.

The above setting gives rise to different two-stage stochastic integer programs; see [16] for details. For the numerical tests in the present paper we have used a planning model where the first-stage variables are given by the states of the separation stages. This places emphasis on the qualitative aspect that a smooth operation of the EPS process is desired, which is achieved by fixing the states of the most sensible part of the plant as early as possible.

Tracing of supported efficient points. To trace the supported parts of nonconvex, nonconnected efficient frontiers we have formulated instances of the EPS problem with 10, 20, 50, and 100 scenarios. With the mentioned extension of Algorithm 4.4 we then have solved to global optimality instances of the mean-risk model $\min\{Q_{\mathbb{E}}(x) + \rho Q_{\mathbb{P}}(x) : x \in X\}$ for suitable values of ρ.

Tracing starts with $\rho = 0$, i.e., with solving $\min\{Q_{\mathbb{E}}(x) : x \in X\}$. If the optimal solution is unique, then it has to be efficient as well. Since we have no indication about unicity of optimal solutions, we solve the mean-risk model again with "small" ρ, say $\underline{\rho} = 0.001$. If the $Q_{\mathbb{E}}$-value of the optimal solution remains the same, the optimal solution has to be efficient, and there are no further supported efficient points for $0 < \rho < \underline{\rho}$.

An analogous procedure is carried out at the "upper end." We solve $\min\{Q_{\mathbb{P}}(x) : x \in X\}$ and check with a "big" ρ ($\bar{\rho} = 1000$) for efficiency.

Suppose the "lower end" and "upper end" procedures have resulted in two distinct efficient points x', x'' with distinct values of $(Q_{\mathbb{E}}, Q_{\mathbb{P}})(x)$. We calculate the normal vector $(1, \hat{\rho})^T$ of the straight line passing through $(Q_{\mathbb{E}}, Q_{\mathbb{P}})(x')$ and $(Q_{\mathbb{E}}, Q_{\mathbb{P}})(x'')$ and solve $\min\{Q_{\mathbb{E}}(x) + \hat{\rho} Q_{\mathbb{P}}(x) : x \in X\}$. The optimal solution x''' is a supported efficient point. If $Q_{\mathbb{E}}(x''') + \hat{\rho} Q_{\mathbb{P}}(x''')$ coincides with the identical values $Q_{\mathbb{E}}(x') + \hat{\rho} Q_{\mathbb{P}}(x') = Q_{\mathbb{E}}(x'') + \hat{\rho} Q_{\mathbb{P}}(x'')$, then it is clear that, up to equality of the value $Q_{\mathbb{E}}(x) + \hat{\rho} Q_{\mathbb{P}}(x)$, there are no further supported efficient points for $\underline{\rho} < \rho < \bar{\rho}$. Otherwise, the search continues with the intervals $\underline{\rho} \le \rho \le \hat{\rho}$ and $\hat{\rho} \le \rho \le \bar{\rho}$.

The procedure is iterated at subintervals for ρ where further supported efficient points still may be expected. It terminates when such intervals no longer exist. With a discrete probability distribution, $Q_{\mathbb{P}}$ attains only finitely many values, implying that the procedure terminates after finitely many steps.

TABLE 1
Computational results for the EPS problem.

Scenarios	Cont/Int/Bin	Constraints	ρ	$(Q_{\mathbb{E}},\ Q_{\mathbb{P}})$	Time (h:mm)	CPLEX
	500/400/352	1370	0	(114.81, 1.00)	0:11	3.23%
	500/400/362	1380	0.001	(114.81, 1.00)	0:12	3.16%
	500/400/362	1380	44.17	(141.31, 0.40)	0:22	14.17%
10	500/400/362	1380	94.78	(141.31, 0.40)	0:21	7.22%
	500/400/362	1380	246.60	(141.31, 0.40)	0:19	11.77%
	500/400/362	1380	1000	(190.64, 0.20)	0:10	0.67%
	500/400/362	1380	$+\infty$	(302.52, 0.20)	0:01	∞
	1000/800/692	2740	0	(120.50, 1.00)	0:21	4.79%
	1000/800/712	2760	0.001	(120.50, 1.00)	0:21	5.51%
	1000/800/712	2760	66.57	(120.50, 1.00)	0:37	11.87%
	1000/800/712	2760	67.80	(137.14, 0.75)	0:38	12.24%
20	1000/800/712	2760	68.68	(161.18, 0.40)	0:38	12.03%
	1000/800/712	2760	177.13	(161.18, 0.40)	0:37	9.68%
	1000/800/712	2760	833.08	(161.18, 0.40)	0:37	23.56%
	1000/800/712	2760	1000	(244.49, 0.30)	0:34	7.18%
	1000/800/712	2760	$+\infty$	(402.55, 0.30)	0:01	∞
	2500/2000/1712	6850	0	(124.00, 1.00)	0:59	5.77%
	2500/2000/1762	6900	0.001	(124.00, 1.00)	1:00	6.23%
	2500/2000/1762	6900	111.25	(166.27, 0.62)	1:53	29.73%
50	2500/2000/1762	6900	222.05	(166.27, 0.62)	1:40	28.03%
	2500/2000/1762	6900	413.43	(257.23, 0.40)	1:48	22.80%
	2500/2000/1762	6900	1000	(257.23, 0.40)	1:04	24.86%
	2500/2000/1762	6900	$+\infty$	(502.07, 0.40)	0:09	∞
	5000/4000/3412	13700	0	(122.10, 1.00)	2:21	5.98%
	5000/4000/3512	13800	0.001	(122.10, 1.00)	2:14	13.40%
	5000/4000/3512	13800	80.33	(153.43, 0.61)	3:27	31.81%
100	5000/4000/3512	13800	153.57	(153.43, 0.61)	3:09	33.11%
	5000/4000/3512	13800	263.42	(221.92, 0.35)	3:18	∞
	5000/4000/3512	13800	1000	(221.92, 0.35)	1:45	∞
	5000/4000/3512	13800	$+\infty$	(452.39, 0.35)	0:22	∞

Table 1 documents our computations. The parameter `M` (cf. (30)) was put to 1000 in all instances. The threshold values φ_o are $102.52, 102.55, 102.07$, and 102.39 for the 10-, 20-, 50-, and 100-scenario instances, respectively.

The ρ column displays the values that were necessary for the search. Let us explain at the 10-scenario instance: $\rho = 0$ and $\rho = +\infty$ correspond to minimizing $Q_{\mathbb{E}}$ and $Q_{\mathbb{P}}$, respectively. The values $\underline{\rho} = 0.001$ and $\bar{\rho} = 1000$ are the mentioned safeguards for efficiency at the "lower ends" and "upper ends." Simultaneously, they serve to initialize the search interval for ρ. The first value for $\hat{\rho}$ is 94.78. It yields an efficient point whose value of $Q_{\mathbb{E}}(x) + \hat{\rho}Q_{\mathbb{P}}(x)$ is distinct from the corresponding value of the two efficient points already found. Hence, the subintervals $[\underline{\rho}, \hat{\rho}]$ and $[\hat{\rho}, \bar{\rho}]$ must be considered further, what is done with the updated $\hat{\rho}$-values 44.17 and 246.60, respectively. In both cases, the optimal values of $Q_{\mathbb{E}}(x) + \hat{\rho}Q_{\mathbb{P}}(x)$ coincide with $(Q_{\mathbb{E}} + \hat{\rho}Q_{\mathbb{P}})$-values of efficient points already known. (In fact, even the optimal solution points coincide with an efficient point already known.) The search terminates. Altogether, we have found three supported efficient points with $(Q_{\mathbb{E}}, Q_{\mathbb{P}})$-values of $(114.81, 1.00), (141.31, 0.40)$, and $(190.64, 0.20)$. Up to equality of values $Q_{\mathbb{E}}(x) + \hat{\rho}Q_{\mathbb{P}}(x)$ for $\hat{\rho} \in \{44.17, 246.60\}$, these are all supported efficient points of the instance.

The time column displays the time needed by our decomposition algorithm to find an optimal point and prove its optimality with a relative gap of 0.001%.

The CPLEX column shows the relative optimality gaps achieved by CPLEX 8.0 (with default parameters) in the time listed in the column before. The symbol ∞

indicates that no feasible solution was found in this time. All computations were carried out on a Linux PC with an AMD Athlon XP 2200+ processor (1.8 GHz) and 1 GB RAM.

The Cont/Int/Bin column has the numbers of continuous, integer (nonbinary), and binary variables in the block-angular mixed-integer linear programs corresponding to the models. The numbers of constraints are listed in the next column.

Altogether, the table confirms that our algorithm is able to trace with reasonable effort supported parts of efficient frontiers of realistic instances of the nonconvex multiobjective program (35).

Acknowledgments. We wish to thank Christoph Helmberg (Technical University of Chemnitz) for giving us access to the implementation of his conic bundle method. Further thanks are given to Andreas Märkert (University Duisburg-Essen) for stimulating discussions and his support in designing and performing the computational tests. Moreover, we are grateful to two anonymous referees and the editorial board member for their comments that helped to improve the final version of our paper.

REFERENCES

[1] Z. Artstein and R. J-B. Wets, *Stability results for stochastic programs and sensors, allowing for discontinuous objective functions*, SIAM J. Optim., 4 (1994), pp. 537–550.
[2] P. Artzner, F. Delbaen, J.-M. Eber, and D. Heath, *Coherent measures of risk*, Math. Finance, 9 (1999), pp. 203–228.
[3] B. Bank, J. Guddat, D. Klatte, B. Kummer, and K. Tammer, *Non-linear Parametric Optimization*, Akademie-Verlag, Berlin, 1982.
[4] B. Bank and R. Mandel, *Parametric Integer Optimization*, Akademie-Verlag, Berlin, 1988.
[5] H. P. Benson, *Multi-objective optimization: Pareto optimal solutions, properties*, in Encyclopedia of Optimization, Vol. III, C. A. Floudas and P. M. Pardalos, eds., Kluwer Academic Publishers, Dordrecht, The Netherlands, 2001, pp. 489–493.
[6] B. Bereanu, *Minimum risk criterion in stochastic optimization*, Econom. Comput. Econom. Cybernet. Stud. Res., 15 (1981), pp. 31–39.
[7] C. Berge, *Topological Spaces*, Macmillan, New York, 1963.
[8] R. N. Bhattacharya and R. Ranga Rao, *Normal Approximation and Asymptotic Expansions*, Wiley, New York, 1976.
[9] P. Billingsley, *Convergence of Probability Measures*, Wiley, New York, 1968.
[10] J. R. Birge and F. Louveaux, *Introduction to Stochastic Programming*, Springer-Verlag, New York, 1997.
[11] C. E. Blair and R. G. Jeroslow, *The value function of a mixed integer program:* I, Discrete Math., 19 (1977), pp. 121–138.
[12] R. C. Buck, *Partition of space*, Amer. Math. Monthly, 50 (1943), pp. 541–544.
[13] C. C. Carøe and R. Schultz, *Dual decomposition in stochastic integer programming*, Oper. Res. Lett., 24 (1999), pp. 37–45.
[14] *ILOG CPLEX* 8.0*, User's Manual*, ILOG, Inc., Mountain View, CA, 2002. Information available online from http://www.cplex.com.
[15] R. M. Dudley, *Real Analysis and Probability*, Wadsworth and Brooks/Cole, Pacific Grove, CA, 1989.
[16] S. Engell, A. Märkert, G. Sand, R. Schultz, and Ch. Schulz, *Online scheduling of multiproduct batch plants under uncertainty*, in Online Optimization of Large Scale Systems, M. Grötschel, S. O. Krumke, and J. Rambau, eds., Springer-Verlag, Berlin, 2001, pp. 649–676.
[17] C. Helmberg and K. C. Kiwiel, *A spectral bundle method with bounds*, Math. Program., 93 (2002), pp. 173–194.
[18] R. Hemmecke and R. Schultz, *Decomposition of test sets in stochastic integer programming*, Math. Program., 94 (2003), pp. 323–341.
[19] P. Kall and S. W. Wallace, *Stochastic Programming*, Wiley, Chichester, 1994.
[20] A. I. Kibzun and Y. S. Kan, *Stochastic Programming Problems with Probability and Quantile Functions*, Wiley, Chichester, 1996.

[21] A. J. King, S. Takriti, and S. Ahmed, *Issues in Risk Modeling for Multi-Stage Systems*, IBM Research Report, RC-20993, Yorktown Heights, NY, 1997.
[22] K. C. Kiwiel, *Proximity control in bundle methods for convex nondifferentiable optimization*, Math. Programming, 46 (1990), pp. 105–122.
[23] K. C. Kiwiel, *User's Guide for NOA* 2.0/3.0: *A Fortran Package for Convex Nondifferentiable Optimization*, Systems Research Institute, Polish Academy of Sciences, Warsaw, Poland, 1994.
[24] D. Klatte, *On the Stability of Local and Global Optimal Solutions in Parametric Problems of Nonlinear Programming, Part* I *and Part* II, Seminarbericht 75, Sektion Mathematik, Humboldt-Universität zu Berlin, Berlin, Germany, 1985, pp. 1–39.
[25] P. Korhonen, *Multiple objective programming support*, in Encyclopedia of Optimization, Vol. III, C. A. Floudas and P. M. Pardalos, eds., Kluwer Academic Publishers, Dordrecht, The Netherlands, 2001, pp. 566–574.
[26] J. Kuelbs and R. M. Dudley, *Log log laws for empirical measures*, Ann. Probab., 8 (1980), pp. 405–418.
[27] A. Märkert and R. Schultz, *Variance and Two-Stage Stochastic Programs*, manuscript.
[28] H. M. Markowitz, *Portfolio selection*, J. Finance, 7 (1952), pp. 77–91.
[29] J. M. Mulvey, R. J. Vanderbei, and S. A. Zenios, *Robust optimization of large-scale systems*, Oper. Res., 43 (1995), pp. 264–281.
[30] G. L. Nemhauser and L. A. Wolsey, *Integer and Combinatorial Optimization*, Wiley, New York, 1988.
[31] W. Ogryczak and A. Ruszczyński, *From stochastic dominance to mean-risk models: Semideviations as risk measures*, European J. Oper. Res., 116 (1999), pp. 33–50.
[32] W. Ogryczak and A. Ruszczyński, *Dual stochastic dominance and related mean-risk models*, SIAM J. Optim., 13 (2002), pp. 60–78.
[33] P. Orlik and H. Terao, *Arrangements of Hyperplanes*, Springer-Verlag, Berlin, 1992.
[34] D. Pollard, *Convergence of Stochastic Processes*, Springer-Verlag, New York, 1984.
[35] A. Prékopa, *Stochastic Programming*, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1995.
[36] S. T. Rachev and W. Römisch, *Quantitative stability in stochastic programming: The method of probability metrics*, Math. Oper. Res., 27 (2002), pp. 792–818.
[37] E. Raik, *Qualitative research into the stochastic nonlinear programming problems*, Eesti NSV Tead. Akad. Toimetised Füüs.-Mat., 20 (1971), pp. 8–14 (in Russian).
[38] E. Raik, *On the stochastic programming problem with the probability and quantile functionals*, Eesti NSV Tead. Akad. Toimetised Füüs.-Mat., 21 (1971), pp. 142–148 (in Russian).
[39] M. Riis and R. Schultz, *Applying the minimum risk criterion in stochastic recourse programs*, Comput. Optim. Appl., 24 (2003), pp. 267–287.
[40] S. M. Robinson, *Local epi-continuity and local optimization*, Math. Programming, 37 (1987), pp. 208–222.
[41] R. T. Rockafellar and R. J-B. Wets, *Variational Analysis*, Springer-Verlag, Berlin, 1998.
[42] W. Römisch and R. Schultz, *Stability analysis for stochastic programs*, Ann. Oper. Res., 30 (1991), pp. 241–266.
[43] A. Ruszczyński and R. J. Vanderbei, *Frontiers of stochastically nondominated portfolios*, Econometrica, to appear.
[44] R. Schultz, *Continuity properties of expectation functions in stochastic integer programming*, Math. Oper. Res., 18 (1993), pp. 578–589.
[45] R. Schultz, *On structure and stability in stochastic programs with random technology matrix and complete integer recourse*, Math. Programming, 70 (1995), pp. 73–89.
[46] R. Schultz, *Rates of convergence in stochastic programs with complete integer recourse*, SIAM J. Optim., 6 (1996), pp. 1138–1152.
[47] G. R. Shorack and J. A. Wellner, *Empirical Processes with Applications to Statistics*, Wiley, New York, 1986.
[48] S. Takriti and S. Ahmed, *On robust optimization of two-stage systems*, Math. Program., to appear.

SIAM J. OPTIM.
Vol. 14, No. 1, pp. 139–172

SEMIDEFINITE PROGRAMMING IN THE SPACE OF PARTIAL POSITIVE SEMIDEFINITE MATRICES*

SAMUEL BURER†

Abstract. We build upon the work of Fukuda et al. [*SIAM J. Optim.*, 11 (2001), pp. 647–674] and Nakata et al. [*Math. Program.*, 95 (2003), pp. 303–327], in which the theory of partial positive semidefinite matrices was applied to the semidefinite programming (SDP) problem as a technique for exploiting sparsity in the data. In contrast to their work, which improved an existing algorithm based on a standard search direction, we present a primal-dual path-following algorithm that is based on a new search direction, which, roughly speaking, is defined completely within the space of partial symmetric matrices. We show that the proposed algorithm computes a primal-dual solution to the SDP problem having duality gap less than a fraction $\varepsilon > 0$ of the initial duality gap in $\mathcal{O}(n \log(\varepsilon^{-1}))$ iterations, where n is the size of the matrices involved. Moreover, we present computational results showing that the algorithm possesses several advantages over other existing implementations.

Key words. semidefinite programming, sparsity, matrix completion, numerical experiments

AMS subject classifications. 90C06, 90C22, 90C51

DOI. 10.1137/S105262340240851X

1. Introduction.

The semidefinite programming (SDP) problem has been studied extensively in recent years, and many different types of algorithms for solving SDPs have been proposed. Various primal-dual interior-point methods for linear programming can be extended to SDP with equivalent iteration complexities, typically $\mathcal{O}(\sqrt{n}\log(\varepsilon^{-1}))$, where n is the size of matrices in the SDP problem and $\varepsilon > 0$ is the desired fractional reduction in the duality gap; for example, see [1, 15, 17, 22, 23, 25, 28, 30, 32]. In practice, these methods have many advantages, including applicability to any standard form SDP, accurate primal-dual optimal solutions in a small number of iterations, and exploitation of sparsity in certain key stages of the algorithm. On the other hand, they also exhibit some notable disadvantages, such as the need to compute, store, and work with dense matrices—in particular, handling the $n \times n$ primal iterate X and the $m \times m$ Schur complement matrix M, where m is the number of linear constraints in the primal SDP, as well as solving the Schur complement equation involving M.

Techniques for dealing with the disadvantages of primal-dual methods have also been developed. For example, to avoid working with the dense matrix X (while maintaining the use of M), Benson, Ye, and Zhang [2] have developed a polynomial-time interior-point method that involves only the dual variables (S, y) and the lower Cholesky factor L of S, since S and L are generally sparse when the SDP data is sparse. In contrast, others have eliminated the need to compute and store M (while maintaining the use of primal-dual iterates (X, S, y)) by using iterative methods such as the preconditioned conjugate gradient method to solve the Schur complement equation (see [20, 27, 29]). When solving the Schur complement equation using an iterative method, however, an inevitable side effect is the increased difficulty of obtaining ac-

*Received by the editors May 29, 2002; accepted for publication (in revised form) January 24, 2003; published electronically July 18, 2003. This work was supported in part by NSF grant CCR-0203426.

http://www.siam.org/journals/siopt/14-1/40851.html

†Department of Management Sciences, University of Iowa, Iowa City, IA 52242-1000 (samuel-burer@uiowa.edu).

curate primal-dual optimal solutions, due to the ill-conditioning of the matrix near optimality.

Other methods, the so-called first-order nonlinear programming algorithms for SDP, depart even more significantly from the standard primal-dual interior-point methods. Generally speaking, these methods solve special classes of SDPs, work in either the primal or dual space, operate on sparse matrices (or compact representations of dense matrices), and sacrifice the underlying theoretical guarantee of polynomial convergence for better opportunities to exploit sparsity and structure. As a result of these algorithmic choices as well as the ill-conditioning that is inherent near optimality, these methods typically can compute optimal solutions of low to medium accuracy in a reasonable balance of iterations and time. See [4, 6, 5, 8, 13, 14] for more information on this class of algorithms.

So far, no one has proposed a method that possesses theoretical polynomial convergence, can solve any standard-form SDP, works in both the primal and dual spaces, and can aggressively exploit sparsity in all stages of computation, including the complete avoidance of dense matrices. In this paper, we propose such a method and explore its theoretical and practical characteristics.

The basic idea of the method presented in this paper is drawn from the recent work of Fukuda et al. [9], in which they show that the theory of partial positive semidefinite matrices can be applied to SDPs to help better take advantage of sparsity. In particular, their "completion method" demonstrates that primal-dual interior-point algorithms can be implemented using a certain "partial" representation of the dense matrix variable X. Computational results given in Nakata et al. [26], which employ the sparse representation of X together with the computation and storage of M in each iteration, indicate the efficiency of the completion method on several classes of problems.

The completion method can be viewed as a computational enhancement of an existing primal-dual path-following implementation that is based on the Helmberg–Rendl–Venderbei–Wolkowicz/Kojima–Shindoh–Hara/Monteiro (or HRVW/KSH/M) search direction (which was first defined in Helmberg et al. [15]). From a theoretical point of view, however, the completion method is not known to converge in polynomial time, with the main obstacle being how to measure the proximity of a partial primal-dual solution to the central path. (See the concluding comments of section 5 in [9], where a polynomial potential-reduction algorithm is discussed but the problem of a path-following algorithm is considered open.) In addition, since the completion method employs the Schur complement matrix M directly, there is a practical limitation to the size of SDP that can be solved by this method. Of course, a simple idea to eliminate the direct use of M would be to use an iterative method to solve the Schur complement equation.

The method of this paper improves upon the completion method of Fukuda et al. in two primary ways. The first is theoretical: the method is a polynomial-time path-following algorithm based entirely on partial positive semidefinite matrices, where the main idea is a reformulation of the central path that yields search directions in the space of partial matrices and that also motivates a new neighborhood of the central path, which has some critical properties when viewed in the context of matrix completion. The second is practical: when the Schur complement equation in our method is solved using an iterative method, our approach provides even more opportunity to take advantage of the sparsity of the SDP data. In section 5, computational results are given to demonstrate this.

Computational results are also given comparing our method with two other successful methods: a primal-dual interior-point method that possesses polynomial convergence but computes and stores both X and M; and a dual-only first-order algorithm that works exclusively with sparse matrices but does not possess polynomial convergence. The overall conclusion of this paper is that our method achieves several advantages that were previously found only in separate algorithms: theoretically strong convergence, applicability to any SDP, a primal-dual framework, and the opportunity to exploit sparsity in all stages of computation.

1.1. Basic notation and terminology. In this paper, $\Re$, $\Re^p$, and $\Re^{p\times q}$ denote the typical Euclidean spaces, and $\mathcal{S}^p$, $\mathcal{S}^p_+$, and $\mathcal{S}^p_{++}$ denote symmetric, symmetric positive semidefinite, and symmetric positive definite matrices, respectively. Lower triangular matrices are denoted by $\mathcal{L}^p$, and those with positive diagonal entries are signified by $\mathcal{L}^p_{++}$. Similarly, we define $\mathcal{U}^p$ and $\mathcal{U}^p_{++}$ for upper triangular matrices.

For any $v \in \Re^p$, v_i is the ith component of v; for any $A \in \Re^{p\times q}$, A_{ij}, $A_{i\cdot}$, and $A_{\cdot j}$ denote the standard subparts of A. For vectors or matrices, the notation $\cdot^T$ denotes the transpose, and for any $A \in \Re^{p\times p}$, $\mathrm{tr}(A) = \sum_{i=1}^p A_{ii}$ denotes the trace function. The standard inner product on $\Re^{p\times q}$ is denoted as $A \bullet B = \mathrm{tr}(A^T B) = \sum_{i=1}^p \sum_{j=1}^q A_{ij}B_{ij}$, and the Frobenius norm, which is induced by the inner product $\bullet$, is denoted $\|A\|_F = (A \bullet A)^{1/2}$. For any $A \in \mathcal{S}^p$ (which necessarily has real eigenvalues), we note that $\mathrm{tr}(A)$ also equals the sum of the eigenvalues of A. The maximum and minimum eigenvalues of A are denoted by $\lambda_{\max}[A]$ and $\lambda_{\min}[A]$. For any $A \in \Re^{p\times q}$, we define the 2-norm of A to be $\|A\| = \sqrt{\lambda_{\max}[A^T A]}$. Some important inequalities are as follows: for all $A, B \in \Re^{p\times q}$, $A \bullet B \le \|A\|_F \|B\|_F$, and for all $A \in \Re^{p\times q}$, $\|A\| \le \|A\|_F$.

If a function f defined between $\Re^p$ and $\Re^q$ is twice differentiable at $v \in \Re^p$, then the first derivative of f at v is denoted as the linear operator $f'(v) : \Re^p \to \Re^q$, which operates on $a \in \Re^p$ as $f'(v)[a]$. In addition, if $q = 1$, then the gradient $\nabla f(v) \in \Re^p$ uniquely satisfies $\nabla f(v)^T a = f'(v)[a]$ for all $a \in \Re^p$. The second derivative of f at v is denoted by the bilinear map $f''(v) : \Re^p \times \Re^p \to \Re^q$, which is written as $f''(v)[a, b]$ for all $(a, b) \in \Re^p \times \Re^p$. If $b = a$, then we write $f''(v)[a]^{(2)}$. If f is also invertible, then f^{-1} denotes its inverse, and if f is linear, then its adjoint is denoted by f^*. In addition, if $p = q$ and f is linear, then f is called positive definite if $v^T f(v) > 0$ for all $v \ne 0$.

The function chol : $\mathcal{S}^p_{++} \to \mathcal{L}^p_{++}$ computes the lower Cholesky factor; that is, $\mathrm{chol}(S) = L$, where $S = LL^T$. Letting M^{-1} denote the matrix inverse of M and inv the matrix inverse function, we have that $\mathrm{inv}'(M)[A] = -M^{-1}AM^{-1}$. We also let argmax denote the unique optimal solution of a given maximization problem, and we define argmin similarly. The matrix I denotes the identity matrix (of appropriate size), and for any $A \in \Re^{p\times p}$, $\mathrm{diag}(A)$ extracts the diagonal of A. Also, for any $M \in \mathcal{S}^p_+$, $M^{1/2}$ denotes the matrix square root of M.

2. The partial SDP problem. We consider the standard-form primal SDP problem

$$(\hat{P}) \qquad \min \left\{ C \bullet \hat{X} : \mathcal{A}(\hat{X}) = b,\ \hat{X} \in \mathcal{S}^n_+ \right\}$$

and its dual

$$(\hat{D}) \qquad \max \left\{ b^T y : \mathcal{A}^*(y) + \hat{S} = C,\ \hat{S} \in \mathcal{S}^n_+ \right\},$$

where the variables are $(\hat{X}, \hat{S}, y) \in \mathcal{S}^n \times \mathcal{S}^n \times \Re^m$ and the data are $C \in \mathcal{S}^n$, $b \in \Re^m$, and $\{A_k\}_{k=1}^m \subset \mathcal{S}^n$. The symbol $\mathcal{A}$ denotes the linear map $\mathcal{A} : \mathcal{S}^n \to \Re^m$ defined by

$\mathcal{A}(\hat{X})_k = A_k \bullet \hat{X}$, and its adjoint $\mathcal{A}^* : \Re^m \to \mathcal{S}^n$ is defined by $\mathcal{A}^*(y) = \sum_{k=1}^m y_k A_k$. We use similar notation for the sets of primal-dual feasible solutions and primal-dual interior feasible solutions as in [22]—$\mathcal{F}(\hat{P}) \times \mathcal{F}(\hat{D})$ and $\mathcal{F}^0(\hat{P}) \times \mathcal{F}^0(\hat{D})$, respectively—and we also make the following standard assumptions:

Â1. the matrices $\{A_k\}_{k=1}^m$ are linearly independent;

Â2. the set of primal-dual interior feasible solutions $\mathcal{F}^0(\hat{P}) \times \mathcal{F}^0(\hat{D})$ is nonempty.

It is well known that, under assumptions Â1 and Â2, both $(\hat{P})$ and $(\hat{D})$ have optimal solutions $\hat{X}^*$ and $(\hat{S}^*, y^*)$, which are characterized by the equivalent conditions that the duality gap $\hat{X}^* \bullet \hat{S}^*$ is zero and that the matrix product $\hat{X}^* \hat{S}^*$ is zero. Moreover, for every $\nu > 0$, there exists a unique primal-dual feasible solution $(\hat{X}_\nu, \hat{S}_\nu, y_\nu)$, which satisfies the perturbed optimality equation $\hat{X}\hat{S} = \nu I$. The set of all solutions $\hat{\mathcal{C}} \equiv \{(\hat{X}_\nu, \hat{S}_\nu, y_\nu) : \nu > 0\}$ is known as the central path, and $\hat{\mathcal{C}}$ serves as the basis for path-following algorithms that solve $(\hat{P})$ and $(\hat{D})$. The basic idea is to construct a sequence $\{(\hat{X}^k, \hat{S}^k, y^k)\}_{k \geq 0} \subset \mathcal{F}^0(\hat{P}) \times \mathcal{F}^0(\hat{D})$ that stays in a neighborhood of the central path such that the duality gap $\hat{X}^k \bullet \hat{S}^k$ goes to zero.

A scaled measure of the duality gap that proves useful in the presentation and analysis of path-following algorithms is

$$\mu(\hat{X}, \hat{S}) \equiv \frac{\hat{X} \bullet \hat{S}}{n} \qquad \forall \ (\hat{X}, \hat{S}) \in \mathcal{S}^n \times \mathcal{S}^n. \tag{2.1}$$

Note that, for all $(\hat{X}, \hat{S}) \in \mathcal{S}^n_+ \times \mathcal{S}^n_+$, we have $\mu(\hat{X}, \hat{S}) > 0$ unless $\hat{X}\hat{S} = 0$. Moreover, $\mu(\hat{X}_\nu, \hat{S}_\nu) = \nu$ for all points $(\hat{X}_\nu, \hat{S}_\nu, y_\nu)$ on the central path.

2.1. The positive semidefinite matrix completion. Recently, Fukuda et al. [9] introduced techniques for exploiting sparsity using ideas from the theory of matrix completions. In this section, we recapitulate their main results and introduce corresponding notation that will be used throughout the paper.

Let $V = \{1, \ldots, n\}$ denote the row and column indices of an $n \times n$ matrix. Also define the *aggregate density pattern* E of the data $\{C\} \cup \{A_k\}_{k=1}^m$ as follows:

$$E \equiv \{(i,j) \in V \times V : \ \exists \ \ y \in \Re^m \ \text{ such that } \ [C - \mathcal{A}^*(y)]_{ij} \neq 0\}.$$

We assume throughout that $\{(i,i) : i \in V\} \subseteq E$, that is, that E contains all of the diagonal entries. Notice also that E is symmetric in the sense that $(i,j) \in E$ if and only if $(j,i) \in E$ because, by definition, $C - \mathcal{A}^*(y) \in \mathcal{S}^n$. (We also remark that the alternative terminology, "aggregate sparsity pattern," has been used in [9] to describe E.)

Given any $(\hat{S}, y) \in \mathcal{F}(\hat{D})$, it is clear from the definition of E that those elements of $V \times V$ that correspond to the nonzeros of $\hat{S}$ are contained in E. Hence, $\bar{E} \equiv V \times V \setminus E$ represents the generic sparsity pattern of the variable $\hat{S}$ of $(\hat{D})$. Unlike $\hat{S}$, the variable $\hat{X}$ of $(\hat{P})$ has no sparsity in general, but the sparsity represented by $\bar{E}$ does affect the primal problem in terms of evaluation of the objective function $C \bullet \hat{X}$ and the constraints $\mathcal{A}(\hat{X})$. In particular, it is not difficult to see that the quantities $C \bullet \hat{X}$ and $A_k \bullet \hat{X}$ are dependent upon only those entries $\hat{X}_{ij}$ of $\hat{X}$, where $(i,j) \in E$. In other words, the entries $\hat{X}_{ij}$ for $(i,j) \in \bar{E}$ are irrelevant for the objective function and constraints, but still, they do impact the positive semidefiniteness constraint $\hat{X} \in \mathcal{S}^n_+$. These were precisely the observations that were exploited in [9], as we detail next.

Given a symmetric $G \subseteq V \times V$, we define the following subset of $\mathcal{S}^n$, which has the density pattern G:

$$\mathcal{S}^G \equiv \{\hat{M} \in \mathcal{S}^n : \hat{M}_{ij} = 0 \ \forall \ (i,j) \notin G\}.$$

We also define the corresponding operator $\pi^G : \mathcal{S}^n \to \mathcal{S}^G$, which performs orthogonal projection onto $\mathcal{S}^G$:

$$[\pi^G(\hat{M})]_{ij} = \begin{cases} \hat{M}_{ij}, & (i,j) \in G, \\ 0, & (i,j) \notin G. \end{cases}$$

We then define the following subsets of $\mathcal{S}^G$:

$$\begin{aligned} \mathcal{S}^G_+ &= \mathcal{S}^G \cap \mathcal{S}^n_+, \\ \mathcal{S}^G_{++} &= \mathcal{S}^G \cap \mathcal{S}^n_{++}, \\ \mathcal{S}^{G?}_+ &= \{M \in \mathcal{S}^G : \exists\ \hat{M} \in \mathcal{S}^n_+ \text{ such that } \pi^G(\hat{M}) = M\}, \\ \mathcal{S}^{G?}_{++} &= \{M \in \mathcal{S}^G : \exists\ \hat{M} \in \mathcal{S}^n_{++} \text{ such that } \pi^G(\hat{M}) = M\}. \end{aligned}$$

In words, we describe the last two sets defined above as follows: $\mathcal{S}^{G?}_+$ and $\mathcal{S}^{G?}_{++}$ consist of those matrices in $\mathcal{S}^G$ that can be completed to matrices in $\mathcal{S}^n_+$ and $\mathcal{S}^n_{++}$, respectively. We use the question mark (?) notation to illustrate the informal idea that, for example, $M \in \mathcal{S}^{G?}_+$ is a positive semidefinite matrix except that the entries M_{ij} for $(i,j) \notin G$ have yet to be specified. In addition, an important observation is that $\mathcal{S}^{G?}_{++}$ is an open subset of $\mathcal{S}^G$, which will play an important role when we investigate the derivatives of functions defined on $\mathcal{S}^{G?}_{++}$.

Using these ideas from matrix completion along with the discussion of E above, it is not difficult to see that problems $(\hat{P})$ and $(\hat{D})$ are equivalent to the following two problems, respectively:

$$\min\left\{C \bullet X : \mathcal{A}(X) = b,\ X \in \mathcal{S}^{E?}_+\right\}, \qquad \max\left\{b^T y : \mathcal{A}^*(y) + S = C,\ S \in \mathcal{S}^E_+\right\}.$$

It is interesting to note that the above equivalence holds even when E is replaced by any symmetric $F \supseteq E$. In fact, for technical reasons that will become clear later, it is desirable to apply this idea with an F that satisfies specific structural properties, as discussed next.

It is straightforward to identify a symmetric $G \subseteq V \times V$ with a simple graph $\tilde{G}$ on V, and we make the following graph theoretic definitions. G is said to be *chordal* if $\tilde{G}$ is chordal, that is, if every cycle in $\tilde{G}$ having length greater than three has a chord. A *perfect elimination ordering* for G is an ordering $(v_1, \ldots, v_n)$ of the vertices $V = \{1, \ldots, n\}$ of $\tilde{G}$ such that, for each $1 \le i \le n-1$, the $\tilde{G}$-neighbors of v_i in $\{v_{i+1}, \ldots, v_n\}$ form a clique in $\tilde{G}$. A fundamental fact (see Fulkerson and Gross [10]) is that G is chordal if and only if it has a perfect elimination ordering.

Now let F be a symmetric extension of E, i.e., $F \supseteq E$, that satisfies two properties: (i) F is chordal; and (ii) the standard ordering $(1, \ldots, n)$ is a perfect elimination ordering for F. We then define the pair of SDP problems

$$(P) \qquad \min\left\{C \bullet X : \mathcal{A}(X) = b,\ X \in \mathcal{S}^{F?}_+\right\}$$

and

$$(D) \qquad \max\left\{b^T y : \mathcal{A}^*(y) + S = C,\ S \in \mathcal{S}^F_+\right\},$$

which, from the discussion above, are equivalent to $(\hat{P})$ and $(\hat{D})$, respectively. It is worthwhile to note that, under the assumption that no numerical cancellations occur during the calculation of the lower Cholesky factor $L \in \mathcal{L}^n_{++}$ of $S \in \mathcal{S}^E_{++}$,

the symmetrized density pattern of L yields a chordal extension F of E such that $(1,\dots,n)$ is a perfect elimination ordering.

How do the standard assumptions Â1 and Â2 translate to the problems (P) and (D)? It is not difficult to see that, with the analogous definitions $\mathcal{F}(P)\times\mathcal{F}(D)$ and $\mathcal{F}^0(P)\times\mathcal{F}^0(D)$, both of the following assumptions hold easily:

A1. the matrices $\{A_k\}_{k=1}^m$ are linearly independent;

A2. the set of interior feasible solutions $\mathcal{F}^0(P)\times\mathcal{F}^0(D)$ is nonempty.

2.2. The partial central path. A critical observation is that the central path equation $\hat{X}\hat{S}=\nu I$ implies that $\hat{X}^{-1}$ has the same density pattern as $\hat{S}$. That is, $\hat{X}^{-1}\in\mathcal{S}^F$, as was proven by Grone et al. in [12] (and used extensively by Fukuda et al. in [9]). The following observation represents an important connection between the spaces $\mathcal{S}^{F?}_{++}$ and $\mathcal{S}^n_{++}$.

THEOREM 2.1. *Let $X\in\mathcal{S}^{F?}_{++}$. Then there exists a unique $\hat{X}\in\mathcal{S}^n_{++}$ satisfying $\pi^F(\hat{X})=X$ and $\hat{X}^{-1}\in\mathcal{S}^F$. Moreover,*

$$\hat{X}=\operatorname{argmax}\big\{\det(\hat{Y}):\ \pi^F(\hat{Y})=X,\ \hat{Y}\in\mathcal{S}^n_{++}\big\}.$$

As in [9], we call $\hat{X}$ the *maximum-determinant positive definite completion* of X, and we also let $\hat{X}:\mathcal{S}^{F?}_{++}\to\mathcal{S}^n_{++}$ denote the function that yields $\hat{X}$ from X, that is, $\hat{X}\equiv\hat{X}(X)$. Using the function $\hat{X}$ and the direct correspondence $\hat{S}=S$ between the spaces of problems $(\hat{D})$ and (D), the central path equation $\hat{X}\hat{S}=\nu I$ can be described in terms of X and S as $\hat{X}(X)S=\nu I$.

For the algorithm presented in this paper, we wish to reformulate the central path equation $\hat{X}(X)S=\nu I$ once more, and so we now introduce some notation and a few small results. We define the following sets of lower triangular matrices, each of which have a density pattern equal to the lower triangular part of F:

$$\begin{aligned}\mathcal{L}^F&\equiv\{L\in\mathcal{L}^n:\ L_{ij}=0\ \forall\ i\geq j\text{ such that }(i,j)\notin F\},\\ \mathcal{L}^F_{++}&\equiv\mathcal{L}^F\cap\mathcal{L}^n_{++}.\end{aligned}$$

Noting the standard fact that the Cholesky factorization has no fill-in when the associated density pattern is chordal and $(1,\dots,n)$ is a perfect elimination ordering, we see that, for all $(X,S)\in\mathcal{S}^{F?}_{++}\times\mathcal{S}^F_{++}$, $\operatorname{chol}(\hat{X}(X)^{-1})\in\mathcal{L}^F_{++}$ and $\operatorname{chol}(S)\in\mathcal{L}^F_{++}$. We thus define

$$V:\mathcal{S}^{F?}_{++}\to\mathcal{L}^F_{++},\qquad V(X)\equiv\operatorname{chol}(\hat{X}(X)^{-1}), \tag{2.2}$$

$$L:\mathcal{S}^F_{++}\to\mathcal{L}^F_{++},\qquad L(S)\equiv\operatorname{chol}(S). \tag{2.3}$$

Using these definitions, it is now possible to further reformulate the central path equation $\hat{X}(X)S=\nu I$.

PROPOSITION 2.2. *Let $(X,S)\in\mathcal{S}^{F?}_{++}\times\mathcal{S}^F_{++}$ and $\nu>0$. Then $\hat{X}(X)S=\nu I$ if and only if $V(X)^{-1}L(S)=\sqrt{\nu}\,I$.*

Proof. Let $\hat{X}\equiv\hat{X}(X)$, $V\equiv V(X)$, and $L\equiv L(S)$. From (2.2) and (2.3), we see that $\hat{X}S=\nu I$ is equivalent to $V^{-1}LL^TV^{-T}=\nu I$, which itself shows that $V^{-1}L$ is the lower Cholesky factor of νI, that is, $V^{-1}L=\sqrt{\nu}\,I$. ∎

Proposition 2.2 now allows us to characterize the point (X_ν,S_ν,y_ν) on the central path $\mathcal{C}$ corresponding to $\nu>0$ as the unique solution of the system

$$\mathcal{A}(X)=b, \tag{2.4a}$$

$$\mathcal{A}^*(y) + S = C, \tag{2.4b}$$

$$V(X)^{-1} L(S) = \sqrt{\nu}\, I. \tag{2.4c}$$

Having expressed the central path in terms of the variables (X, S), we now wish to express the duality gap in terms of X and S as well. Given $(\hat{X}, \hat{S})$ and defining $(X, S) = (\pi^F(\hat{X}), \hat{S})$, we have

$$\hat{X} \bullet \hat{S} = \hat{X} \bullet S = \pi^F(\hat{X}) \bullet S = X \bullet S.$$

Alternatively, given (X, S) and letting $\hat{X}$ be any completion of X, we see that the equality also holds. Hence, $X \bullet S$ is the appropriate measure of the duality gap in the space $\mathcal{F}(P) \times \mathcal{F}(D)$. Furthermore, from (2.1) and the above equality, we have $\mu(\hat{X}, \hat{S}) = \mu(X, S)$.

The equation of the previous paragraph introduces a simple but important idea that will be used several times throughout the paper, and so we now give it a verbal description in order to make it simpler to refer to. Given $A, B \in \mathcal{S}^F$ and $\hat{A} \in \mathcal{S}^n$ such that $\pi^F(\hat{A}) = A$, we see that $\hat{A} \bullet B = \pi^F(\hat{A}) \bullet B = A \bullet B$, and we say that $(\hat{A}, A, B)$ are *trace-compatible*.

Given our usage of (2.4c) in this paper, we also wish to define the square root of the scaled duality gap measure $\mu(\cdot, \cdot)$:

$$\rho(X, S) \equiv \mu(X, S)^{1/2} \qquad \forall \;\; (X, S) \in \mathcal{S}_+^{F?} \times \mathcal{S}_+^F. \tag{2.5}$$

Note that, using the fact that $(\hat{X}(X), X, S)$ are trace-compatible, (2.2), (2.3), and (2.1), along with standard properties of the trace function and the Frobenius norm, we easily have that

$$\rho(X, S) = \frac{\|V(X)^{-1} L(S)\|_F}{\sqrt{n}}. \tag{2.6}$$

Equation (2.6) will come in handy throughout the presentation of this paper.

2.3. Nonsingularity of the partial central path. In section 4, we will develop a primal-dual path-following algorithm based on the central path equations (2.4), and so in this subsection we consider the nonsingularity of the Jacobian of the equations defining the central path, which will be necessary for the existence of the SDP direction proposed in section 4.

Noting that $V(X)^{-1}$ is generically dense, it is not difficult to see that the left-hand sides of the equations (2.4) are not "square" since they map a point $\mathcal{S}_{++}^{F?} \times \mathcal{S}_{++}^F \times \Re^m$ to $\Re^m \times \mathcal{S}^F \times \mathcal{L}^n$. As has become standard in the SDP literature, however, we can reconfigure the central path equations to obtain a square system. In this case, we replace (2.4c) with $L(S) - \sqrt{\nu}\, V(X) = 0$, which yields a system of equations $H(X, S, y) = (0, 0, 0)$, where $H : \mathcal{S}_{++}^{F?} \times \mathcal{S}_{++}^F \times \Re^m \to \Re^m \times \mathcal{S}^F \times \mathcal{L}^F$ is given by

$$H(X, S, y) \equiv \begin{pmatrix} \mathcal{A}(X) \\ \mathcal{A}^*(y) + S \\ L(S) - \sqrt{\nu}\, V(X) \end{pmatrix}. \tag{2.7}$$

Note that the definition of H is dependent on $\nu > 0$, which we consider fixed in the subsequent discussion. We remark that (2.7) is reminiscent of the central path

equation $\hat{S} - \nu\hat{X}^{-1} = 0$, where the usual complementarity equation $\hat{X}\hat{S} = \nu I$ has been reconfigured. This formulation is not appropriate for primal-dual path-following algorithms because, at any $(\hat{X}, \hat{S}, y) \in \mathcal{S}^n_{++} \times \mathcal{S}^n_{++} \times \Re^m$, the resulting Newton direction $(\Delta\hat{X}, \Delta\hat{S}, \Delta y)$ has the property that $\Delta\hat{X}$ depends only on $\hat{X}$ and not on $(\hat{S}, y)$; i.e., the Newton direction is not "primal-dual." In fact, an analogous system for partial matrices uses $S - \nu\hat{X}(X)^{-1} = 0$ instead of $\hat{X}(X)S = \nu I$, but this system also suffers a similar drawback. Hence, we have chosen to model the central path as in (2.7), in part because, in section 4, (2.7) will yield a primal-dual Newton direction due to the special structure of the functions L and V.

We now wish to investigate the Jacobian of H and to determine whether it is nonsingular (perhaps under suitable conditions). Since the derivative of H clearly depends on the derivatives of $V(\cdot)$ and $L(\cdot)$, we first describe these in the set of propositions and corollaries below (whose proofs are not difficult and are hence left to the reader).

PROPOSITION 2.3. *Let $M \in \mathcal{S}^n_{++}$. Then the first derivative of chol$(\cdot)$ at M is given by the invertible, linear map $chol'(M) : \mathcal{S}^n \to \mathcal{L}^n$, which is defined by the following: for all $N \in \mathcal{S}^n$, $K' \equiv chol'(M)[N] \in \mathcal{L}^n$ is the unique solution of the equation $N = K'K^T + K(K')^T$, where $K \equiv chol(M)$.*

COROLLARY 2.4. *Let $S \in \mathcal{S}^F_{++}$. Then the first derivative of $L(\cdot)$ at S is the invertible, linear map $L'(S) : \mathcal{S}^F \to \mathcal{L}^F$, which is defined by the following: for all $B \in \mathcal{S}^F$, $L' \equiv L'(S)[B] \in \mathcal{L}^F$ is the unique solution of the equation*

$$B = L'L^T + L(L')^T, \tag{2.8}$$

where $L \equiv L(S)$.

PROPOSITION 2.5. *Let $X \in \mathcal{S}^{F?}_{++}$. Then the linear map $\hat{X}'(X) : \mathcal{S}^F \to \mathcal{S}^n$ is defined by the following: for all $A \in \mathcal{S}^F$, $\hat{X}' \equiv \hat{X}'(X)[A] \in \mathcal{S}^n$ uniquely satisfies the requirements*

$$\pi^F(\hat{X}') = A, \qquad \hat{X}^{-1}\hat{X}'\hat{X}^{-1} \in \mathcal{S}^F, \tag{2.9}$$

where $\hat{X} \equiv \hat{X}(X)$.

COROLLARY 2.6. *Let $X \in \mathcal{S}^{F?}_{++}$. Then the linear map $V'(X) : \mathcal{S}^F \to \mathcal{L}^F$ is defined by the following: for all $A \in \mathcal{S}^F$, $V' \equiv V'(X)[A] \in \mathcal{L}^F$ is the unique solution of the equation*

$$-\hat{X}^{-1}\hat{X}'\hat{X}^{-1} = V'V^T + V(V')^T, \tag{2.10}$$

where $V \equiv V(X)$, $\hat{X} = \hat{X}(X)$, and $\hat{X}' \equiv \hat{X}'(X)[A]$. In addition, $V'(X)$ is invertible.

Having described the derivatives of $V(\cdot)$ and $L(\cdot)$, we now turn to the derivative of H. From (2.7), we see that the linear map $H' : \mathcal{S}^F \times \mathcal{S}^F \times \Re^m \to \Re^m \times \mathcal{S}^F \times \mathcal{L}^F$ is defined by

$$H'(X, S, y)[A, B, c] = \begin{pmatrix} \mathcal{A}(A) \\ \mathcal{A}^*(c) + B \\ L'(S)[B] - \sqrt{\nu}\, V'(X)[A] \end{pmatrix}. \tag{2.11}$$

In Lemma 2.8, Corollary 2.9, and Theorem 2.10, we show that $H'(X, S, y)$ is invertible as long as the product $V(X)^{-1}L(S)$ is sufficiently close to some positive multiple of the identity matrix, but first we need a technical lemma whose proof is straightforward that will prove useful below and also throughout the paper.

LEMMA 2.7. *Let $J \in \mathcal{L}^n$. Then $\|J\|_F \leq \|J + J^T\|_F/\sqrt{2}$, with equality holding if and only if J is strictly lower triangular.*

LEMMA 2.8. *Let $(X, S) \in \mathcal{S}^{F?}_{++} \times \mathcal{S}^F_{++}$. Define $V \equiv V(X)$ and $L \equiv L(S)$, and let $V'(X)$ and $L'(S)$ be as in Corollaries* 2.6 *and* 2.4*, respectively. Then, for all $A \in \mathcal{S}^F$ and for all $\beta > 0$,*

$$-A \bullet \left(V'(X)^{-1} \circ L'(S)\right)[A] \geq \left\|L^{-1}AL^{-T}\right\|_F^2 \left((1+\sqrt{2})\beta^3 - \sqrt{2}(\beta + \|Q\|_F)^3\right), \tag{2.12}$$

where $Q \equiv V^{-1}L - \beta I$.

Proof. With $L' \equiv L'(S)[A]$, we see from (2.8) that

$$A = L'L^T + L(L')^T. \tag{2.13}$$

Also defining $\hat{X} \equiv \hat{X}(X)$ and $\hat{X}' \equiv \hat{X}'(X)[V'(X)^{-1}[L']]$, we see from (2.9) and (2.10) that

$$\pi^F(\hat{X}') = V'(X)^{-1}[L'], \qquad -\hat{X}^{-1}\hat{X}'\hat{X}^{-1} = L'V^T + V(L')^T. \tag{2.14}$$

Now using (2.14), (2.2), and the trace-compatibility of $(\hat{X}', V'(X)^{-1}[L'], A)$, we see that the left-hand side of (2.12) equals

$$\begin{aligned} -A \bullet V'(X)^{-1}[L'] &= -A \bullet \hat{X}' = A \bullet \hat{X}\left(L'V^T + V(L')^T\right)\hat{X} \\ &= 2\, A \bullet V^{-T}V^{-1}L'V^{-1}. \end{aligned} \tag{2.15}$$

Introducing the notation $\tilde{A} \equiv L^{-1}AL^{-T}$ and $\tilde{L} \equiv L^{-1}L'$ so that $\tilde{A} = \tilde{L} + \tilde{L}^T$ by (2.13) and using the definition of Q, we observe that

$$\begin{aligned} A \bullet V^{-T}V^{-1}L'V^{-1} &= \tilde{A} \bullet (V^{-1}L)^T(V^{-1}L)\tilde{L}(V^{-1}L) \\ &= \tilde{A} \bullet (Q+\beta I)^T (Q+\beta I)\,\tilde{L}\,(Q+\beta I). \end{aligned} \tag{2.16}$$

Expanding the right-hand argument of the inner-product just obtained, we see that

$$\begin{aligned} (Q+\beta I)^T (Q+\beta I)\,\tilde{L}\,(Q+\beta I) = {} & Q^TQ\tilde{L}Q + \beta Q^TQ\tilde{L} + \beta Q^T\tilde{L}Q + \beta^2 Q^T\tilde{L} \\ & + \beta Q\tilde{L}Q + \beta^2 Q\tilde{L} + \beta^2\tilde{L}Q + \beta^3\tilde{L}. \end{aligned} \tag{2.17}$$

Now combining (2.15), (2.16), and (2.17), applying Lemma 2.7 with $J = \tilde{L}$, and using standard properties of the trace function and the Frobenius norm, we have

$$\begin{aligned} -A \bullet V'(X)^{-1}[L'] &= 2\,\tilde{A} \bullet Q^TQ\tilde{L}Q + 2\,\beta\,\tilde{A} \bullet Q^TQ\tilde{L} + 2\,\beta\,\tilde{A} \bullet Q^T\tilde{L}Q + 2\,\beta^2\,\tilde{A} \bullet Q^T\tilde{L} \\ &\quad + 2\,\beta\;\tilde{A} \bullet Q\tilde{L}Q + 2\,\beta^2\,\tilde{A} \bullet Q\tilde{L} + 2\;\tilde{A} \bullet \beta^2\tilde{L}Q + 2\,\beta^3\,\tilde{A} \bullet \tilde{L} \\ &\geq \beta^3\|\tilde{A}\|_F^2 - \sqrt{2}\|\tilde{A}\|_F^2\|Q\|_F^3 - 3\sqrt{2}\beta\|\tilde{A}\|_F^2\|Q\|_F^2 - 3\sqrt{2}\beta^2\|\tilde{A}\|_F^2\|Q\|_F \\ &= \|\tilde{A}\|_F^2\left((1+\sqrt{2})\beta^3 - \sqrt{2}\left(\|Q\|_F^3 + 3\,\beta\,\|Q\|_F^2 + 3\,\beta^2\,\|Q\|_F + \beta^3\right)\right) \\ &= \|\tilde{A}\|_F^2\left((1+\sqrt{2})\beta^3 - \sqrt{2}\,(\|Q\|_F + \beta)^3\right), \end{aligned}$$

which proves the lemma. (Note that, within the inequality, we have also used the equality $2\,\beta^3\,\tilde{A} \bullet \tilde{L} = \beta^3\,\|\tilde{A}\|_F^2$.) ∎

COROLLARY 2.9. *Let* $(X,S) \in \mathcal{S}^{F?}_{++} \times \mathcal{S}^{F}_{++}$. *Define* $V \equiv V(X)$ *and* $L \equiv L(S)$, *and let* $V'(X)$ *and* $L'(S)$ *be as in Corollaries* 2.6 *and* 2.4, *respectively. Then, if* $\|V^{-1}L - \beta\, I\|_F \leq \beta/6$ *for some* $\beta > 0$, *the operator* $-V'(X)^{-1} \circ L'(S)$ *is positive definite.*

Proof. By (2.12), $-V'(X)^{-1} \circ L'(S)$ is positive definite as long there exists some β such that

$$(1+\sqrt{2})\beta^3 - \sqrt{2}\,(\|Q\|_F + \beta)^3 > 0,$$

where $Q \equiv V^{-1}L - \beta I$, which is equivalent to

$$\|Q\|_F < \left[2^{-1/6}(1+\sqrt{2})^{1/3} - 1\right]\beta.$$

Since the coefficient in front of β is approximately 0.1951, the result follows. ∎

Corollary 2.9 now allows us to prove that H' is nonsingular under certain conditions.

THEOREM 2.10. *Let* $(X,S,y) \in \mathcal{S}^{F?}_{++} \times \mathcal{S}^{F}_{++} \times \Re^m$, *and suppose there exists some* $\beta > 0$ *such that* $\|V^{-1}L - \beta\, I\|_F \leq \beta/6$, *where* $V \equiv V(X)$ *and* $L \equiv L(S)$. *Then the linear map* $H' : \mathcal{S}^F \times \mathcal{S}^F \times \Re^m \to \Re^m \times \mathcal{S}^F \times \mathcal{L}^F$ *defined by* (2.11) *is invertible.*

Proof. To show that H' is invertible, we show that $(0,0,0)$ is the only solution of the equation $H'(X,S,y)[A,B,c] = (0,0,0)$, where $(A,B,c) \in \mathcal{S}^F \times \mathcal{S}^F \times \Re^m$. As is standard in the SDP literature, it is not difficult to see that this equation can be reduced to the $m \times m$ system

$$\left(\mathcal{A} \circ V'(X)^{-1} \circ L'(S) \circ \mathcal{A}^*\right)(c) = 0.$$

Since $V'(X)^{-1} \circ L'(S)$ is negative definite by Corollary 2.9 and since $\mathcal{A}^*$ is injective by assumption A1, we conclude that $c = 0$, which immediately implies that $(A,B) = (0,0)$, as desired. ∎

The above theorem will help us establish the existence of the Newton direction that will be used as the basis for our algorithm to solve problems (P) and (D) in section 4. Moreover, the theorem motivates the need for the neighborhood condition $\|V^{-1}L - \beta I\|_F \leq \beta/6$ that we will formally introduce next in section 3.

3. Technical results. In this section, we prove several results that will be used for establishing the polynomial convergence of the algorithm that we propose in section 4.

3.1. Properties of the partial central path map. Given $\gamma \in [0, 1/6]$, we define a feasible neighborhood of the central path as follows:

(3.1)
$$\mathcal{N}(\gamma) \equiv \left\{(X,S,y) \in \mathcal{F}^0(P) \times \mathcal{F}^0(D) :\ \|V(X)^{-1}L(S) - \rho(X,S)\, I\|_F \leq \gamma\, \rho(X,S)\right\}.$$

Clearly $\mathcal{N}(\gamma)$ is nonempty as $(X_\nu, S_\nu, y_\nu) \in \mathcal{N}(\gamma)$ for all $\nu > 0$. Note also, by Theorem 2.10 with $\beta = \rho(X,S)$ as well as the fact that $\gamma \leq 1/6$, that $H'(X,S,y)$ is invertible for all $(X,S,y) \in \mathcal{N}(\gamma)$.

We now wish to establish several fundamental results concerning both the central path function $V(X)^{-1}L(S)$ and the neighborhood $\mathcal{N}(\gamma)$. The first result establishes how the neighborhood condition can be restated simply as an inequality on $\mathrm{tr}(V(X)^{-1}L(S))$.

PROPOSITION 3.1. *$(X, S, y) \in \mathcal{N}(\gamma)$ if and only if (X, S, y) is primal-dual feasible and*

$$\operatorname{tr}\left(V(X)^{-1}L(S)\right) \geq \Gamma\,\rho(X,S), \tag{3.2}$$

where $\Gamma \equiv n - \gamma^2/2$. Moreover, for any $(X, S) \in \mathcal{S}^{F?}_{++} \times \mathcal{S}^{F}_{++}$, it holds that $\operatorname{tr}(V(X)^{-1}L(S)) \leq n\,\rho(X, S)$.

Proof. Letting $V \equiv V(X)$, $L \equiv L(S)$, and $\rho \equiv \rho(X, S)$ and using (2.6), we have

$$\begin{aligned}\|V^{-1}L - \rho I\|_F^2 &= (V^{-1}L - \rho I) \bullet (V^{-1}L - \rho I) = V^{-1}L \bullet V^{-1}L - 2\rho V^{-1}L \bullet I + \rho^2 I \bullet I \\ &= \|V^{-1}L\|_F^2 - 2\,\rho\operatorname{tr}(V^{-1}L) + n\,\rho^2 = 2\,n\,\rho^2 - 2\,\rho\operatorname{tr}(V^{-1}L),\end{aligned}$$

from which the first statement of the proposition follows, using (3.1). The second statement of the proposition also follows from the above equations, which imply $\|V^{-1}L - \rho\, I\|_F^2/(2\,\rho) = n\,\rho - \operatorname{tr}(V^{-1}L)$. ∎

The next proposition establishes some results concerning the second derivative of the function $\hat{V}(\hat{X})^{-1}\hat{L}(\hat{S})$, which, as described in (3.3), is analogous to $V(X)^{-1}L(S)$ but is defined on all of $\mathcal{S}^n_{++} \times \mathcal{S}^n_{++}$.

PROPOSITION 3.2. *Let $\hat{V} : \mathcal{S}^n_{++} \to \mathcal{L}^n_{++}$ and $\hat{L} : \mathcal{S}^n_{++} \to \mathcal{L}^n_{++}$ be defined by*

$$\hat{V}(\hat{X}) = \mathit{chol}(\hat{X}^{-1}), \qquad \hat{L}(\hat{S}) = \mathit{chol}(\hat{S}) \qquad \forall\ \hat{X}, \hat{S} \in \mathcal{S}^n_{++}. \tag{3.3}$$

Then the function $\Phi : \mathcal{S}^n_{++} \times \mathcal{S}^n_{++} \to \mathcal{L}^n_{++}$ defined by $\Phi(\hat{X}, \hat{S}) = \hat{V}(\hat{X})^{-1}\hat{L}(\hat{S})$ satisfies the following:

(i) *for any fixed $\hat{S} \in \mathcal{S}^n_{++}$, $\operatorname{tr}(\Phi(\cdot, \hat{S}))$ is a strictly concave function;*

(ii) *for all $\hat{A}, \hat{B} \in \mathcal{S}^n$,*

$$\left\|\Phi''(\hat{X}, \hat{S})[\hat{A}, \hat{B}]^{(2)}\right\|_F \leq \frac{1}{\sqrt{2}}\,\|\hat{V}^{-1}\hat{L}\|\left(\|\hat{V}^T\hat{A}\hat{V}\|_F + \|\hat{L}^{-1}\hat{B}\hat{L}^{-T}\|_F\right)^2, \tag{3.4}$$

where $\hat{V} \equiv \hat{V}(\hat{X})$ and $\hat{L} \equiv \hat{L}(\hat{S})$.

Proof. In certain places throughout this proof, we will avoid the use of the hat ($\hat{\cdot}$) notation, which indicates fully dense matrices, in order to simplify the notation; the meanings of the expressions will be clear from the context. Also to simplify notation, we define $\hat{U} : \mathcal{S}^n_{++} \to \mathcal{U}^n_{++}$ as being defined by $\hat{U}(\hat{X}) = \hat{V}(\hat{X})^{-T}$. With this definition, we see that $\Phi(\hat{X}, \hat{S}) = \hat{U}(\hat{X})^T\hat{L}(\hat{S})$ and that $\hat{X} = \hat{U}(\hat{X})\hat{U}(\hat{X})^T$.

To prove both (i) and (ii), we consider the second derivative of Φ. Using (3.3) along with arguments similar to those found in the derivation of $V'(X)$ and $L'(S)$ in section 2.3, we see that, for all $A, B \in \mathcal{S}^n$,

$$\Phi' \equiv \Phi'(\hat{X}, \hat{S})[A, B] = (U')^T L + U^T L', \tag{3.5}$$

where $U \equiv \hat{U}(\hat{X})$ and $L \equiv \hat{L}(\hat{S})$ and $U' \equiv \hat{U}'(\hat{X})[A] \in \mathcal{U}^n$ and $L' \equiv \hat{L}'(\hat{S})[B] \in \mathcal{L}^n$ are, respectively, the unique solutions of the equations

$$A = U'U^T + U(U')^T, \qquad B = L'L^T + L(L')^T. \tag{3.6}$$

Differentiating once again, we see that

$$\Phi'' \equiv \Phi''(\hat{X}, \hat{S})[A, B]^{(2)} = (U'')^T L + 2\,(U')^T L' + U^T L'', \tag{3.7}$$

where $U'' \equiv \hat{U}''(\hat{X})[A]^{(2)} \in \mathcal{U}^n$ and $L'' \equiv \hat{L}''(\hat{S})[B]^{(2)} \in \mathcal{L}^n$ are, respectively, the unique solutions of the equations

$$0 = U''U^T + 2\,U'(U')^T + U(U'')^T, \qquad 0 = L''L^T + 2\,L'(L')^T + L(L'')^T. \tag{3.8}$$

We now prove (i). Letting h denote the function $\mathrm{tr}(\Phi(\cdot, \hat{S}))$, where $\hat{S}$ is fixed, it is straightforward to verify that $h''(\hat{X})[A]^{(2)} = \mathrm{tr}((U'')^T L)$, where U'' and L are defined as above. From (3.8), we have

$$U^{-1}U'' + (U'')^T U^{-T} = -2(U^{-1}U')(U^{-1}U')^T, \tag{3.9}$$

which implies that $\mathrm{diag}(U^{-1}U'') \leq 0$, since the right-hand side of (3.9) is negative semidefinite, which in turn implies that $\mathrm{diag}(U'') \leq 0$, since $U^{-1} \in \mathcal{U}^n_{++}$. It follows that $h''(\hat{X})[A]^{(2)} < 0$ unless $\mathrm{diag}(U'') = 0$. So suppose $\mathrm{diag}(U'') = 0$. Then, by (3.9), we see

$$\mathrm{diag}\left((U^{-1}U')(U^{-1}U')^T\right) = 0 \quad \Longleftrightarrow \quad U^{-1}U' = 0 \quad \Longleftrightarrow \quad U' = 0 \quad \Longleftrightarrow \quad A = 0.$$

Thus, we conclude that for all $A \neq 0$, $h''(\hat{X})[A]^{(2)} < 0$. This proves that h is strictly concave.

We now prove (ii). Using (3.5)–(3.8), Lemma 2.7, and standard properties of the 2-norm and the Frobenius norm, we have

$$\begin{aligned}
\|\Phi''\|_F &\leq \|(U'')^T L\|_F + 2\|(U')^T L'\|_F + \|U^T L''\|_F \\
&\leq \|U^T L\| \left(\|U^{-1}U''\|_F + 2\|U^{-1}U'\|_F \|L^{-1}L'\|_F + \|L^{-1}L''\|_F\right) \\
&\leq \|U^T L\|(\sqrt{2}\,\|(U^{-1}U')(U^{-1}U')^T\|_F + 2\|U^{-1}U'\|_F\|L^{-1}L'\|_F \\
&\qquad + \sqrt{2}\|(L^{-1}L')(L^{-1}L')^T\|_F) \\
&\leq \|U^T L\|(\sqrt{2}\,\|U^{-1}U'\|_F^2 + 2\|U^{-1}U'\|_F\|L^{-1}L'\|_F + \sqrt{2}\|L^{-1}L'\|_F^2) \\
&\leq \|U^T L\| \left(\frac{1}{\sqrt{2}}\,\|U^{-1}AU^{-T}\|_F^2 + \|U^{-1}AU^{-T}\|_F\|L^{-1}BL^{-T}\|_F \right. \\
&\qquad \left. + \frac{1}{\sqrt{2}}\|L^{-1}BL^{-T}\|_F^2\right) \\
&\leq \frac{1}{\sqrt{2}}\|U^T L\| \left(\|U^{-1}AU^{-T}\|_F + \|L^{-1}BL^{-T}\|_F\right)^2.
\end{aligned}$$

The result now follows from the definition of $\hat{U}(\cdot)$. ∎

The next result plays a crucial role in the analysis of section 4. In words, the theorem says that, given a fixed pair $(X, S) \in \mathcal{S}^{F?}_{++} \times \mathcal{S}^F_{++}$, the maximum-determinant completion $\hat{X}(X)$ also maximizes the function $\mathrm{tr}(\hat{V}(\cdot)^{-1}L(S))$ among all positive definite completions of X.

THEOREM 3.3. *Let $(X, S) \in \mathcal{S}^{F?}_{++} \times \mathcal{S}^F_{++}$, and let $\hat{V} : \mathcal{S}^n_{++} \to \mathcal{L}^n_{++}$ be defined as in Proposition* 3.2. *Then*

$$\hat{X}(X) = \mathrm{argmax}\left\{\mathrm{tr}\big(\hat{V}(\hat{Y})^{-1}L(S)\big) : \ \pi^F(\hat{Y}) = X, \ \hat{Y} \in \mathcal{S}^n_{++}\right\}.$$

Proof. Noting that $\hat{L}(S) = L(S)$ and letting $L \equiv L(S)$, we have from Proposition 3.2 that $h(\hat{Y}) \equiv \mathrm{tr}(\hat{V}(\hat{Y})^{-1}L)$ is a strictly concave function of $\hat{Y}$. Hence, since the constraints of the optimization problem under consideration are convex, any stationary point of this problem is a unique global maximum, and so we prove the theorem by showing that $\hat{X} \equiv \hat{X}(X)$ is a stationary point.

The derivative $h'(\hat{Y}) : \mathcal{S}^n \to \Re$ of h at $\hat{Y}$ is given by

$$h' \equiv h'(\hat{Y})[\hat{A}] = -\mathrm{tr}\big(\hat{V}^{-1}\hat{V}'\hat{V}^{-1}L\big) \tag{3.10}$$

for all $\hat{A} \in \mathcal{S}^n$, where $\hat{V} \equiv \hat{V}(\hat{Y})$ and $\hat{V}' \equiv \hat{V}'(\hat{Y})[\hat{A}]$ is the unique solution of the system

$$-\hat{Y}^{-1}\hat{A}\hat{Y}^{-1} = \hat{V}'\hat{V}^T + \hat{V}(\hat{V}')^T. \tag{3.11}$$

Premultiplying (3.11) by $\hat{V}^{-1}$, postmultiplying by $\hat{V}^{-T}$, and using the fact that $\hat{Y}^{-1} = \hat{V}\hat{V}^T$, we see that $-\hat{V}^T\hat{A}\hat{V} = \hat{V}^{-1}\hat{V}' + (\hat{V}')^T\hat{V}^{-T}$, which shows that $-\operatorname{diag}(\hat{V}^T\hat{A}\hat{V}) = 2\operatorname{diag}(\hat{V}^{-1}\hat{V}')$. Applying this equality to (3.10) and letting $W \in \mathcal{S}^n_{++}$ be the diagonal matrix defined by $W_{jj} = \hat{V}_{jj}^{-1}L_{jj}/2$, we deduce that

$$h' = \frac{1}{2}\operatorname{diag}(\hat{V}^T\hat{A}\hat{V})^T\operatorname{diag}(\hat{V}^{-1}L) = W \bullet \hat{V}^T\hat{A}\hat{V} = \hat{V}W\hat{V}^T \bullet \hat{A},$$

which implies that $\nabla h(\hat{Y}) = \hat{V}W\hat{V}^T$.

Let $\bar{F} \equiv V \times V \setminus F$. Considering that the variables $\hat{Y}_{ij}$ for $(i,j) \in F$ can be eliminated by the equation $\pi^F(\hat{Y}) = X$, we see that a stationary point $\hat{Y}$ of the optimization problem satisfies $\pi^{\bar{F}}(\nabla h(\hat{Y})) = 0$, that is, $\nabla h(\hat{Y}) = \hat{V}W\hat{V}^T \in \mathcal{S}^F$. Since $W \in \mathcal{S}^n_{++}$ is diagonal, $\hat{V}W\hat{V}^T \in \mathcal{S}^F$ is equivalent to $\hat{V}\hat{V}^T = \hat{Y}^{-1} \in \mathcal{S}^F$, which is precisely the condition that $\hat{X}$ satisfies uniquely by Theorem 2.1. So $\hat{X}$ is a stationary point of the optimization problem, which completes the proof. ∎

3.2. Miscellaneous results. In this subsection, we catalog a few results that will prove useful in section 4. The first two results give details about the system $H'(X,S,y)[A,B,c] = (0,0,R)$, where H' is given as in (2.11).

LEMMA 3.4. *Let $(X,S,y) \in \mathcal{S}^{F?}_{++} \times \mathcal{S}^F_{++} \times \Re^m$ and $R \in \mathcal{L}^F$ be given, and suppose that $(A,B,c) \in \mathcal{S}^F \times \mathcal{S}^F \times \Re^m$ satisfies $H'(X,S,y)[A,B,c] = (0,0,R)$. Then*

$$A \bullet B = (V^{-1}V' + (V')^TV^{-T}) \bullet V^{-1}BV^{-T} = 0, \tag{3.12}$$

where $V \equiv V(X)$ and $V' \equiv V'(X)[A]$.

Proof. From (2.11), we see that $\mathcal{A}(A) = 0$ and $\mathcal{A}^*(c) + B = 0$. Hence, $A \bullet B = -A \bullet \mathcal{A}^*(c) = -c^T\mathcal{A}(A) = 0$. Also, letting $\hat{X} \equiv \hat{X}(X)$, $\hat{X}' \equiv \hat{X}'(X)[A]$ and using (2.9), (2.10), (2.2), and the trace-compatibility of $(\hat{X}', A, B)$, we see that

$$A \bullet B = \hat{X}' \bullet B = -\hat{X}(V'V^T + V(V')^T)\hat{X} \bullet B = (V^{-1}V' + (V')^TV^{-T}) \bullet V^{-1}BV^{-T},$$

which completes the proof. ∎

PROPOSITION 3.5. *Let the conditions of Lemma 3.4 hold, and define $Q \equiv V^{-1}L - \sqrt{\nu}\,I$, where $L \equiv L(S)$ and $\nu > 0$ is as in (2.7). Suppose that $\|Q\|_F < \sqrt{\nu}/\sqrt{2}$. Then*

$$\|V^{-1}V'\|_F \le \frac{1}{\sqrt{2\nu}}\big(\sqrt{\nu} - \sqrt{2}\,\|Q\|_F\big)^{-1}\|V^{-1}(RL^T + LR^T)V^{-T}\|_F, \tag{3.13}$$

$$\|V^{-1}BV^{-T}\|_F \le \sqrt{\nu}(\sqrt{\nu} - \sqrt{2}\,\|Q\|_F)^{-1}\|V^{-1}(RL^T + LR^T)V^{-T}\|_F. \tag{3.14}$$

Proof. From (2.11), we have $B = L'(S)^{-1}[R + \sqrt{\nu}\,V'] = (R + \sqrt{\nu}\,V')L^T + L(R + \sqrt{\nu}\,V')^T$. Hence, letting $\tilde{R} \equiv RL^T + LR^T$,

$$V^{-1}BV^{-T} = V^{-1}\tilde{R}V^{-T} + \sqrt{\nu}\,V^{-1}V'(V^{-1}L)^T + \sqrt{\nu}\,V^{-1}L(V^{-1}V')^T.$$

It follows that

$$\begin{aligned}\frac{1}{\sqrt{\nu}}V^{-1}BV^{-T} - \sqrt{\nu}\left(V^{-1}V' + (V')^TV^{-T}\right) \\ = \frac{1}{\sqrt{\nu}}V^{-1}\tilde{R}V^{-T} + V^{-1}V'Q^T + Q(V^{-1}V')^T,\end{aligned}$$

which, from (3.12), implies

$$\begin{aligned}&\max\left\{\frac{1}{\sqrt{\nu}}\left\|V^{-1}BV^{-T}\right\|_F, \sqrt{\nu}\left\|V^{-1}V' + (V')^TV^{-T}\right\|_F\right\} \\ &\le \left\|\frac{1}{\sqrt{\nu}}V^{-1}BV^{-T} - \sqrt{\nu}\left(V^{-1}V' + (V')^TV^{-T}\right)\right\|_F \\ &\le \frac{1}{\sqrt{\nu}}\|V^{-1}\tilde{R}V^{-T}\|_F + 2\|Q\|_F\|V^{-1}V'\|_F.\end{aligned} \tag{3.15}$$

Applying Lemma 2.7 with $J = V^{-1}V'$ to (3.15), we see that

$$\sqrt{2}\big(\sqrt{\nu} - \sqrt{2}\|Q\|_F\big)\|V^{-1}V'\|_F \le \frac{1}{\sqrt{\nu}}\|V^{-1}\tilde{R}V^{-T}\|_F,$$

which proves (3.13). To prove (3.14), we combine (3.13) and (3.15) to obtain

$$\begin{aligned}\frac{1}{\sqrt{\nu}}\|V^{-1}BV^{-T}\|_F &\le \frac{1}{\sqrt{\nu}}\|V^{-1}\tilde{R}V^{-T}\|_F\left(1 + \frac{\sqrt{2}\|Q\|_F}{\sqrt{\nu} - \sqrt{2}\|Q\|_F}\right) \\ &= \big(\sqrt{\nu} - \sqrt{2}\|Q\|_F\big)^{-1}\|V^{-1}\tilde{R}V^{-T}\|_F. \qquad \square\end{aligned}$$

We next establish several inequalities that relate to bounds on the maximum and minimum eigenvalues of $\hat{X}(X)S$ for $(X,S) \in \mathcal{S}^{F?}_{++} \times \mathcal{S}^F_{++}$.

PROPOSITION 3.6. *Let $(X,S) \in \mathcal{S}^{F?}_{++} \times \mathcal{S}^F_{++}$, and define $\hat{X} \equiv \hat{X}(X)$, $V \equiv V(X)$, $L \equiv L(S)$, and $\rho \equiv \rho(X,S)$. Suppose $\|V^{-1}L - \rho I\| \le \gamma\rho$ for some $\gamma \ge 0$ satisfying $\gamma^2 + 2\gamma < 1$. Then the following hold:*

(i) $\|V^{-1}L\| \le (\gamma+1)\rho$;
(ii) $\|L^{-1}V\| \le (1 - (\gamma^2+2\gamma))^{-1/2}\rho^{-1}$.

Proof. Let $Q \equiv V^{-1}L - \rho I$, and note that

$$\begin{aligned}\|Q\|^2 = \|Q^TQ\| &= \|L^TV^{-T}V^{-1}L - \rho V^{-1}L - \rho L^TV^{-T} + \rho^2 I\| \\ &= \|L^TV^{-T}V^{-1}L - \rho^2 I - \rho(Q + Q^T)\|.\end{aligned}$$

Hence, from (2.2), (2.3), standard properties of $\|\cdot\|$, and the assumptions of the proposition,

$$\begin{aligned}\|\hat{X}S - \rho^2 I\| &= \|L^TV^{-T}V^{-1}L - \rho^2 I\| \\ &= \|L^TV^{-T}V^{-1}L - \rho^2 I - \rho(Q+Q^T) + \rho(Q+Q^T)\| \\ &\le \|Q\|^2 + 2\rho\|Q\| \le (\gamma^2+2\gamma)\rho^2.\end{aligned} \tag{3.16}$$

Recall from the definition of ρ that $\hat{X} \bullet S = X \bullet S = n\rho^2$, that is, $n\rho^2$ equals the sum of the eigenvalues of $\hat{X}S$. It follows that $\lambda_{\max}[\hat{X}S] \ge \rho^2$ and that $\lambda_{\min}[\hat{X}S] \le \rho^2$. Hence,

$$\|\hat{X}S - \rho^2 I\| = \max\{\lambda_{\max}[\hat{X}S] - \rho^2, \rho^2 - \lambda_{\min}[\hat{X}S]\}.$$

Thus, using (3.16), (2.2), and (2.3), we have

$$\|V^{-1}L\|^2 = \|L^T V^{-T} V^{-1} L\| = \|\hat{X}S\| = \lambda_{\max}[\hat{X}S] \le \|\hat{X}S - \rho^2 I\| + \rho^2 \le (\gamma^2 + 2\gamma + 1)\rho^2,$$

which proves (i). Similarly,

$$\begin{aligned}\|L^{-1}V\|^2 &= \|V^T L^{-T} L^{-1} V\| = \|S^{-1}\hat{X}^{-1}\| = \lambda_{\min}[\hat{X}S]^{-1} \\ &\le \left(\rho^2 - \|\hat{X}S - \rho^2 I\|\right)^{-1} \le \left(\rho^2 - (\gamma^2 + 2\gamma)\rho^2\right)^{-1},\end{aligned}$$

which proves (ii). □

We remark that the condition $(X, S, y) \in \mathcal{N}(\gamma)$ implies that the hypotheses of Proposition 3.6 hold since $\|V^{-1}L - \rho I\| \le \|V^{-1}L - \rho I\|_F$.

Finally, we state the following proposition, which follows as a direct extension of Lemmas 3.4 and 3.5 of Monteiro and Tsuchiya [24].

PROPOSITION 3.7. *Let $(X, S) \in \mathcal{S}^{F?}_{++} \times \mathcal{S}^{F}_{++}$ and $\bar{X}, \bar{S} \in \mathcal{S}^n$. Define $\hat{X} \equiv \hat{X}(X)$, $V \equiv V(X)$, and $L \equiv L(S)$. Suppose that there exists some $\tau \in (0, 1)$ such that*

$$\max\left\{\|V^T(\bar{X} - \hat{X})V\|, \|L^{-1}(\bar{S} - S)L^{-T}\|\right\} \le \tau.$$

Then $\bar{X}, \bar{S} \in \mathcal{S}^n_{++}$,

$$\max\left\{\|V^{-1}\bar{V}\|, \|\bar{V}^{-1}V\|, \|L^{-1}\bar{L}\|, \|\bar{L}^{-1}L\|\right\} \le \frac{1}{\sqrt{1-\tau}}, \quad \text{and} \quad \|\bar{V}^{-1}\bar{L}\| \le \frac{\|V^{-1}L\|}{1-\tau},$$

where $\bar{V} = chol(\bar{X}^{-1})$ and $\bar{L} = chol(\bar{S})$.

4. The partial primal-dual algorithm. The algorithm described in this section is based on the same ideas that typical path-following algorithms are based on—namely, the use of a Newton direction to decrease the duality gap, and a bound on the step-size to ensure proximity to the central path. Using these ideas, we establish the polynomiality of the algorithm in Theorem 4.9.

Suppose that $(X, S, y) \in \mathcal{N}(\gamma)$, where $\gamma \in [0, 1/6]$. Then, for a fixed constant $0 \le \sigma < 1$, we define the Newton direction $(\Delta X, \Delta S, \Delta y)$ at (X, S, y) as the solution of the system

$$H'(X, S, y)[\Delta X, \Delta S, \Delta y] = (0, 0, \sigma\rho V - L), \tag{4.1}$$

where $\nu \equiv \rho^2$ is implicit in the definition of H, $\rho \equiv \rho(X, S)$, $V \equiv V(X)$, and $L \equiv L(S)$. Note that $(\Delta X, \Delta S, \Delta y)$ is well defined by Theorem 2.10. We also make the following definitions for all $\alpha \in \Re$ such that $(X_\alpha, S_\alpha) \in \mathcal{S}^{F?}_{+} \times \mathcal{S}^{F}_{+}$:

$$\begin{aligned}(X_\alpha, S_\alpha, y_\alpha) &\equiv (X, S, y) + \alpha(\Delta X, \Delta S, \Delta y), \\ \mu_\alpha &\equiv \mu(X_\alpha, S_\alpha), \\ \rho_\alpha &\equiv \rho(X_\alpha, S_\alpha).\end{aligned}$$

We have the following proposition.

PROPOSITION 4.1. *Let $(X, S, y) \in N(\gamma)$, and define $\mu \equiv \mu(X, S)$, $\rho \equiv \rho(X, S)$, and $\zeta = (\Delta X \bullet S + X \bullet \Delta S)(n\mu)^{-1}$. Then for all $\alpha \in \Re$ such that $(X_\alpha, S_\alpha, y_\alpha) \in \mathcal{S}^{F?}_{+} \times \mathcal{S}^{F}_{+}$,*

$$\mu_\alpha = \mu(1 + \alpha\zeta), \qquad \rho_\alpha \le \rho\Big(1 + \frac{\alpha\zeta}{2}\Big). \tag{4.2}$$

Proof. Note that $\Delta X \bullet \Delta S = 0$ from (3.12). Using (2.1), we thus have

$$\mu_\alpha = \frac{(X + \alpha\Delta X) \bullet (S + \alpha\Delta S)}{n} = \frac{X \bullet S + \alpha\zeta n\mu}{n} = \mu + \alpha\zeta\mu,$$

which proves the equality. Similarly, using (2.5), we see that

$$\rho_\alpha = \rho(1 + \alpha\,\zeta)^{1/2} \leq \rho\Big(1 + \frac{\alpha\,\zeta}{2}\Big),$$

where the inequality follows from the real-number relation $1 + 2x \leq 1 + 2x + x^2$. ∎

With regard to the above proposition, it is important to mention that we anticipate that ζ is negative due to the fact that $\sigma < 1$, which would imply that $\mu_\alpha < \mu$ and $\rho_\alpha < \rho$; that is, the duality gap decreases along the direction $(\Delta X, \Delta S, \Delta y)$. This, however, must be proven under certain assumptions, as will be shown below. For the discussion of the generic algorithm next, it would be useful for the reader to assume that $\zeta < 0$.

We now state the generic primal-dual path-following algorithm that we study in this section.

Algorithm SDP.

Let $\varepsilon > 0$, $\gamma \in [0, 1/6]$, $\sigma \in [0, 1)$, and $(X^0, S^0, y^0) \in \mathcal{N}(\gamma)$ be given. Set $k = 0$.

Repeat until $X^k \bullet S^k \leq \varepsilon$ **do**

1. Set $(X, S, y) \equiv (X^k, S^k, y^k)$ and $\rho \equiv \rho(X, S)$.
2. Compute the Newton direction $(\Delta X, \Delta S, \Delta y)$ at (X, S, y).
3. Choose $\alpha \geq 0$ such that $(X_\alpha, S_\alpha, y_\alpha) \in \mathcal{N}(\gamma)$.
4. Set $(X^{k+1}, S^{k+1}, y^{k+1}) = (X_\alpha, S_\alpha, y_\alpha)$ and increment k by 1.

End

The remainder of this section is devoted to determining constants γ and σ and a constant step-size α such that Algorithm SDP terminates within a polynomial number of loops, where the polynomial depends on n, ε, and $X^0 \bullet S^0$.

To this end, we introduce constants $\gamma \geq 0$, $\delta \in [0, \sqrt{n})$, and $\tau \in (0, 1)$ satisfying

$$\gamma \leq \frac{1}{6}, \qquad 0 < 1 - (\gamma^2 + 2\gamma) \leq \frac{1}{\sqrt{2}} \tag{4.3}$$

and

$$2\,(\gamma + 1)\left(\delta^2 + \gamma^2\right)^{1/2} \leq \left(1 - \sqrt{2}\,\gamma\right)\left(1 - (\gamma^2 + 2\,\gamma)\right)\tau. \tag{4.4}$$

Note that, for example, the triple $(\gamma, \delta, \tau) = (0.138, 0.138, 0.79)$ satisfies (4.3) and (4.4) irrespective of the value of n. We also define

$$\sigma \equiv 1 - \frac{\delta}{\sqrt{n}}. \tag{4.5}$$

In addition, we make the mild assumption that n is large enough so that

$$\delta\sqrt{n} \geq \tau\gamma(\gamma + 1), \tag{4.6}$$

$$\sigma \geq \tau\theta, \tag{4.7}$$

$$\sqrt{n} \geq \frac{\gamma^2(\sigma - \tau\theta)(1 - \tau)}{2\sqrt{2}\,\tau^2(\gamma + 1)}, \tag{4.8}$$

where $\theta \equiv 1 + \gamma(\gamma + 1)/(2n)$. In fact, taking $(\gamma, \delta, \tau) = (0.138, 0.138, 0.79)$ as above shows that (4.6)–(4.8) are also satisfied for all values of n.

LEMMA 4.2. *Define $V \equiv V(X)$, $L \equiv L(S)$, and $\rho \equiv \rho(X,S)$, and suppose that $(X,S) \in \mathcal{S}^{F?}_{++} \times \mathcal{S}^{F}_{++}$ satisfies $\|V^{-1}L - \rho I\|_F \le \gamma\rho$. Then*

$$\|\sigma\rho I - V^{-1}L\|_F \le (\delta^2 + \gamma^2)^{1/2}\rho.$$

Proof. Using the definition of the Frobenius norm, we have

$$\begin{aligned}\|\sigma\rho I - V^{-1}L\|_F^2 &= \|\rho I - V^{-1}L + (\sigma-1)\rho I\|_F^2 \\ &= \|\rho I - V^{-1}L\|_F^2 + 2(\sigma-1)\rho(\rho I - V^{-1}L) \bullet I + (\sigma-1)^2 n\rho^2, \\ &= \|\rho I - V^{-1}L\|_F^2 + 2(\sigma-1)\rho\left(n\rho - \operatorname{tr}(V^{-1}L)\right) + (\sigma-1)^2 n\rho^2 \\ &\le \gamma^2\rho^2 + (\sigma-1)^2 n\rho^2,\end{aligned}$$

where the inequality follows by assumption and by the second statement of Proposition 3.1. Since $(\sigma-1)^2 n = \delta^2$ by (4.5), the result follows. □

PROPOSITION 4.3. *Let $(X,S,y) \in \mathcal{N}(\gamma)$. Then*

$$\max\left\{\|V^T\hat{X}'V\|_F, \|L^{-1}\Delta S L^{-T}\|_F\right\} \le \tau,$$

where $\hat{X}' \equiv \hat{X}'(X)[\Delta X]$. As a result, $\hat{X} + \hat{X}' \in \mathcal{S}^n_{++}$ and $S + \Delta S \in \mathcal{S}^F_{++}$, where $\hat{X} \equiv \hat{X}(X)$.

Proof. Let $V \equiv V(X)$, $L \equiv L(S)$, and $\rho \equiv \rho(X,S)$. Also, letting $V' \equiv V'(X)[\Delta X]$ and using (2.10) and (2.2), we see that

$$-V^T\hat{X}'V = V^{-1}V' + (V')^T V^{-T} \qquad \Longrightarrow \qquad \|V^T\hat{X}'V\|_F \le 2\,\|V^{-1}V'\|_F.$$

Note that the hypotheses of Proposition 3.5 hold, with $R \equiv \sigma\rho V - L$, $\nu \equiv \rho^2$, and $Q \equiv V^{-1}L - \rho I$. Hence, using (3.13), standard properties of norms, (3.1), Lemma 4.2, Proposition 3.6(i), (4.3), and (4.4), we have

$$\begin{aligned}2\,\|V^{-1}V'\|_F &\le \frac{\sqrt{2}}{\rho}\big(\rho - \sqrt{2}\,\|Q\|_F\big)^{-1}\|V^{-1}((\sigma\rho V - L)L^T + L(\sigma\rho V - L)^T)V^{-T}\|_F \\ &\le \frac{2\sqrt{2}}{\rho}\big(\rho - \sqrt{2}\,\gamma\,\rho\big)^{-1}\|V^{-1}(\sigma\rho V - L)L^T V^{-T}\|_F \\ &\le \frac{2\sqrt{2}}{\rho^2}\big(1 - \sqrt{2}\,\gamma\big)^{-1}\|\sigma\rho I - V^{-1}L\|_F\|V^{-1}L\| \\ &\le 2\sqrt{2}\big(1 - \sqrt{2}\,\gamma\big)^{-1}\big(\delta^2 + \gamma^2\big)^{1/2}(\gamma+1) \\ &\le 2\big(1 - (\gamma^2 + 2\,\gamma)\big)^{-1}\big(1 - \sqrt{2}\,\gamma\big)^{-1}\big(\delta^2 + \gamma^2\big)^{1/2}(\gamma+1) \le \tau. \qquad (4.9)\end{aligned}$$

Now using (3.14) and Proposition 3.6(ii) along with similar arguments, we have

$$\begin{aligned}\|L^{-1}\Delta S L^{-T}\|_F &\le \|L^{-1}V\|^2\,\|V^{-1}\Delta S V^{-T}\|_F \\ &\le \|L^{-1}V\|^2\,\rho\big(\rho - \sqrt{2}\,\|Q\|_F\big)^{-1}\|V^{-1}((\sigma\rho V - L)L^T + L(\sigma\rho V - L)^T)V^{-T}\|_F \\ &\le 2\left(1 - (\gamma^2 + 2\gamma)\right)^{-1}\left(1 - \sqrt{2}\,\gamma\right)^{-1}\left(\delta^2 + \gamma^2\right)^{1/2}(\gamma+1) \le \tau,\end{aligned}$$

which concludes the proof of the first statement of the proposition. The second statement follows from Proposition 3.7, with $\bar{X} = \hat{X} + \hat{X}'$ and $\bar{S} = S + \Delta S$ (and the fact that $\Delta S \in \mathcal{S}^F$). □

COROLLARY 4.4. *Let $(X,S,y) \in \mathcal{N}(\gamma)$. Then for all $0 \le \alpha \le 1$, $(X_\alpha, S_\alpha) \in \mathcal{S}^{F?}_{++} \times \mathcal{S}^F_{++}$.*

Proof. By Proposition 4.3, we see that $S_1 = S + \Delta S \in \mathcal{S}^F_{++}$. Since S_α is a convex combination of S and S_1 for any $0 \le \alpha \le 1$, it follows that $S_\alpha \in \mathcal{S}^F_{++}$. Likewise, $\hat{X} + \alpha\hat{X}' \in \mathcal{S}^n_{++}$ for all $0 \le \alpha \le 1$. Noting that, by Theorem 2.1 and (2.9),

$$\pi^F(\hat{X} + \alpha\hat{X}') = X + \alpha\Delta X = X_\alpha,$$

we see that $\hat{X} + \alpha\hat{X}'$ is a positive definite completion of X_α, which implies $X_\alpha \in \mathcal{S}^{F?}_{++}$. ∎

LEMMA 4.5. *Let $(X, S, y) \in \mathcal{N}(\gamma)$, and define $V \equiv V(X)$, $L \equiv L(S)$, $\rho \equiv \rho(X, S)$, $V' \equiv V'(X)[\Delta X]$, and $L' \equiv L'(S)[\Delta S]$. Also let Γ be defined as in Proposition* 3.1. *Then*

$$V^{-1}L \bullet \left(V^{-1}L' - V^{-1}V'V^{-1}L\right) \le \left((\sigma-1)n + \frac{1}{2}\tau(\gamma+1)\gamma\right)\rho^2, \tag{4.10}$$

$$\frac{\Gamma}{n}V^{-1}L \bullet \left(V^{-1}L' - V^{-1}V'V^{-1}L\right) - \rho\,\mathrm{tr}\left(V^{-1}V'(\rho I - V^{-1}L)\right) \le \left((\sigma-1)\Gamma + \frac{1}{2}\tau\gamma^2\left(1 + \frac{1}{2n}\gamma(\gamma+1)\right)\right)\rho^2. \tag{4.11}$$

Proof. We first prove some simple bounds that will allow us to prove (4.10) and (4.11) more easily. Defining $P \equiv V^{-1}V'\left(\rho I - V^{-1}L\right)$ and using standard properties of $\mathrm{tr}(\cdot)$, $\|\cdot\|$, and $\|\cdot\|_F$ along with (3.1), Proposition 3.6(i), and (4.9) (which appears inside the proof of Proposition 4.3), we see that

$$\begin{aligned}|V^{-1}L \bullet P| &= |(V^{-1}V')^T(V^{-1}L) \bullet (\rho I - V^{-1}L)| \le \|V^{-1}V'\|_F\|V^{-1}L\|\,\|\rho I - V^{-1}L\|_F \\ &\le \frac{1}{2}\tau(\gamma+1)\gamma\rho^2. \end{aligned} \tag{4.12}$$

Similarly,

$$|(V^{-1}L - \rho I) \bullet P| \le \|V^{-1}V'\|_F\|V^{-1}L - \rho I\|_F^2 \le \frac{1}{2}\tau\gamma^2\rho^2. \tag{4.13}$$

Now, the equation (4.1) for the Newton direction $(\Delta X, \Delta S, \Delta y)$ shows that $L' - \rho V' = \sigma\rho V - L$, which implies $V^{-1}L' = \sigma\rho I - V^{-1}L + \rho V^{-1}V'$. Substituting for $V^{-1}L'$ in the left-hand side of (4.10) and using (2.6), the second statement of Proposition 3.1, and (4.12), we see that

$$\begin{aligned} V^{-1}L \bullet \left(V^{-1}L' - V^{-1}V'V^{-1}L\right) &= V^{-1}L \bullet \left(\sigma\rho I - V^{-1}L + \rho V^{-1}V' - V^{-1}V'V^{-1}L\right) \\ &= \sigma\rho\,\mathrm{tr}(V^{-1}L) - \|V^{-1}L\|_F^2 + V^{-1}L \bullet P \\ &\le (\sigma-1)n\rho^2 + \frac{1}{2}\tau(\gamma+1)\gamma\rho^2, \end{aligned}$$

as desired. Using similar arguments along with (4.13) and the fact that $\Gamma/n = 1 - \gamma^2/(2n)$, (4.11) is proven as follows:

$$\begin{aligned} &\frac{\Gamma}{n}V^{-1}L \bullet \left(V^{-1}L' - V^{-1}V'V^{-1}L\right) - \rho\,\mathrm{tr}\left(V^{-1}V'(\rho I - V^{-1}L)\right) \\ &\quad= \frac{\Gamma}{n}\sigma\rho\,\mathrm{tr}(V^{-1}L) - \frac{\Gamma}{n}\|V^{-1}L\|_F^2 + \left(\frac{\Gamma}{n}V^{-1}L - \rho I\right) \bullet P \\ &\quad= \frac{\Gamma}{n}\sigma\rho\,\mathrm{tr}(V^{-1}L) - \frac{\Gamma}{n}\|V^{-1}L\|_F^2 + \left(V^{-1}L - \rho I\right) \bullet P - \frac{\gamma^2}{2n}V^{-1}L \bullet P \\ &\quad\le \Gamma\sigma\rho^2 - \Gamma\rho^2 + \frac{1}{2}\tau\gamma^2\rho^2 + \frac{1}{4n}\tau(\gamma+1)\gamma^3\rho^2. \end{aligned}$$

∎

PROPOSITION 4.6. *Let $(X, S, y) \in \mathcal{N}(\gamma)$, and define $\mu \equiv \mu(X, S)$. Then*

$$\zeta \equiv \frac{\Delta X \bullet S + X \bullet \Delta S}{n\mu} \leq -\frac{\delta}{\sqrt{n}}.$$

Hence, $\mu(\cdot,\cdot)$ and $\rho(\cdot,\cdot)$ decrease from (X, S, y) along the direction $(\Delta X, \Delta S, \Delta y)$.

Proof. Let $\hat{X} \equiv \hat{X}(X)$, $\hat{X}' \equiv \hat{X}'(X)[\Delta X]$, $V \equiv V(X)$, $V' \equiv V'(X)[\Delta X]$, $L \equiv L(S)$, $L' \equiv L'(S)[\Delta S]$, and $\rho \equiv \rho(X, S)$. Then using Theorem 2.1, (2.9) with $A = \Delta X$, the fact that $S \in \mathcal{S}^F$ and $\Delta S \in \mathcal{S}^F$, (2.10) with $A = \Delta X$, (2.8) with $B = \Delta S$, (2.2), and (2.3), we see that

$$\begin{aligned}
\Delta X \bullet S + X \bullet \Delta S &= \hat{X}' \bullet S + \hat{X} \bullet \Delta S \\
&= \big(-\hat{X}(V'V^T + V(V')^T)\hat{X}\big) \bullet LL^T + V^{-T}V^{-1} \bullet (L'L^T + L(L')^T) \\
&= 2\,V^{-T}V^{-1} \bullet L'L^T - 2\,\hat{X}V'V^T\hat{X} \bullet LL^T \\
&= 2\,V^{-T}V^{-1} \bullet L'L^T - 2\,V^{-T}V^{-1}V'V^{-1} \bullet LL^T \\
&= 2\,V^{-1}L \bullet \big(V^{-1}L' - V^{-1}V'V^{-1}L\big) \qquad (4.14) \\
&\leq 2\left((\sigma-1)n + \frac{1}{2}\tau(\gamma+1)\gamma\right)\rho^2,
\end{aligned}$$

where the inequality follows from (4.10). By the definition of ζ, the inequality just proven, (4.5), and (4.6), we have

$$\zeta \leq 2(\sigma-1) + \tau(\gamma+1)\frac{\gamma}{n} \leq -\frac{2\delta}{\sqrt{n}} + \frac{\delta}{\sqrt{n}} = -\frac{\delta}{\sqrt{n}}.$$

The conclusion that the duality gap measures $\mu(\cdot,\cdot)$ and $\rho(\cdot,\cdot)$ both decrease along $(\Delta X, \Delta S, \Delta y)$ can now be seen from Proposition 4.1. □

PROPOSITION 4.7. *Let $(X, S, y) \in \mathcal{N}(\gamma)$, and define the constant Γ as in Proposition* 3.1 *and the functions Φ, $\hat{V}$, and $\hat{L}$ as in Proposition* 3.2. *Suppose that $\alpha \geq 0$ satisfies the inequality*

$$\alpha \leq \frac{\gamma^2(\sigma - \tau\theta)(1-\tau)}{2\sqrt{2}\,\tau^2(\gamma+1)}\frac{1}{\sqrt{n}}, \qquad (4.15)$$

where $\theta \equiv 1 + \gamma(\gamma+1)/(2n)$. Then

$$\mathrm{tr}(\Phi(\hat{X} + \alpha\hat{X}', S_\alpha)) = \mathrm{tr}(\hat{V}(\hat{X} + \alpha\hat{X}')^{-1}\hat{L}(S_\alpha)) \geq \Gamma\rho_\alpha,$$

where $\hat{X} \equiv \hat{X}(X)$ and $\hat{X}' \equiv \hat{X}'(X)[\Delta X]$.

Proof. We first remark that the right-hand side of (4.15) is nonnegative by (4.7), and is less than or equal to 1 by (4.8), and thus $0 \leq \alpha \leq 1$, which clearly shows that $\hat{X} + \alpha\hat{X}' \in \mathcal{S}^n_{++}$ and $S_\alpha \in \mathcal{S}^n_{++}$ by Proposition 4.3. Hence, Φ is defined at $(\hat{X} + \alpha\hat{X}', S_\alpha)$.

Define $\hat{V} \equiv \hat{V}(\hat{X})$, $\hat{V}' \equiv \hat{V}'(\hat{X})[\Delta X]$, $\hat{L} \equiv \hat{L}(S)$, and $\hat{L} \equiv \hat{L}'(S)[\Delta S]$. In addition, define $V \equiv V(X)$, $V' \equiv V'(X)[\Delta X]$, $L \equiv L(S)$, and $L' \equiv L'(S)[\Delta S]$. Noting that $V(\cdot) = \hat{V}(\hat{X}(\cdot))$ and that $\hat{L}(\cdot)$ is identical to $L(\cdot)$ on the domain $\mathcal{S}^F_{++}$, we see that

$$V = \hat{V}, \quad V' = \hat{V}', \quad L = \hat{L}, \quad L' = \hat{L}'. \qquad (4.16)$$

The Taylor integral formula implies that

$$\Phi(\hat{X} + \alpha\hat{X}', S_\alpha) = \Phi(\hat{X}, S) + \alpha\,\Phi'(\hat{X}, S)[\hat{X}', \Delta S] + \alpha^2\,T_\alpha, \qquad (4.17)$$

where

$$T_\alpha \equiv \int_0^1 (1-t)\Phi''(\hat{X} + t\alpha\hat{X}', S_{t\alpha})[\hat{X}', \Delta S]^{(2)} dt. \tag{4.18}$$

Analyzing the first two components of (4.17), we first see by (4.16) that $\Phi(\hat{X}, S) = \hat{V}^{-1}\hat{L} = V^{-1}L$. Secondly, letting $\rho \equiv \rho(X, S)$, we have

$$\begin{aligned}
\Phi'(\hat{X}, S)[\hat{X}', \Delta S] &= -\hat{V}^{-1}\hat{V}'\hat{V}^{-1}\hat{L} + \hat{V}^{-1}\hat{L}' \\
&= V^{-1}L' - V^{-1}V'V^{-1}L \\
&= \sigma\rho I - V^{-1}L + V^{-1}V'(\rho I - V^{-1}L),
\end{aligned}$$

where the third equality comes from substituting for $V^{-1}L'$ as was done in the proof of Lemma 4.5. Hence, we can rewrite (4.17) as

$$\begin{aligned}
\Phi(\hat{X} + \alpha\hat{X}', S_\alpha) &= V^{-1}L + \alpha\left(\sigma\rho I - V^{-1}L + V^{-1}V'(\rho I - V^{-1}L)\right) + \alpha^2 T_\alpha \\
&= (1-\alpha)V^{-1}L + \alpha\sigma\rho I + \alpha P + \alpha^2 T_\alpha,
\end{aligned} \tag{4.19}$$

where $P \equiv V^{-1}V'(\rho I - V^{-1}L)$. Taking the trace of (4.19) and using Proposition 3.1 and Proposition 4.1, where $\zeta \equiv (\Delta X \bullet S + X \bullet \Delta S)/(n\rho^2)$, we see

$$\begin{aligned}
\operatorname{tr}\big(\Phi(\hat{X} + \alpha\hat{X}', S_\alpha)\big) &\geq (1-\alpha)\Gamma\rho + \alpha\sigma n\rho + \alpha\operatorname{tr}(P) + \alpha^2 \operatorname{tr}(T_\alpha) \\
&\geq \Gamma\Big(\rho_\alpha - \alpha\rho\Big(1 + \frac{\zeta}{2}\Big)\Big) + \alpha\sigma n\rho + \alpha\operatorname{tr}(P) + \alpha^2 \operatorname{tr}(T_\alpha) \\
&= \Gamma\rho_\alpha - \alpha\Gamma\rho\left(1 + \frac{1}{2n\rho^2}\left(\Delta X \bullet S + X \bullet \Delta S\right)\right) + \alpha\sigma n\rho + \alpha\operatorname{tr}(P) + \alpha^2 \operatorname{tr}(T_\alpha) \\
&= \Gamma\rho_\alpha - \alpha\Gamma\rho + \alpha\sigma n\rho + \alpha\rho^{-1}\left(\rho\operatorname{tr}(P) - \frac{\Gamma}{2n}\left(\Delta X \bullet S + X \bullet \Delta S\right)\right) + \alpha^2 \operatorname{tr}(T_\alpha).
\end{aligned}$$

From (4.14) (which is inside the proof of Proposition 4.6), we have $\Delta X \bullet S + X \bullet \Delta S = 2\,V^{-1}L \bullet (V^{-1}L' - V^{-1}V'V^{-1}L)$. Thus, letting $\theta \equiv 1 + \gamma(\gamma+1)/(2n)$, we can apply (4.11) to the above inequality to get

$$\begin{aligned}
\operatorname{tr}\big(\Phi(\hat{X} + \alpha\hat{X}', S_\alpha)\big) &\geq \Gamma\rho_\alpha - \alpha\Gamma\rho + \alpha\sigma n\rho + \alpha\rho\left(-(\sigma-1)\Gamma - \frac{\tau\gamma^2\theta}{2}\right) + \alpha^2 \operatorname{tr}(T_\alpha) \\
&= \Gamma\rho_\alpha + \alpha\sigma n\rho + \alpha\rho\left(-\sigma\Gamma - \frac{\tau\gamma^2\theta}{2}\right) + \alpha^2 \operatorname{tr}(T_\alpha) \\
&= \Gamma\rho_\alpha + \alpha\rho\left(\frac{\sigma\gamma^2}{2} - \frac{\tau\gamma^2\theta}{2}\right) + \alpha^2 \operatorname{tr}(T_\alpha) \\
&= \Gamma\rho_\alpha + \frac{\alpha\rho\gamma^2(\sigma - \tau\theta)}{2} + \alpha^2 \operatorname{tr}(T_\alpha),
\end{aligned}$$

where the second equality comes from the definition $\Gamma \equiv n - \gamma^2/2$.

From the above inequality, the statement of the proposition will hold if

$$\frac{\rho\gamma^2(\sigma - \tau\theta)}{2} + \alpha\operatorname{tr}(T_\alpha) \geq 0, \tag{4.20}$$

and so we now devote our efforts to establishing (4.20). We start by providing a bound on $\|T_\alpha\|_F$, which can be obtained as follows from (4.18), the standard properties of

integration, and Proposition 3.2(ii):

$$\begin{aligned}\|T_\alpha\|_F &\le \int_0^1 (1-t)\big\|\Phi''(\hat{X}+t\alpha\hat{X}', S_{t\alpha})[\hat{X}', \Delta S]^{(2)}\big\|_F dt \\ &\le \frac{1}{\sqrt{2}}\int_0^1 (1-t)\|\hat{V}_{t\alpha}^{-1}\hat{L}_{t\alpha}\|\big(\|\hat{V}_{t\alpha}^T\hat{X}'\hat{V}_{t\alpha}\|_F + \|\hat{L}_{t\alpha}^{-1}\Delta S\hat{L}_{t\alpha}^{-T}\|_F\big)^2 dt \\ &\le \frac{1}{\sqrt{2}}\int_0^1 (1-t)\|\hat{V}_{t\alpha}^{-1}\hat{L}_{t\alpha}\|\big(\|V^{-1}\hat{V}_{t\alpha}\|^2\|V^T\hat{X}'V\|_F \\ &\qquad + \|\hat{L}_{t\alpha}^{-1}L\|^2\|L^{-1}\Delta SL^{-T}\|_F\big)^2 dt,\end{aligned}$$

where $\hat{V}_{t\alpha} \equiv \hat{V}(\hat{X}+t\alpha\hat{X}')$ and $\hat{L}_{t\alpha} \equiv \hat{L}(S_{t\alpha})$. We note that the hypotheses of Proposition 3.7 hold with $(\hat{X}, S)$, $(\hat{X}+t\alpha\hat{X}', S_{t\alpha})$, and the scalar $t\alpha\tau$ due to Proposition 4.3. Now using Propositions 4.3, 3.7, and 3.6(i) as well as (4.16) and simple integration with respect to t, the above inequality shows that

$$\begin{aligned}\|T_\alpha\|_F &\le \frac{1}{\sqrt{2}}\int_0^1 (1-t)\frac{\|\hat{V}^{-1}\hat{L}\|}{1-t\alpha\tau}\left((1-t\alpha\tau)^{-1}\tau + (1-t\alpha\tau)^{-1}\tau\right)^2 dt \\ &= 2\sqrt{2}\,\tau^2\|V^{-1}L\|\int_0^1 \frac{1-t}{(1-t\alpha\tau)^3}\,dt \\ &\le 2\sqrt{2}\,\tau^2\left((\gamma+1)\rho\right)\left(2(1-\alpha\tau)\right)^{-1} \\ &\le \sqrt{2}\,\tau^2(1-\tau)^{-1}(\gamma+1)\rho.\end{aligned}$$

Hence,

$$\begin{aligned}\frac{\rho\gamma^2(\sigma-\tau\theta)}{2} + \alpha\mathrm{tr}(T_\alpha) &\ge \frac{\rho\gamma^2(\sigma-\tau\theta)}{2} - \alpha\left|\mathrm{tr}(T_\alpha)\right| \ge \frac{\rho\gamma^2(\sigma-\tau\theta)}{2} - \alpha\sqrt{n}\,\|T_\alpha\|_F \\ &\ge \frac{\rho\gamma^2(\sigma-\tau\theta)}{2} - \alpha\sqrt{2}\tau^2(1-\tau)^{-1}(\gamma+1)\sqrt{n}\rho \\ &= \rho\left(\frac{\gamma^2(\sigma-\tau\theta)}{2} - \alpha\sqrt{2}\tau^2(1-\tau)^{-1}(\gamma+1)\sqrt{n}\right) \ge 0,\end{aligned}$$

where the last inequality follows from (4.15). This completes the proof of the proposition. □

COROLLARY 4.8. *Let $(X, S, y) \in \mathcal{N}(\gamma)$, and suppose that $\alpha \ge 0$ satisfies* (4.15). *Then $(X_\alpha, S_\alpha, y_\alpha) \in \mathcal{N}(\gamma)$.*

Proof. As discussed in the proof of Proposition 4.7, we have $\alpha \le 1$, and so $(X_\alpha, S_\alpha) \in \mathcal{S}_{++}^{F?} \times \mathcal{S}_{++}^{F}$ by Corollary 4.4. In addition, the Newton system defining $(\Delta X, \Delta S, \Delta y)$ clearly shows that $(X_\alpha, S_\alpha, y_\alpha) \in \mathcal{F}^0(P) \times \mathcal{F}^0(D)$.

As was shown in the proof of Corollary 4.4, $\hat{X}+\alpha\hat{X}'$ is a positive definite completion of X_α, where $\hat{X} \equiv \hat{X}(X)$ and $\hat{X}' \equiv \hat{X}'(X)[\Delta X]$. By Theorem 3.3, we have that, among all positive definite completions of X_α, the maximum-determinant completion $\hat{X}(X_\alpha)$ of X_α maximizes the function $\mathrm{tr}(\hat{V}(\cdot)^{-1}\hat{L}(S_\alpha))$, where $\hat{V}(\cdot)$ and $\hat{L}(\cdot)$ are as in Proposition 3.2. Thus, using the fact that $V(\cdot) = \hat{V}(\hat{X}(\cdot))$ and combining Theorem 3.3 and Proposition 4.7, we see

$$\mathrm{tr}(V(X_\alpha)^{-1}L(S_\alpha)) = \mathrm{tr}(\hat{V}(\hat{X}(X_\alpha))^{-1}\hat{L}(S_\alpha)) \ge \mathrm{tr}(\hat{V}(\hat{X}+\alpha\hat{X}')^{-1}\hat{L}(S_\alpha)) \ge \Gamma\rho_\alpha.$$

Combining this inequality with the fact that $(X_\alpha, S_\alpha, y_\alpha) \in \mathcal{F}^0(P) \times \mathcal{F}^0(D)$, and applying Proposition 3.1, we conclude that $(X_\alpha, S_\alpha, y_\alpha) \in \mathcal{N}(\gamma)$. □

THEOREM 4.9. *Let $n \geq 1$ and constants $\gamma \geq 0$, $\delta \in [0, \sqrt{n})$, and $\tau \in (0,1)$ be given such that (4.3)–(4.8) are satisfied. Suppose that Algorithm SDP is initialized with $(X^0, S^0, y^0) \in \mathcal{N}(\gamma)$ and a tolerance $\varepsilon \geq 0$, and suppose that the constant step-size $\alpha \geq 0$ is used in each iteration of the algorithm, where α is given by the right-hand side of (4.15). Then, for each $k \geq 0$, the sequence $\{(X^k, S^k, y^k)\}_{k \geq 0}$ produced by the algorithm satisfies the following:*

(i) $(X^k, S^k, y^k) \in \mathcal{N}(\gamma)$;
(ii) $\mu(X^k, S^k) \leq (1 - \alpha\,\delta/\sqrt{n})^k \mu(X^0, S^0)$.

As a consequence, Algorithm SDP terminates with a point (X^k, S^k, y^k) satisfying $X^k \bullet S^k \leq \varepsilon$ in at most $\mathcal{O}(n \log(X^0 \bullet S^0/\varepsilon))$ iterations.

Proof. Item (i) follows from Corollary 4.8 and induction on k. Likewise, item (ii) follows from the combination of Propositions 4.1 and 4.6 along with induction on k. The bound on the number of iterations now follows from (ii) and the standard argument that shows that the duality gap is reduced in each iteration by a factor on the order of $\mathcal{O}(1 - \alpha\,\delta/\sqrt{n})$, which is $\mathcal{O}(1 - 1/n)$ due to (4.15). ∎

5. Computational issues and results. In this section, we discuss several computational issues related to Algorithm SDP, and then present computational results comparing our method with three other SDP implementations.

5.1. Implementation features of Algorithm SDP. We now demonstrate how the theoretical presentation of Algorithm SDP in section 4 can be specified into a practical implementation.

First, as is typical with practical primal-dual interior-point algorithms, we implement Algorithm SDP as an infeasible method. Thus, we do not require full primal-dual feasibility of the iterates (X, S, y) but rather require only $X \in \mathcal{S}^{F?}_{++}$ and $S \in \mathcal{S}^{F}_{++}$ and then define the Newton direction $(\Delta X, \Delta S, \Delta y)$ by

$$H'(X,S,y)[\Delta X, \Delta S, \Delta y] = (b - \mathcal{A}(X), C - \mathcal{A}^*(y) - S, \sigma\,\rho\,V - L), \tag{5.1}$$

where $\rho \equiv \rho(X,S)$ and $0 \leq \sigma < 1$, instead of by (4.1). In particular, this makes the choice of initial iterate trivial. (In fact it is chosen as described in the SDPT3 user's guide [31] in all of the computational results presented in the following subsections.)

Second, we choose a different stopping criterion than that originally given for Algorithm SDP. A more practical stopping criterion is based on the relative duality gap and relative feasibility of the current iterate (X, S, y), which we define respectively as

$$\frac{X \bullet S}{1 + (|C \bullet X| + |b^T y|)/2} \qquad \text{and} \qquad \max\left\{ \frac{\|b - \mathcal{A}(X)\|}{1 + \|b\|}, \frac{\|C - \mathcal{A}^*(y) - S\|_F}{1 + \|C\|_F} \right\}.$$

Target values for the gap and feasibility can then be specified at run time.

Third, in our implementation of Algorithm SDP, we do not bother to stay in the neighborhood $\mathcal{N}(\gamma)$, since our computational experience indicates that this does not yield a substantial practical improvement. Instead, we essentially take α as large as possible, while keeping $(X_\alpha, S_\alpha, y_\alpha)$ in $\mathcal{S}^{F?}_{++} \times \mathcal{S}^{F}_{++} \times \Re^m$. In fact, as is common in SDP implementations, we differentiate two step-sizes, α_p for the primal and α_d for the dual. Then, for the primal and dual separately, the actual step-size is calculated by estimating the infeasible boundary step-size $\bar{\alpha}$ to within an absolute accuracy of 1.0e−2 using a simple bisection method and then choosing the step-size slightly less than $\min(\bar{\alpha}, 1)$.

Fourth, we have found it advantageous for reducing the duality gap in the early stages of Algorithm SDP to update X and S in each iteration by performing an alternative linesearch in the spaces of V and L. More specifically, we choose $\alpha_p \leq 1$ and $\alpha_d \leq 1$ so that $L_{\alpha_d} \equiv L(S) + \alpha_d L'(S)[\Delta S]$ and $V_{\alpha_p} \equiv V(X) + \alpha_p V'(X)[\Delta X]$ are close to the boundary of $\mathcal{L}^F_{++}$, and we then define $X_{\alpha_p} = \pi^F(V_{\alpha_p}^{-T} V_{\alpha_p}^{-1})$ and $S_{\alpha_d} = L_{\alpha_d} L_{\alpha_d}^T$. (We note that the calculation of X_{α_p} and S_{α_d} can be done efficiently; see below.) This update method, however, does not effectively achieve feasibility, and so it is always necessary to revert back to the typical linesearch in the space of X and S.

Fifth, our choice of σ in each iteration is adaptive rather than constant as in the statement of Algorithm SDP. Roughly speaking, we choose σ conservatively whenever we are experiencing small step-sizes, but then more aggressively when our step-sizes are larger. In particular, we set $\sigma = 1 - \gamma \min(\alpha_p, \alpha_d)$, where α_p and α_d are the successful step-sizes of the preceding iteration and $\gamma \in [0.1, 0.3]$ is an adaptive parameter that is incrementally increased when step-sizes satisfying $\alpha_p = \alpha_d = 1$ are encountered, and incrementally reduced otherwise. As such, our typical values for σ are smaller than those for other SDP implementations. (For example, SDPT3 employs a constant $\gamma = 0.9$.) In addition, it is not immediately clear whether a predictor-corrector approach will improve the choice of σ or even whether such an approach would be computationally efficient (see following subsections), and so our current implementation of Algorithm SDP does not use a predictor-corrector strategy.

Finally, we mention some details regarding the calculation of the Newton direction $(\Delta X, \Delta S, \Delta y)$. As with other primal-dual path-following algorithms, the calculation can be reduced to the solution of the system $M\Delta y = h$, where M is the so-called Schur complement matrix and h is in accordance with the system (5.1). Here, M is the $m \times m$ matrix representation of the linear operator $-\mathcal{A} \circ V'(X)^{-1} \circ L'(S) \circ \mathcal{A}^*$, so that M is positive definite by Corollary 2.9. Two fundamental techniques for calculating Δy can then be considered: either (1) solution of the system via forward and backward substitution after the direct formation and factorization of M; or (2) solution of the equation via an iterative method. We present numerical results for both methods in later subsections. It is important to note, however, that M has no inherent sparsity (as is typical with other methods) and that M is nonsymmetric (which is atypical). Hence, the first method for calculating Δy requires Gaussian elimination with pivoting, and the second requires an efficient iterative method for nonsymmetric systems (like BiCGSTAB, which we have chosen in the computational results). Thus, the ill-conditioning of M near optimality can have a negative impact upon both methods. A natural way to reduce this impact in the case of BiCGSTAB is to perform some pre-conditioning of the linear system, but this has not been implemented in the current version of Algorithm SDP since more investigation is necessary to develop reasonable preconditioning strategies.

Having described the key implementation choices of Algorithm SDP, we now consider the basic operations of the algorithm and in particular discuss their computational complexities. From the statement of Algorithm SDP and the definition of the Newton direction, we see that the main operations are checking $X \in \mathcal{S}^{F?}_{++}$ and $S \in \mathcal{S}^F_{++}$ and evaluating $V(X)$, $V'(X)^{-1}[N]$, $L(S)$, and $L'(S)[B]$ for any $N \in \mathcal{L}^F$ and $B \in \mathcal{S}^F$. To describe the complexity of these operations, we introduce a few definitions. For each $j = 1, \ldots, n$ we define

$$K_j \equiv \{i \in V : (i,j) \in F,\ i \geq j\}.$$

That is, for each $j = 1, \ldots, n$, K_j is the set of row indices of the jth column of the lower part of F. We have the following fact (detailed in Fukuda et al. [9]), which expresses the chordal structure F as a union of dense blocks, or cliques:

$$F = \bigcup_{j=1}^{n} K_j \times K_j. \tag{5.2}$$

We also define

$$f_2 \equiv \sum_{j=1}^{n} |K_j|^2.$$

A common way to check whether S is in $\mathcal{S}^F_{++}$ is to simply attempt the calculation of $L(S)$ (then $S \in \mathcal{S}^F_{++}$ if and only if the calculation is successful), and standard methods for calculating $L(S)$ show that the time required is $\mathcal{O}(f_2)$. Moreover, in a similar manner, the defining equation (2.8) shows that $L'(S)[B]$ can also be calculated in time $\mathcal{O}(f_2)$. Hence, each of the key operations involving S is $\mathcal{O}(f_2)$.

To calculate the times required for the key operations involving X, we introduce some additional notation and ideas. First, for any $P, Q \subseteq V$ and any $W \in \mathcal{S}^n$, we let $W_{PQ} \in \Re^{|P|\times|Q|}$ denote the matrix obtained from W by deleting all rows $p \notin P$ and all columns $q \notin Q$. Second, it is clear that (5.2) can be simplified to $F = \cup_{r=1}^{\ell} C_r \times C_r$, where $\{C_r\}_{r=1}^{\ell}$ are the maximal members of $\{K_j\}_{j=1}^{n}$, i.e., those members of $\{K_j\}_{j=1}^{n}$ that are not properly contained in any other members. We then define

$$f_3 \equiv \sum_{r=1}^{\ell} |C_r|^3$$

and have the following critical theorem, proved in [12].

THEOREM 5.1. *Let $X \in \mathcal{S}^F$. Then*

(i) *$X \in \mathcal{S}^{F?}_{+}$ if and only if $X_{C_rC_r} \in \mathcal{S}^{|C_r|}_{+}$ for all $r = 1, \ldots, \ell$;*

(ii) *$X \in \mathcal{S}^{F?}_{++}$ if and only if $X_{C_rC_r} \in \mathcal{S}^{|C_r|}_{++}$ for all $r = 1, \ldots, \ell$.*

This theorem shows immediately that testing whether $X \in \mathcal{S}^{F?}_{++}$ can be done in time $\mathcal{O}(f_3)$ by simply attempting the factorizations of the submatrices $X_{C_rC_r}$ of X.

Next, we determine the time required for $V(X)$ and $V'(X)^{-1}[A]$ by considering the following proposition, which gives a formula for $V(X)$ that will be convenient for computation. In the proposition, π^{Fl} is the operator that is defined similarly to π^F except that it projects onto $\mathcal{L}^F$ instead of $\mathcal{S}^F$.

PROPOSITION 5.2. *Let $X \in \mathcal{S}^{F?}_{++}$. Then $V \equiv V(X) \in \mathcal{L}^F_{++}$ uniquely satisfies the equation*

$$\pi^{Fl}(XV) = \pi^{Fl}(V^{-T}). \tag{5.3}$$

Proof. We first consider solving the simpler equation $\pi^{Fl}(XQ) = 0$ for $Q \in \mathcal{L}^F$. Because the jth column of this equation can be expressed compactly as $X_{K_jK_j}q_j = 0$, where q_j is the nonzero part of $Q_{\cdot j}$, we conclude from Theorem 5.1(ii) that $q_j = 0$, which implies that $Q = 0$.

Now suppose that $V_1, V_2 \in \mathcal{L}^F_{++}$ each satisfy (5.3), and note that the right-hand side of (5.3) is a diagonal matrix. For $i = 1, 2$, let D_i be the diagonal matrix defined by the diagonal entries of V_i. Then (5.3) implies that $\pi^{Fl}(XV_1D_1) = \pi^{Fl}(XV_2D_2)$ or, equivalently, that $\pi^{Fl}(X(V_1D_1 - V_2D_2)) = 0$. Using the result of the previous

paragraph, this shows $V_1 D_1 = V_2 D_2$. Examining the jjth position of this equation and employing the definition of D_i, we see that $[V_1]_{jj} = [V_2]_{jj}$, which in turn implies that $D_1 = D_2$ and hence $V_1 = V_2$. In other words, at most one $V \in \mathcal{L}^F_{++}$ satisfies (5.3).

We now show that $V \equiv V(X)$ satisfies (5.3), which will complete the proof of the proposition. For $i \geq j$ such that $(i,j) \in F$, consider the ijth entry of the matrix $(\hat{X} - X)V$, where $\hat{X} \equiv \hat{X}(X)$:

$$[(\hat{X} - X)V]_{ij} = \sum_{k=j}^{n} [\hat{X} - X]_{ik} V_{kj}.$$

We claim that the above expression equals zero. So suppose for contradiction that it is nonzero. Then there exists $k \geq j$ such that $[\hat{X} - X]_{ik} \neq 0$ and $V_{kj} \neq 0$. Because $\pi^F(\hat{X}) = X$ and $V \in \mathcal{L}^F_{++}$, this implies that $(i,k) \notin F$ and $(k,j) \in F$. However, due to the chordal structure of F and the fact that the ordering $(1, \ldots, n)$ is a perfect elimination ordering for F, we have that $(i,j), (k,j) \in F$ imply $(i,k) \in F$, which is a contradiction. Hence, the above expression equals 0. Said differently, we have $\pi^{Fl}((\hat{X} - X)V) = 0$, which implies $\pi^{Fl}(XV) = \pi^{Fl}(\hat{X}V) = \pi^{Fl}(V^{-T})$, as desired. □

As the proposition and its proof indicate, the nonzero part $v_j \in \Re^{|K_j|}$ of the jth column of V is simply the solution of the system $X_{K_j K_j} v_j = V_{jj}^{-1} e_1$, where $e_1 \in \Re^{|K_j|}$ has a one in its first position and zeros elsewhere. Hence, as long as the factorizations of $X_{K_j K_j}$ for $j = 1, \ldots, n$ are readily available, the calculation of V can be done in time $\mathcal{O}(f_2)$.

Are these factorizations readily available, however? The operation to verify $X \in \mathcal{S}^{F?}_{++}$ yields the factorizations of $X_{C_r C_r}$ only for $r = 1, \ldots, \ell$, and so the factorizations of $X_{K_j K_j}$ are not explicitly available. This is not a significant obstacle, however, since it is possible to reorder the vertices V (in a preprocessing phase, for example) so that factorizations of the matrices $X_{K_j K_j}$ are embedded in a natural manner in the upper Cholesky factorizations of the matrices $X_{C_r C_r}$. Moreover, this reordering can be done without altering the chordal structure of F or the property that $(1, \ldots, n)$ is a perfect elimination ordering. This property is discussed in detail in section 2.1 of [9] and section 2.2 of [26], where it is described as a *perfect elimination ordering induced from an ordering of maximal cliques satisfying the running intersection property*, and this feature has been incorporated into Algorithm SDP.

Differentiating (5.3) with respect to X in the direction $A \in \mathcal{S}^F$ and defining $N \equiv V'(X)[A]$, we see that

$$\pi^{Fl}(AV) = -\pi^{Fl}(V^{-T} N^T V^{-T}) - \pi^{Fl}(XN).$$

Note that the first term on the right-hand side does not require the full matrix $V^{-T} N^T V^{-T}$ but rather just its diagonal. The above equation also provides a convenient form for calculating $A = V'(X)^{-1}[N]$ for an arbitrary $N \in \mathcal{L}^F$, once X and V are available. In fact, it is not difficult to see that A can be computed from N in time $\mathcal{O}(f_2)$.

We summarize the complexities obtained from the preceding discussion in the following proposition.

PROPOSITION 5.3. *Let $X, S \in \mathcal{S}^F$. Determining whether X is in $\mathcal{S}^{F?}_{++}$ requires time $\mathcal{O}(f_3)$, and under the mild assumption that certain calculations are stored upon the successful determination of $X \in \mathcal{S}^{F?}_{++}$, calculation of $V(X)$ and $V'(X)^{-1}[N]$, for*

any $N \in \mathcal{L}^F$, requires time $\mathcal{O}(f_2)$. Determining whether S is in $\mathcal{S}^F_{++}$ is performed by attempting the computation of $L(S)$, which requires time $\mathcal{O}(f_2)$. Upon the successful computation of $L(S)$, the calculation of $L'(S)[B]$, for any $B \in \mathcal{S}^F$, requires time $\mathcal{O}(f_2)$.

5.2. Comparison with a standard primal-dual method. In order to see how Algorithm SDP compares with other primal-dual path-following implementations, in this subsection we compare Algorithm SDP with SDPT3 version 3.0, a successful implementation by Tütüncü, Toh, and Todd (see [31]). We have chosen to run SDPT3 with the HRVW/KSH/M direction using the Mehrotra predictor-corrector strategy. Both algorithms use the same starting point and terminate when the relative duality gap is less than 1.0e−4 and the relative feasibility is less than 1.0e−5.

We remark that these moderate values for the target gap and feasibility have been chosen for two reasons. First, it makes our presentation consistent with later subsections where larger SDPs are considered and solved to the same accuracies. Second, we have chosen moderate values in keeping with our discussion of the previous subsection concerning how the ill-conditioning of M near optimality affects the calculation of Δy in Algorithm SDP. In fact, for all but a few of the problems presented below (most notably the *control* and *qap* instances), Algorithm SDP can easily obtain even higher accuracy.

For these computational results, we solve the system $M\Delta y = h$ in Algorithm SDP using either Gaussian elimination or BiCGSTAB, depending on the stage of the algorithm. In the early stages of the algorithm when the conditioning of M is good, we employ BiCGSTAB. As soon as the number of multiplications by M (or equivalently, the number of evaluations of $-\mathcal{A} \circ V'(X)^{-1} \circ L'(S) \circ \mathcal{A}^*$) exceeds m during one call to the BiCGSTAB subroutine, however, we switch to Gaussian elimination. Assuming that evaluations of the functions $\mathcal{A}$ and $\mathcal{A}^*$ require time $\mathcal{O}(f_2)$ (as is the case for most of the test problems below), the direct method requires time $\mathcal{O}(mf_2)$ to form the matrix M and then an additional $\mathcal{O}(m^3)$ time to factor M and solve for Δy. We have thus chosen problems having $mf_2 + m^3 \leq$ 1.0e+9 as a heuristic guide for selecting problems for which the direct method is not too computationally intensive. In particular, no problems having $m > 1000$ have been selected.

The test problems come from the SDPLIB collection of problems maintained by Borchers [3], and their statistics are listed in Table 1. The first three columns are self-explanatory, and the last two give the percentage of nonzeros in the iterates S and X represented by the density patterns E and F. Complete computational results on a 2.4 GHz Pentium 4 computer are given in Table 2 and are self-explanatory, with times given in seconds.

We remark that, although Algorithm SDP is designed primarily for sparse problems, i.e., when F is a relatively small subset of $V \times V$, it can of course be applied to dense problems with $F = V \times V$, as with a few of the test problems in Table 1. We have included these problems because we feel it is instructive to compare the performance of Algorithm SDP and SDPT3 on such instances.

The results indicate several interesting points. First and foremost, both methods were able to solve all problems to the desired accuracy in a reasonable amount of time. Second, Algorithm SDP outperformed SDPT3 on several sets problems (the *arch*, *max*, *qp*, and *ss* problems, as well as a subset of the *mcp* problems) for which the chordal pattern F was very small, indicating Algorithm SDP's capability for exploiting this structure. On the other hand, SDPT3 significantly outperformed Algorithm SDP on the *control* and *qap* problems, as Algorithm SDP was challenged by the

TABLE 1
Statistics for SDPLIB test problems.

Problem	n	m	Dens E (%)	Dens F (%)
arch2	335	174	3.29	6.87
arch4	335	174	3.29	6.87
arch8	335	174	3.29	6.87
control5	75	351	45.61	46.88
control6	90	496	45.42	46.76
control7	105	666	45.28	46.68
control8	120	861	45.18	46.43
gpp100	100	101	100.00	100.00
gpp124-2	124	125	100.00	100.00
gpp124-3	124	125	100.00	100.00
gpp124-4	124	125	100.00	100.00
maxG11	800	800	0.75	2.61
mcp250-2	250	250	2.75	15.27
mcp250-3	250	250	4.89	36.76
mcp250-4	250	250	8.51	58.63
mcp500-1	500	500	0.90	2.58
mcp500-2	500	500	1.38	11.99
qap7	50	358	100.00	100.00
qap8	65	529	100.00	100.00
qap9	82	748	100.00	100.00
qpG11	1600	800	0.25	0.73
ss30	426	132	4.43	9.85
theta1	50	104	100.00	100.00
theta2	100	498	100.00	100.00
truss5	331	208	3.31	3.31
truss6	451	172	0.88	0.88
truss7	301	86	0.99	0.99
truss8	628	496	3.18	3.18

conditioning and density of these problems. In addition, SDPT3 was faster on the remaining *mcp* problems, most likely due to the related fact that Algorithm SDP consistently required more iterations than SDPT3, which itself is indicative of the strong convergence properties of the HRVW/KSH/M direction when combined with the Merhrotra predictor-corrector strategy.

5.3. Comparison with the completion method. In this subsection, we compare Algorithm SDP with the completion method (CM) of Fukuda et al. on problems for which the large size of m requires the solution of the Schur complement equation by an iterative method. As in the previous subsection, Algorithm SDP is run with BiCGSTAB as the iterative solver, and the SDPs are solved to an accuracy of 1.0e−4 for the relative duality gap and 1.0e−5 for relative feasibility.

As described in [9, 26], CM stores X and S in the same sparse format as Algorithm SDP does, and the search direction in each iteration is the sparse projection of the HRVW/KSH/M direction. Moreover, the sparsity of X and S is exploited in the formation of the Schur complement matrix M, which is then factored directly to solve for Δy. Here, however, we have implemented our own version of CM which computes Δy using an iterative method, namely, the conjugate gradient method, which is appropriate since, in this case, M is symmetric positive definite. Other algorithmic choices for our implementation of CM mimic those of SDPT3 in the previous subsection, except that the predictor-corrector method has not been implemented due to its need to solve an extra $m \times m$ system in each iteration.

TABLE 2
Results of Algorithm SDP (AS) and SDPT3 (T3) on the SDPLIB test problems.

PROBLEM	OBJECTIVE VALUE				R-GAP		R-FEAS		ITER		TIME (s)	
	AS		T3		AS	T3	AS	T3	AS	T3	AS	T3
	PRIMAL	DUAL	PRIMAL	DUAL								
arch2	−6.71397e−01	−6.71546e−01	−6.71485e−01	−6.71523e−01	8.9e−05	3.8e−05	7.6e−13	8.4e−12	50	17	8	20
arch4	−9.72501e−01	−9.72657e−01	−9.72561e−01	−9.72644e−01	7.9e−05	8.3e−05	1.4e−11	2.9e−11	52	18	10	20
arch8	−7.05642e+00	−7.05716e+00	−7.05689e+00	−7.05702e+00	9.2e−05	1.7e−05	1.4e−10	3.1e−10	42	17	8	19
control5	−1.68824e+01	−1.68838e+01	−1.68835e+01	−1.68836e+01	7.8e−05	1.1e−05	1.4e−08	1.3e−07	52	18	37	11
control6	−3.73018e+01	−3.73048e+01	−3.73035e+01	−3.73047e+01	7.8e−05	3.0e−05	1.7e−08	4.4e−08	52	19	95	21
control7	−2.06237e+01	−2.06253e+01	−2.06249e+01	−2.06251e+01	7.6e−05	1.3e−05	2.7e−08	1.2e−07	50	21	207	43
control8	−2.02856e+01	−2.02865e+01	−2.02861e+01	−2.02865e+01	4.6e−05	2.0e−05	1.5e−08	5.5e−08	51	20	429	71
gpp100	4.49437e+01	4.49435e+01	4.49445e+01	4.49435e+01	2.2e−06	2.3e−05	8.2e−06	3.1e−10	42	12	7	1
gpp124-2	4.68623e+01	4.68623e+01	4.68639e+01	4.68623e+01	8.1e−07	3.5e−05	7.0e−06	4.2e−10	44	12	15	2
gpp124-3	1.53014e+02	1.53014e+02	1.53017e+02	1.53014e+02	2.0e−06	2.1e−05	7.8e−06	4.1e−10	41	11	15	2
gpp124-4	4.18990e+02	4.18987e+02	4.19008e+02	4.18987e+02	5.7e−06	5.1e−05	9.1e−06	1.5e−09	38	12	16	2
maxG11	−6.29127e+02	−6.29165e+02	−6.29127e+02	−6.29165e+02	6.1e−05	6.0e−05	5.6e−06	9.8e−16	39	11	20	65
mcp250-2	−5.31888e+02	−5.31932e+02	−5.31926e+02	−5.31930e+02	8.2e−05	8.3e−06	2.3e−06	7.1e−16	34	10	5	4
mcp250-3	−9.81095e+02	−9.81175e+02	−9.81168e+02	−9.81173e+02	8.1e−05	5.0e−06	3.2e−06	1.3e−15	34	10	26	4
mcp250-4	−1.68183e+03	−1.68196e+03	−1.68188e+03	−1.68196e+03	8.0e−05	5.0e−05	8.5e−06	1.8e−15	34	9	52	5
mcp500-1	−5.98103e+02	−5.98154e+02	−5.98122e+02	−5.98154e+02	8.5e−05	5.4e−05	2.9e−06	7.9e−16	35	11	3	19
mcp500-2	−1.06996e+03	−1.07006e+03	−1.07003e+03	−1.07006e+03	9.0e−05	2.3e−05	2.0e−06	4.0e−15	35	11	57	23
qap7	4.24833e+02	4.24811e+02	4.24763e+02	4.24787e+02	7.1e−05	2.1e−05	4.3e−06	2.7e−10	40	12	29	2
qap8	7.56972e+02	7.56941e+02	7.56778e+02	7.56863e+02	6.6e−05	1.3e−05	9.3e−06	3.1e−10	39	12	33	4
qap9	1.40999e+03	1.40991e+03	1.40988e+03	1.40986e+03	7.1e−05	7.6e−05	8.0e−06	4.3e−10	39	11	100	7
qpG11	−2.44858e+03	−2.44866e+03	−2.44864e+03	−2.44866e+03	3.1e−05	8.6e−06	5.4e−06	2.5e−15	39	12	27	281
ss30	−2.02378e+01	−2.02396e+01	−2.02389e+01	−2.02396e+01	8.7e−05	3.2e−05	3.0e−06	5.2e−11	46	17	7	71
theta1	−2.29992e+01	−2.30002e+01	−2.29998e+01	−2.30000e+01	4.5e−05	1.1e−05	4.9e−06	2.6e−13	39	9	1	0
theta2	−3.28782e+01	−3.28793e+01	−3.28776e+01	−3.28795e+01	3.3e−05	5.8e−05	1.2e−15	1.9e−12	42	9	48	3
truss5	1.32638e+02	1.32630e+02	1.32636e+02	1.32635e+02	6.4e−05	1.3e−05	6.5e−10	5.7e−08	45	15	5	5
truss6	9.01022e+02	9.00985e+02	9.00999e+02	9.00999e+02	4.1e−05	6.8e−06	2.6e−08	6.0e−07	43	21	2	3
truss7	9.00038e+02	8.99999e+02	9.00012e+02	9.00000e+02	4.4e−05	1.5e−05	1.3e−11	2.2e−07	42	22	0	2
truss8	1.33117e+02	1.33105e+02	1.33114e+02	1.33113e+02	8.8e−05	1.1e−05	5.1e−09	3.7e−07	48	14	66	27

TABLE 3
Statistics for test problems with large m.

Problem	n	m	Dens E (%)	Dens F (%)
brock200-1.co	200	5067	100.00	100.00
brock200-4.co	200	6812	100.00	100.00
c-fat200-1.co	200	18367	100.00	100.00
hamming6-4.co	64	1313	100.00	100.00
hamming8-4.co	256	11777	100.00	100.00
johnson16-2-4.co	120	1681	100.00	100.00
keller4.co	171	5101	100.00	100.00
san200-0.7-1.co	200	5971	100.00	100.00
sanr200-0.7.co	200	6033	100.00	100.00
MANN-a27.co	379	1081	2.03	3.12
vibra3	1185	544	0.62	1.53
vibra4	2545	1200	0.30	0.90
vibra5	6801	3280	0.11	0.45
copo14	560	1275	1.17	1.17
copo23	2300	5820	0.31	0.31

In CM, multiplication by M is equivalent to an evaluation of the operator $\mathcal{A}(\hat{X}\mathcal{A}^*(\cdot)S^{-1})$, where $\hat{X} \equiv \hat{X}(X)$. It is not difficult to see that, using the sparse matrices V and L and assuming that evaluations of $\mathcal{A}$ and $\mathcal{A}^*$ require time $\mathcal{O}(f_2)$ (as assumed in the previous subsection), this operator can be evaluated in time $\mathcal{O}(nf)$, where $f = \sum_{j=1}^{n} |K_j|$, by using sparse triangular solves. We thus have that

$$f_2 = \sum_{j=1}^{n} |K_j|^2 < \sum_{j=1}^{n} n|K_j| = nf.$$

Since f_2 is the time required to multiply by M in Algorithm SDP, this demonstrates an advantage of Algorithm SDP over CM.

For comparing Algorithm SDP with CM, we have chosen fifteen test problems from several sources; the problems are listed in Table 3. The first ten are Lovász theta SDPs based on graphs used in the Second DIMACS Challenge [16]. The first nine use the original SDP formulation of Lovász (see [21]) in which both E and F are completely dense due to the fact that C is the matrix of all ones, while the tenth employs a different formulation (see [19]), which better respects the sparsity of the underlying graph. Experimentally, we have found that the general conditioning of the first formulation is better than that of the second, and so the first formulation should be preferred whenever the underlying graph is relatively dense (which is the case for the first nine but not the tenth). In addition, in the case of the first nine problems, we have that f_2 and nf are both $\mathcal{O}(n^3)$, which allows us to compare Algorithm SDP and CM on a more equal footing for these problems. The next three, the so-called *vibra* problems, were studied in [18], and the final two were test problems in the Seventh DIMACS Challenge [7]. Both methods were given an upper bound of 5 hours computation time and so were terminated after the first iteration past this time limit, if necessary. The computational results are listed in Table 4.

The table shows that Algorithm SDP converged on all problems except the *vibra* problems, for which achieving both small gap and small feasibility was difficult. (We remark that Algorithm SDP was terminated when progress became slow.) Interestingly, however, on all three *vibra* problems, the objective values achieved by Algorithm SDP were close to optimal (see [18]). CM also failed to converge on the *vibra* problems as well as *copo*23. In terms of running times, Algorithm SDP outperformed CM in

TABLE 4
Results of Algorithm SDP (AS) and the Completion Method (CM) on test problems with large m.

PROBLEM	OBJECTIVE VALUE				R-GAP		R-FEAS		ITER		TIME (s)	
	AS		CM		AS	CM	AS	CM	AS	CM	AS	CM
	PRIMAL	DUAL	PRIMAL	DUAL								
brock200-1.co	−2.74560e+01	−2.74570e+01	−2.74554e+01	−2.74571e+01	3.7e−05	6.0e−05	9.4e−06	4.8e−07	44	20	1170	2647
brock200-4.co	−2.12929e+01	−2.12938e+01	−2.12926e+01	−2.12940e+01	4.1e−05	6.5e−05	8.8e−06	4.6e−07	44	18	1165	2466
c-fat200-1.co	−1.19995e+01	−1.20000e+01	−1.19996e+01	−1.20000e+01	4.4e−05	3.7e−05	5.0e−06	4.8e−07	44	17	717	6097
hamming6-4.co	−5.33333e+00	−5.33334e+00	−5.33320e+00	−5.33340e+00	1.7e−06	3.3e−05	9.6e−06	9.9e−08	41	12	1	0
hamming8-4.co	−1.60000e+01	−1.60000e+01	−1.59999e+01	−1.60001e+01	5.0e−06	1.1e−05	7.9e−06	3.2e−07	46	18	510	89
johnson16-2-4.co	−7.99993e+00	−8.00035e+00	−7.99999e+00	−8.00010e+00	4.5e−05	1.3e−05	3.7e−06	3.0e−14	41	12	21	3
keller4.co	−1.40119e+01	−1.40125e+01	−1.40121e+01	−1.40123e+01	4.5e−05	1.4e−05	7.1e−06	4.7e−07	44	19	861	1663
san200-0.7-1.co	−3.00001e+01	−3.00000e+01	−2.99991e+01	−3.00000e+01	1.6e−05	3.1e−05	7.5e−06	4.3e−07	45	22	311	433
sanr200-0.7.co	−2.38355e+01	−2.38365e+01	−2.38357e+01	−2.38364e+01	3.9e−05	3.2e−05	8.5e−06	4.2e−07	44	19	1191	3696
MANN-a27.co	−1.32761e+02	−1.32763e+02	−1.32751e+02	−1.32764e+02	1.8e−05	9.7e−05	1.8e−06	2.3e−06	42	31	3	270
vibra3	−1.72333e+02	−1.72684e+02	−1.72458e+02	−1.72648e+02	2.0e−03	1.1e−03	4.9e−06	8.3e−07	58	98	94	19071
vibra4	−1.65349e+02	−1.65724e+02	−1.24071e+02	−1.89703e+02	2.2e−03	4.2e−01	4.3e−09	4.4e−06	64	57	1044	18188
vibra5	−1.65570e+02	−1.66063e+02	−9.34045e+00	−7.96259e+02	4.0e−03	8.2e+02	1.5e+00	8.1e+02	56	8	5500	18598
copo14	2.65890e−05	−7.14668e−05	1.11364e−05	−2.92301e−05	9.9e−05	4.0e−05	1.3e−06	4.2e−07	50	21	36	603
copo23	−5.27584e−05	−6.03132e−05	2.52874e−02	−4.42279e−02	8.5e−05	6.7e−02	6.8e−06	3.8e−07	63	23	822	22684

TABLE 5
Statistics for Maxcut and theta test problems.

Problem	n	m	Dens E (%)	Dens F (%)
G43	1001	10991	2.39	55.46
G51	1001	6910	1.58	16.12
brock400-1.co	401	20478	25.90	90.92
p-hat300-1.co	301	34218	75.95	98.28
toruspm3-8-50	512	512	1.56	15.60
torusg3-8	512	512	1.56	15.60
toruspm3-15-50	3375	3375	0.24	6.08
torusg3-15	3375	3375	0.24	6.08
hamming-07-5-6	129	1921	24.44	73.71
hamming-08-3-4	257	16385	50.19	93.27
hamming-09-5-6	513	54273	41.55	92.82
hamming-09-8	513	2817	2.53	15.98
hamming-10-2	1025	24065	4.77	38.40
hamming-11-2	2049	58369	2.88	36.57

nearly all instances—it was approximately twice as fast for the first nine and an order of magnitude or more faster for the last six problems.

5.4. Comparison with a first-order method. Finally, we compare Algorithm SDP with the first-order method (BMZ) of Burer, Monteiro, and Zhang [6]. BMZ is a dual-only method that solves a special class of so-called fixed-diagonal SDPs by optimizing the log-barrier function for a decreasing sequence of barrier parameters $\{\nu_k\}_{k\geq 0}$ using a first-order gradient-based nonlinear programming approach. Two of the key features of BMZ are that it works only with S and L and that its function and gradient evaluations each take time $\mathcal{O}(f_2)$, which matches the complexity of the fundamental operations of Algorithm SDP.

We compare Algorithm SDP and BMZ on the fourteen problems shown in Table 5. The first two problems (which come from the Gset test problem suite [14]) and the last six problems (which come from the Seventh DIMACS Challenge) are Lovász theta SDPs and employ the sparse formulation (as mentioned in the previous subsection). The remaining problems are maximum cut SDP relaxations (see [11]) and also come from the Seventh DIMACS Challenge. Each method was given an upper bound of 5 hours running time on each problem and was thus terminated upon completion of the first iteration after 5 hours, if necessary. BMZ was stopped once the log-barrier subproblem corresponding to $\nu = 1.0\text{e}{-4}$ was solved, which yielded a comparable accuracy to Algorithm SDP's stopping criterion. The computational results are shown in Table 6.

The results show that both methods had difficulty solving such large problems in the time allotted. Even still, when comparing objective values, on twelve of the fourteen problems Algorithm SDP made more progress towards optimality than BMZ. An advantage of BMZ, of course, is that each iterate is dual feasible, while an advantage of Algorithm SDP is that primal information is produced in addition to dual information.

6. Concluding remarks. The results of this paper involve both theoretical and practical aspects of solving SDPs. Theoretically, we have shown that it is possible to express the central path using sparse equations rather than the usual dense ones, and we have, moreover, shown how to measure the proximity of a partial primal-dual solution to the central path, which was a question left open in [9]. Combining these ideas, we have also shown how to solve the SDP in polynomial time using a "partial"

TABLE 6
Results of Algorithm SDP (AS) and the Burer–Monteiro–Zhang first-order method (BMZ) on Maxcut and theta test problems.

PROBLEM	OBJECTIVE VALUE			GAP MEAS		R-FEAS		ITER		TIME (s)	
	AS		BMZ	AS (r-gap)	BMZ (ν)	AS	BMZ	AS	BMZ	AS	BMZ
	PRIMAL	DUAL									
G43	−2.82795e+02	−2.81556e+02	−4.83252e+02	5.3e−02	1.0e−02	1.6e−01	0.0e+00	29	737	18221	18029
G51	−3.62057e+02	−3.50632e+02	−3.49226e+02	2.4e−02	1.0e−04	1.6e−01	0.0e+00	30	4783	20317	18003
brock400-1.co	−3.96791e+01	−3.97140e+01	−4.05332e+01	8.6e−04	1.0e−04	3.9e−06	0.0e+00	37	7401	20819	18001
p-hat300-1.co	−9.66567e+00	−1.01670e+01	−1.93734e+01	1.1e−03	1.0e−03	8.7e−05	0.0e+00	37	15455	19039	18003
toruspm3-8-50	−5.27764e+02	−5.27809e+02	−5.27810e+02	8.7e−05	1.0e−05	5.1e−06	0.0e+00	36	7814	88	510
torusg3-8	−4.57317e+07	−4.57359e+07	−4.57360e+07	9.2e−05	1.0e−05	5.0e−06	0.0e+00	36	5692	96	521
toruspm3-15-50	−3.47481e+03	−3.47513e+03	−3.47516e+03	9.2e−05	1.0e−04	5.4e−06	0.0e+00	39	2559	21460	18008
torusg3-15	−3.13381e+08	−3.13457e+08	−3.13458e+08	2.4e−04	1.0e−04	9.7e−06	0.0e+00	38	2829	20996	18005
hamming-07-5-6	−4.26654e+01	−4.26667e+01	−4.26677e+01	1.3e−05	1.0e−05	6.0e−06	0.0e+00	37	7405	38	117
hamming-08-3-4	−2.55381e+01	−2.56004e+01	−2.56700e+01	4.7e−05	1.0e−04	6.2e−06	0.0e+00	38	27447	2794	18001
hamming-09-5-6	−8.47006e+01	−8.54005e+01	−1.11574e+02	3.3e−02	1.0e−02	4.1e−02	0.0e+00	30	4091	18112	18011
hamming-09-8	−2.23997e+02	−2.24000e+02	−2.24003e+02	1.3e−05	1.0e−05	8.4e−06	0.0e+00	39	5674	78	334
hamming-10-2	−1.02183e+02	−1.02402e+02	−1.02427e+02	4.3e−05	1.0e−05	3.1e−06	0.0e+00	40	4431	14463	18002
hamming-11-2	−1.22074e+02	−1.65422e+03	−2.22378e+02	2.1e+00	1.0e−03	1.5e+03	0.0e+00	24	404	19540	18029

Newton direction. Even so, there seem to be many interesting theoretical questions left open by the ideas presented in this paper. For example, can the nonsingularity of H' be established under conditions weaker than those presented in Theorem 2.10? Or can a wider neighborhood of the central path be used to improve the iteration complexity of the method? (The relatively small step-size established in section 4 was forced by the neighborhood, not by the positive semidefiniteness of the new iterates.) Or can other directions with better properties be defined?

Of course, one of the most appealing aspects of applying the idea of matrix completions to SDP is the prospect of actually solving sparse SDPs more efficiently, and the results of section 5 indicate that the algorithm proposed in this paper is highly effective on varying classes of problems—especially for those having a small density pattern F. An area of further investigation for Algorithm SDP is the conditioning of the Schur complement matrix M near optimality, particularly as it affects the convergence of the BiCGSTAB subroutine. Currently, it is unclear how preconditioning techniques can best be employed to mitigate the inevitable ill-conditioning.

Overall, we feel that Algorithm SDP makes a significant contribution to the existing algorithms for SDP, allowing one to solve any SDP in a primal-dual framework while taking advantage of sparsity in all stages of computation.

Acknowledgments. The author would like to thank Stephen Vavasis and Kim Chuan Toh for helpful comments that improved the presentation of section 5. In addition, the author is in debt to two anonymous referees, who provided extremely detailed and insightful comments on the first and second drafts of this paper.

REFERENCES

[1] F. Alizadeh, J.-P. A. Haeberly, and M. L. Overton, *Primal-dual interior-point methods for semidefinite programming: Convergence rates, stability and numerical results*, SIAM J. Optim., 8 (1998), pp. 746–768.

[2] S. J. Benson, Y. Ye, and X. Zhang, *Solving large-scale sparse semidefinite programs for combinatorial optimization*, SIAM J. Optim., 10 (2000), pp. 443–461.

[3] B. Borchers, *SDPLIB* 1.2*, A library of semidefinite programming test problems*, Optim. Methods Softw., 11 (1999), pp. 683–690.

[4] S. Burer and R. D. C. Monteiro, *A nonlinear programming algorithm for solving semidefinite programs via low-rank factorization*, Math. Program., 95 (2003), pp. 329–357.

[5] S. Burer, R. D. C. Monteiro, and Y. Zhang, *A computational study of a gradient-based log-barrier algorithm for a class of large-scale SDPs*, Math. Program., 95 (2003), pp. 359–379.

[6] S. Burer, R. Monteiro, and Y. Zhang, *Solving a class of semidefinite programs via nonlinear programming*, Math. Program., 93 (2002), pp. 97–122.

[7] *Seventh DIMACS Implementation Challenge on Semidefinite and Related Optimization Problems*, 2000, information online at http://dimacs.rutgers.edu/Challenges/Seventh.

[8] M. Fukuda and M. Kojima, *Interior-Point Methods for Lagrangian Duals of Semidefinite Programs*, manuscript, Department of Mathematical and Computing Sciences, Tokyo Institute of Technology, Tokyo, Japan, 2000.

[9] M. Fukuda, M. Kojima, K. Murota, and K. Nakata, *Exploiting sparsity in semidefinite programming via matrix completion* I: *General framework*, SIAM J. Optim., 11 (2001), pp. 647–674.

[10] D. Fulkerson and O. Gross, *Incidence matrices and interval graphs*, Pacific J. Math., 15 (1965), pp. 835–855.

[11] M. Goemans and D. Williamson, *Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming*, J. ACM, 42 (1995), pp. 1115–1145.

[12] R. Grone, C. Johnson, E. Sá, and H. Wolkowicz, *Positive definite completions of partial Hermitian matrices*, Linear Algebra Appl., 58 (1984), pp. 109–124.

[13] C. Helmberg and K. C. Kiwiel, *A spectral bundle method with bounds*, Math. Program., 93 (2002), pp. 173–194.

[14] C. Helmberg and F. Rendl, *A spectral bundle method for semidefinite programming*, SIAM J. Optim., 10 (2000), pp. 673–696.
[15] C. Helmberg, F. Rendl, R. J. Vanderbei, and H. Wolkowicz, *An interior-point method for semidefinite programming*, SIAM J. Optim., 6 (1996), pp. 342–361.
[16] D. S. Johnson and M. A. Trick, eds., *Cliques, Coloring, and Satisfiability: Second DIMACS Implementation Challenge* (1993), American Mathematical Society, Providence, RI, 1996.
[17] M. Kojima, S. Shindoh, and S. Hara, *Interior-point methods for the monotone semidefinite linear complementarity problem in symmetric matrices*, SIAM J. Optim., 7 (1997), pp. 86–125.
[18] M. Kočvara and M. Stingl, *Pennon: A code for convex nonlinear and semidefinite programming*, Optim. Methods Softw., to appear.
[19] M. Laurent, S. Poljak, and F. Rendl, *Connections between semidefinite relaxations of the max-cut and stable set problems*, Math. Program., 77 (1997), pp. 225–246.
[20] C.-J. Lin and R. Saigal, *On Solving Large Scale Semidefinite Programming Problems: A Case Study of Quadratic Assignment Problem*, Technical report, Department of Industrial and Operations Engineering, The University of Michigan, Ann Arbor, MI, 1997.
[21] L. Lovász, *On the Shannon capacity of a graph*, IEEE Trans. Inform. Theory, IT-25 (1979), pp. 1–7.
[22] R. D. C. Monteiro, *Primal–dual path-following algorithms for semidefinite programming*, SIAM J. Optim., 7 (1997), pp. 663–678.
[23] R. D. C. Monteiro, *Polynomial convergence of primal-dual algorithms for semidefinite programming based on the Monteiro and Zhang family of directions*, SIAM J. Optim., 8 (1998), pp. 797–812.
[24] R. D. C. Monteiro and T. Tsuchiya, *Polynomial convergence of a new family of primal-dual algorithms for semidefinite programming*, SIAM J. Optim., 9 (1999), pp. 551–577.
[25] R. Monteiro and Y. Zhang, *A unified analysis for a class of path-following primal-dual interior-point algorithms for semidefinite programming*, Math. Program., 81 (1998), pp. 281–299.
[26] K. Nakata, K. Fujisawa, M. Fukuda, M. Kojima, and K. Murota, *Exploiting sparsity in semidefinite programming via matrix completion* II: *Implementation and numerical results*, Math. Program., 95 (2003), pp. 303–327.
[27] K. Nakata, K. Fujisawa, and M. Kojima, *Using the conjugate gradient method in interior-points for semidefinite programs*, Proc. Inst. Statist. Math., 46 (1998), pp. 297–316 (in Japanese).
[28] Y. E. Nesterov and M. J. Todd, *Primal-dual interior-point methods for self-scaled cones*, SIAM J. Optim., 8 (1998), pp. 324–364.
[29] K. Toh and M. Kojima, *Solving some large scale semidefinite programs via the conjugate residual method*, SIAM J. Optim., 12 (2002), pp. 669–691.
[30] L. Tuncel, *Potential reduction and primal-dual methods*, in Handbook of Semidefinite Programming, H. Wolkowicz, R. Saigal, and L. Vandenberghe, eds., Kluwer Academic Publishers, 2000, Norwell, MA, pp. 235–265.
[31] R. H. Tütüncü, K. C. Toh, and M. J. Todd, *SDPT*3: *A Matlab Software Package for Semidefinite-Quadratic-Linear Programming, Version* 3.0, 2001; available online from http://www.math.nus.edu.sg/~mattohkc/sdpt3.html.
[32] Y. Zhang, *On extending some primal-dual interior-point algorithms from linear programming to semidefinite programming*, SIAM J. Optim., 8 (1998), pp. 365–386.

SIAM J. OPTIM.
Vol. 14, No. 1, pp. 173–199

A PRIMAL-DUAL INTERIOR-POINT METHOD FOR NONLINEAR PROGRAMMING WITH STRONG GLOBAL AND LOCAL CONVERGENCE PROPERTIES*

ANDRÉ L. TITS†, ANDREAS WÄCHTER‡, SASAN BAKHTIARI†, THOMAS J. URBAN§, AND CRAIG T. LAWRENCE¶

Abstract. An exact-penalty-function-based scheme—inspired from an old idea due to Mayne and Polak [*Math. Program.*, 11 (1976), pp. 67–80]—is proposed for extending to general smooth constrained optimization problems any given feasible interior-point method for inequality constrained problems. It is shown that the primal-dual interior-point framework allows for a simpler penalty parameter update rule than the one discussed and analyzed by the originators of the scheme in the context of first order methods of feasible direction. Strong global and local convergence results are proved under mild assumptions. In particular, (i) the proposed algorithm does not suffer a common pitfall recently pointed out by Wächter and Biegler [*Math. Program.*, 88 (2000), pp. 565–574]; and (ii) the positive definiteness assumption on the Hessian estimate, made in the original version of the algorithm, is relaxed, allowing for the use of exact Hessian information, resulting in local quadratic convergence. Promising numerical results are reported.

Key words. constrained optimization, nonlinear programming, primal-dual interior-point methods, feasibility

AMS subject classifications. 49M37, 65K05, 65K10, 90C30, 90C53

DOI. 10.1137/S1052623401392123

1. Introduction.

Consider the problem

$$\text{(P)} \qquad \begin{array}{ll} \min\limits_{x\in\mathcal{R}^n} & f(x) \\ \text{s.t.} & c_j(x) = 0, \quad j = 1,\dots,m_{\mathrm{e}}, \\ & d_j(x) \geq 0, \quad j = 1,\dots,m_{\mathrm{i}}, \end{array}$$

where $f : \mathcal{R}^n \to \mathcal{R}$, $c_j : \mathcal{R}^n \to \mathcal{R}$, $j = 1,\dots,m_{\mathrm{e}}$, and $d_j : \mathcal{R}^n \to \mathcal{R}$, $j = 1,\dots,m_{\mathrm{i}}$, are smooth. No convexity assumptions are made. A number of primal-dual interior-point methods have been proposed to tackle such problems; see, e.g., [26, 27, 5, 8, 7, 3, 4, 22]. While all of these methods make use of a search direction generated by a Newton or quasi-Newton iteration on a perturbed version of some first order necessary conditions of optimality, they differ in many respects. For example, some algorithms enforce feasibility of all iterates with respect to inequality constraints [8, 7], while others, sometimes referred to as "infeasible," sidestep that requirement via the introduction of slack variables [26, 27, 5, 3, 4, 22]. As for equality constraints, some schemes include them "as is" in the perturbed optimality conditions [26, 27, 5, 8, 3, 4], while

*Received by the editors July 10, 2001; accepted for publication (in revised form) March 4, 2003; published electronically July 18, 2003. This work was supported in part by the National Science Foundation under grant DMI-9813057.

http://www.siam.org/journals/siopt/14-1/39212.html

†Department of Electrical and Computer Engineering and Institute for Systems Research, University of Maryland, College Park, MD 20742 (andre@umd.edu, sasanb@umd.edu).

‡IBM T.J. Watson Research Center, P.O. Box 218, Yorktown Heights, NY 10598 (andreasw@watson.ibm.com).

§Applied Physics Laboratory, The Johns Hopkins University, 11100 Johns Hopkins Road, Laurel, MD 20723-6099 (thomas.urban@jhuapl.edu).

¶Alphatech, 3811 North Fairfax Drive, Arlington, VA 22203 (craigl@dc.alphatech.com).

some soften this condition by making use of two sets of slack variables [22] or by introducing a quadratic penalty function, yielding optimality conditions involving a perturbed version of "$c(x) = 0$" [7]. Also, some proposed algorithms (e.g., [27, 3, 4]) involve a trust region mechanism. In many cases (e.g., [27, 8, 22]), promising numerical results have been obtained. In some cases (e.g., [26, 27, 5, 3]), convergence properties have been proved under certain assumptions. Often, however, it is not proved that the line search eventually accepts a step size close enough to one to allow fast local convergence; i.e., a Maratos-like effect [14] is not ruled out. An exception is [27], but rather strong assumptions are used there.

Recently, Wächter and Biegler [23] showed that many of the proposed algorithms suffer a major drawback in that for problems with two or more equality constraints and a total number of constraints in excess of the dimension of the space, the constructed primal sequence can converge to spurious, infeasible points. They produced a simple, seemingly innocuous example where such behavior is observed when starting from rather arbitrary initial points. They pointed out that where global convergence had been proved, it was under a linear independence assumption that often fails to hold in the case of problems with such a number of constraints. One exception to this is [3], where the proposed trust-region-based method is proved to converge globally under fairly mild assumptions; another is the recent paper [24].

In this paper, we propose a line-search-based primal-dual interior-point algorithm of the "feasible" variety for which global and fast local convergence are proved to hold under rather mild assumptions. In particular, it involves a scheme to circumvent Maratos-like effects and is immune to the phenomenon observed in [23]. A distinguishing feature of the proposed algorithm is that it makes use of both a barrier parameter and an "exterior" penalty parameter, both of which are adaptively adjusted to ensure global and fast local convergence. The algorithm originates in two papers dating back more than one and two decades, respectively: [17] and [15]. The former proposed a feasible interior-point method for inequality constrained problems, proven to converge globally and locally superlinearly under appropriate assumptions. The latter offered a scheme for dealing with equality constraints in the context of a (largely arbitrary) algorithm for inequality constraint optimization.

In the 1980s, a feasible-iterate algorithm for solving (P) was proposed for the case without equality constraints, based on the following idea. First, given strictly feasible estimates $\hat{x}$ of a solution and $\hat{z}$ of the corresponding Karush–Kuhn–Tucker (KKT) multiplier vector, compute the Newton (or a quasi-Newton) direction for the solution of the equations in the KKT first order necessary conditions of optimality. Then solve again the same system of equations but with the right-hand side appropriately perturbed so as to tilt the primal direction away from the constraint boundaries into the feasible set. The amount of perturbation is determined from the solution of the unperturbed system. Both the original and tilted primal directions are directions of descent for f. Decrease of f is then enforced by the line search to ensure global convergence. Maratos-like effects are avoided by means of a second order correction (adapted from an idea of Mayne and Polak [16]), allowing for fast local convergence to take place. These ideas were put forth in [17]. The central idea in the algorithm of [17] originated in earlier work by Segenreich, Zouain, and Herskovits [20] and Herskovits [10, 11]; see [21] for a detailed historical account. Ideas were also borrowed from [6] and [18].

In the mid-1970s Mayne and Polak proposed an ingenious scheme to incorporate equality constraints in methods of feasible directions [15]. The idea is to (1) relax

each equality constraint ($c_j(x) = 0$) by replacing it with an inequality constraint ($c_j(x) \geq 0$) and (2) penalize departure from the constraint boundaries associated with these relaxed constraints by adding a simple penalty term ($\rho \sum c_j(x)$, $\rho > 0$) to the cost function. For fixed values of the penalty parameter ρ, the feasible direction method under consideration is used. It is readily shown that, locally, convergence to KKT points of the original problem takes place, provided the penalty parameter is increased to a value larger than the magnitude of the most negative equality constraint multiplier (for the original problem) at the solution. Accordingly, in [15] the penalty parameter is adaptively increased based on estimates of these multipliers. While [15] is concerned with classical first order feasible directions methods, it is pointed out in the introduction of that paper that the proposed scheme can convert "*any* (emphasis from [15]) interior point algorithm for inequality constrained optimization problems into an algorithm for optimization subject to combined equality and inequality constraints."

A careful examination of the proposed algorithm, however, reveals two shortcomings. The first one concerns the computation of multiplier estimates. In [15], this is done by solving a linear least squares problem for all equality constraint multipliers and all multipliers associated with ϵ-active inequality constraints (that is, with inequality constraints whose current value is less than some fixed, prescribed ϵ—denoted ϵ' in [15]). The price to pay is that if ϵ is "large," then (1) the computational overhead may become significant and (2) the set of active constraints may be overestimated, leading to incorrect multiplier estimates. On the other hand, if ϵ is selected to be very small, the set of active constraints will be underestimated, again yielding incorrect multipler estimates. The second shortcoming is that global convergence is proved under the strong assumption that at every point in the extended feasible set (where one-side violation of equality constraints is allowed) the gradients of *all* equality constraints and of the active inequality constraints are linearly independent. Indeed, as pointed out in [23], such an assumption does not hold in the example discussed there, and it is typically violated on entire manifolds in problems with two or more equality constraints and a total number of constraints in excess of n.[1] In [11] it is suggested that the idea introduced in [15] could be readily applied to the interior-point algorithm proposed there, but no details are given. The Mayne–Polak idea was used in [13] in the context of feasible SQP. The ready availability of multiplier estimates (for the penalized problem) in that context allows an improved multiplier estimation scheme (for the original problem), thus improving on the first shortcoming just pointed out; however, no attempt is made in [13] to dispense with the strong linear independence assumption.

In the 1980s and 1990s, other penalty parameter update rules were proposed for ℓ_1 (as in [15]) or ℓ_∞ exact penalty functions in the context of SQP and trust region methods, among others. (See, e.g., [16, 19, 2, 28].) In most cases, just like in [15] and [13], the updating rule involves multiplier estimates whose computation requires the solution of a linear system of equations or even that of a linear program. An exception is [28], where the following simple rule is used: at iteration k, increase ρ if the constraint is far from being satisfied, specifically if $\|c(x_k)\| > \upsilon_k$, where υ_k appropriately decreases to zero as k goes to infinity. This rule is proposed in the context of a trust region method, and υ_k involves the model decrease. A challenge when extending it to other contexts is that if υ_k is chosen too small, ρ will increase unnecessarily, perhaps without bound.

[1] See the discussion following the statement of Assumption 3 in section 3 below.

The contributions of the present paper are as follows. First it is shown that all the convergence results proved in [17] for the algorithm proposed in that paper still hold if the positive definiteness assumption on the Hessian estimate is relaxed and replaced with a significantly milder assumption. In particular, the new assumption allows for use of the exact Hessian. Subject to a minor modification of the algorithm, local quadratic convergence in the primal-dual space is proved when the exact Hessian is indeed used. Second, the algorithm is extended to general constrained problems by incorporating a modified Mayne–Polak scheme. Specifically, a new, simple penalty parameter update rule is introduced involving no additional computation. Such a rule is made possible by the availability of multiplier estimates for the penalized problem through the primal-dual iteration. The resulting algorithm converges globally and locally superlinearly without the requirement that a strong regularity assumption be satisfied, thus avoiding the pitfall observed in [23].

The balance of the paper is organized as follows. In section 2, the algorithm from [17] is described in "modern" terms from a barrier function perspective. It is shown how certain assumptions made in [17] can be relaxed, and quadratic convergence is shown for the case when the "exact Hessian" is used. The overall algorithm is then motivated and described in section 3. In section 4, global and local superlinear convergence are proved. Preliminary numerical results are reported in section 5, starting with results on the example discussed in [23]. Finally, section 6 is devoted to concluding remarks. Throughout, $\|\cdot\|$ denotes the Euclidean norm or corresponding operator norm and, given two vectors v_1 and v_2, inequalities such as $v_1 \leq v_2$ and $v_1 < v_2$ are to be understood componentwise. Much of our notation is borrowed from [8].

2. Problems without equality constraints. We briefly review the algorithm of [17] in the primal-dual interior-point formalism and then point out how the assumptions made in [17] can be relaxed without affecting the convergence theorems.[2]

2.1. Brief review of [17]. Consider problem (P) with $m_{\mathrm{e}} = 0$, i.e.,

$$\begin{array}{ll} \min\limits_{x\in\mathcal{R}^n} & f(x) \\ \text{s.t.} & d_j(x) \geq 0, \quad j = 1, \ldots, m_{\mathrm{i}} \, . \end{array} \tag{2.1}$$

The algorithm proposed in [17] for problems such as (2.1) can equivalently be stated based on the logarithmic barrier function

$$\beta(x,\mu) = f(x) - \sum_{j=1}^{m_{\mathrm{i}}} \mu^{(j)} \log d_j(x), \tag{2.2}$$

where $\mu = [\mu^{(1)}, \ldots, \mu^{(m_{\mathrm{i}})}]^{\mathrm{T}} \in \mathcal{R}^{m_{\mathrm{i}}}$ and the μ_js are positive. The barrier gradient is given by

$$\nabla_x \beta(x,\mu) = g(x) - B(x)^{\mathrm{T}} D(x)^{-1} \mu, \tag{2.3}$$

where g denotes the gradient of f, B the Jacobian of d, and $D(x)$ the diagonal matrix $\mathrm{diag}(d_j(x))$.

Problem (2.1) can be tackled via a sequence of unconstrained minimizations of $\beta(x,\mu)$ with $\mu \to 0$. In view of (2.3), $z = D(x)^{-1}\mu$ can be viewed as an approximation

[2] While the present paper was under review, this algorithm was further enhanced; see [1].

to the KKT multiplier vector associated with a solution of (2.1) and the right-hand side of (2.3) as the value at (x, z) of the gradient (with respect to x) of the Lagrangian

$$\mathcal{L}(x, z) = f(x) - \langle z, d(x)\rangle.$$

Accordingly, and in the spirit of primal-dual interior-point methods, consider using a (quasi-)Newton iteration for the solution of the system of equations in (x, z)

$$g(x) - B(x)^{\mathrm{T}} z = 0, \tag{2.4}$$

$$D(x)z = \mu, \tag{2.5}$$

i.e.,

$$\begin{bmatrix} -W & B(x)^{\mathrm{T}} \\ ZB(x) & D(x) \end{bmatrix} \begin{bmatrix} \Delta x \\ \Delta z \end{bmatrix} = \begin{bmatrix} g(x) - B(x)^{\mathrm{T}} z \\ \mu - D(x)z \end{bmatrix}, \tag{2.6}$$

where $Z = \mathrm{diag}(z^{(j)})$ and where W is equal to, or approximates, the Hessian (with respect to x) of the Lagrangian $\mathcal{L}(x, z)$. When $\mu = 0$, a primal-dual feasible solution to (2.4)–(2.5) is a KKT point for (2.1). Moreover, under the assumption made in [17] that W is positive definite[3] and given any strictly feasible primal-dual pair (x, z), the primal direction Δx^0 obtained by solving (2.6) with $\mu = 0$ is a descent direction for f at x. In [17], such a property is sought for the search direction and used in the line search. On the other hand, while any primal direction is "feasible" when starting from an interior point, Δx^0 is not necessarily a direction of ascent for "almost active" constraints, whereas when the components of μ are chosen to be strictly positive, such desirable ascent property is guaranteed, but descent for f may be lost. Thus, the components of μ should

- be positive enough to prevent the primal step length from collapsing due to infeasibility,
- be small enough that significant descent for f is maintained, and
- go to zero fast enough to preserve the fast local convergence properties associated with the (quasi-)Newton iteration for (2.4)–(2.5) with $\mu = 0$.

This is achieved in [17] by selecting

$$\mu = \varphi \|\Delta x^0\|^{\nu} z, \tag{2.7}$$

with $\varphi \in (0, 1]$ as large as possible subject to the constraint

$$\langle g(x), \Delta x\rangle \leq \theta \langle g(x), \Delta x^0\rangle, \tag{2.8}$$

where $\nu > 2$ and $\theta \in (0, 1)$ are prespecified;[4] condition (2.8) ensures that Δx is still a descent direction for f.

In [17] primal and dual strict feasibility is enforced at each iteration. An arc search is performed to select a next primal iterate x^+. The search criterion includes decrease of f and strict primal feasibility. It involves a second order correction $\Delta\tilde{x}$ to allow a full Newton (or quasi-Newton) step to be taken near the solution. With index sets I and J defined by

$$I = \{j : d_j(x) \leq z^{(j)} + \Delta z^{(j)}\},$$

[3]Below (section 2.2) we show that this assumption can be relaxed.

[4]Note that Δx depends on φ affinely and thus Δx is computed at no extra cost once (2.6) has been solved with, say, $\mu = \|\Delta x^0\|^{\nu} z$.

$$J = \{j : z^{(j)} + \Delta z^{(j)} \leq -d_j(x)\},$$

$\Delta\tilde{x}$ is the solution of the linear least squares problem

$$\min \frac{1}{2}\langle \Delta\tilde{x}, W\Delta\tilde{x}\rangle \text{ s.t. } d_j(x+\Delta x) + \langle \nabla d_j(x), \Delta\tilde{x}\rangle = \psi \quad \forall j \in I, \tag{2.9}$$

where

$$\psi = \max\left\{ \|\Delta x\|^{\tau}, \max_{j\in I}\left|\frac{\Delta z^{(j)}}{z^{(j)}+\Delta z^{(j)}}\right|^{\kappa} \|\Delta x\|^2 \right\}, \tag{2.10}$$

with $\tau \in (2,3)$ and $\kappa \in (0,1)$ prespecified. If $J \neq \emptyset$ or (2.9) is infeasible or unbounded or $\|\Delta\tilde{x}\| > \|\Delta x\|$, $\Delta\tilde{x}$ is set to 0. The rationale for the first of these three conditions is that computing the Maratos correction involves some cost, and it is known to be of help only close to a solution: when $J \neq \emptyset$, the correction is not computed. Note that I estimates the active index set and that J (multipliers of "wrong" sign) should be empty near the solution when strict complementarity holds. An (Armijo-type) arc search is then performed as follows: given $\eta \in (0,1)$, compute the first number α in the sequence $\{1, \eta, \eta^2, \dots\}$ such that

$$f\left(x + \alpha\Delta x + \alpha^2\Delta\tilde{x}\right) \leq f(x) + \xi\alpha\langle g(x), \Delta x\rangle, \tag{2.11}$$

$$d_j\left(x + \alpha\Delta x + \alpha^2\Delta\tilde{x}\right) > 0 \quad \forall j, \tag{2.12}$$

$$d_j\left(x + \alpha\Delta x + \alpha^2\Delta\tilde{x}\right) \geq d_j(x) \quad \forall j \in J, \tag{2.13}$$

where $\xi \in (0, 1/2)$ is prespecified. The third inequality is introduced to prevent convergence to points with negative multipliers. The next primal iterate is then set to

$$x^+ = x + \alpha\Delta x + \alpha^2\Delta\tilde{x}.$$

Finally, the dual variable z is reinitialized whenever $J \neq \emptyset$; if $J = \emptyset$, the new value $z^{+,(j)}$ of $z^{(j)}$ is set to

$$z^{+,(j)} = \min\{\max\{\|\Delta x\|, z^{(j)} + \Delta z^{(j)}\}, z_{\max}\}, \tag{2.14}$$

where $z_{\max} > 0$ is prespecified. Thus $z^{+,(j)}$ is allowed to be close to 0 only if $\|\Delta x\|$ is small, indicating proximity to a solution.

It is observed in [17, section 5] that the total work per iteration (in addition to function evaluations) is essentially one Cholesky decomposition of size m_{i} and one Cholesky decomposition of size equal to the number of active constraints at the solution.[5]

On the issue of global convergence, it is shown in [17] that given an initial strictly feasible primal-dual pair (x_0, z_0) and given a sequence of symmetric matrices $\{W_k\}$, uniformly bounded and uniformly positive definite, the primal sequence $\{x_k\}$ constructed by the algorithm just described (with W_k used as W at the kth iteration) converges to KKT points for (2.1), provided the following assumptions hold: (i) $\{x : f(x) \leq f(x_0), d(x) \geq 0\}$ is bounded so that the primal sequence remains

[5] There are two misprints in [17, section 5]: in equation (5.3) (statement of Proposition 5.1) as well as in the last displayed equation in the proof of Proposition 5.1 (expression for λ_k^0), $M_k B_k^{-1}$ should be $B_k^{-1} M_k$.

bounded, (ii) for all feasible x the vectors $\nabla d_j(x)$, $j \in \{j : d_j(x) = 0\}$, are linearly independent, and (iii) the set of feasible points x for which (2.4)–(2.5) hold for some z (with no restriction on the sign of the components of z)[6] is finite.

Superlinear convergence of the primal sequence—in particular, eventual acceptance of the full step of one by the arc search—is also proved in [17] under appropriate second order assumptions, provided that none of the KKT multipliers at the solution are larger than $z_{\max}$ and that, asymptotically, W_k suitably approximates the Hessian of the Lagrangian at the solution on the tangent plane to the active constraints.

Finally, stronger convergence results hold for a variation of the present algorithm, under weaker assumptions, in the LP and convex QP cases. In particular, global convergence to the solution set X^* takes place whenever X^* is nonempty and bounded, the feasible set X has a nonempty interior, and for every $x \in X$ the gradients of the active constraints at x are linearly independent. See [21] for details.

2.2. Global convergence under milder assumptions. Two assumptions made in [17] can be relaxed without affecting the convergence results proved there. First, Assumption A4 (x_0 is the initial point):

> The set $X \cap \{x \text{ s.t. } f(x) \leq f(x_0)\}$ is compact

can be eliminated altogether. Indeed, this assumption is invoked only in the proof of Lemmas 3.8 and 3.9 of [17]. The former is readily proved without such assumption: convergence of $\{x_{k-1}\}$ on K directly follows from the assumed convergence on K of $\{x_k\}$ and $\{d_{k-1}\}$ (in the notation of [17]) and from the last displayed equation in the proof. As for the latter, a weaker statement by which K is selected under the additional restriction that $\{x_k\}$ converges on K is sufficient for the use made of that lemma in Proposition 3.10 and Theorem 3.11.

Second and more significantly, Assumption A6 of [17] (in the notation of this paper):

> There exist $\sigma_1, \sigma_2 > 0$ such that $\sigma_1\|v\|^2 \leq \langle v, W_k v\rangle \leq \sigma_2\|v\|^2$, for all k, for all $v \in \mathcal{R}^n$

can be replaced with the following milder assumption.

Assumption PTH-A6*. *Given any index set K such that $\{x_k\}_{k\in K}$ is bounded, there exist $\sigma_1, \sigma_2 > 0$ such that, for all $k \in K$,*

$$\|W_k\| \leq \sigma_2$$

and

$$\left\langle v, \left(W_k + \sum_{i=1}^{m_{\mathrm{i}}} \frac{z_k^{(i)}}{d_i(x_k)} \nabla d_i(x_k) \nabla d_i(x_k)^{\mathrm{T}} \right) v \right\rangle \geq \sigma_1 \|v\|^2 \quad \forall v \in R^n.$$

(Here $\{x_k\}$, $\{z_k\}$, and $\{W_k\}$ are the sequences of values of x, z, and W generated by the algorithm outlined above. The restriction of this assumption to bounded subsequences of $\{x_k\}$ is made in connection with our dropping Assumption A4.)

The difference with Assumption A6 of [17] is significant because, as is readily verified, the (exact) Hessian of the Lagrangian satisfies the relaxed assumption in the neighborhood of any solution of (2.1) at which strong second order sufficiency conditions of optimality hold. It is shown in the appendix that all the results proved in [17] still hold under the new assumption. In particular, it is proven that the direction Δx^0 (in the notation of this paper) is still well defined and is a direction

[6] Such points are referred to in [17] as *stationary points*.

of descent for f. It should be noted that when W is not positive definite, there are two ways (rather than one) in which (2.9) can fail to have a solution: when its feasible set is nonempty, its cost function could possibly be unbounded from below. As observed in the appendix, the analysis of [17] still implies that locally around a "strong" minimizer, (2.9) still has a solution.

2.3. Local quadratic convergence. As noted at the end of subsection 2.1, superlinear convergence of $\{x_k\}$ is proved in [17] under appropriate local assumptions. Here we show that under the further assumption that, eventually, W_k is the Hessian evaluated at (x_k, z_k) of the Lagrangian associated with problem (2.1), the pair (x_k, z_k) converges Q-quadratically, provided the following minor modification is made to the algorithm of [17]: replace (2.14) with

$$z^{+,(j)} = \min\{\max\{\|\Delta x\|^2, z^{(j)} + \Delta z^{(j)}\}, z_{\max}\},$$

i.e., allow z_k to go to zero like $\|\Delta x_k\|^2$ rather than merely $\|\Delta x_k\|$. It can be checked that this modification does not affect the analysis carried out in [17].

The proof is based on Proposition 3.10 of [21], which we restate here for ease of reference. (Related result are obtained in [5] and [25].)

LEMMA 2.1. *Let $F : \mathcal{R}^\ell \to \mathcal{R}^\ell$ be twice continuously differentiable, and let $w^* \in \mathcal{R}^\ell$ and $r > 0$ be such that $F(w^*) = 0$ and $\frac{\partial F}{\partial w}(w)$ is nonsingular whenever $w \in B(w^*, r) := \{w : \|w^* - w\| \le r\}$. Let $v^N : B(w^*, r) \to \mathcal{R}^\ell$ be defined by $v^N(w) = -\left(\frac{\partial F}{\partial w}(w)\right)^{-1} F(w)$. Then given any $\Gamma_1 > 0$ there exists $\Gamma_2 > 0$ such that*

$$\|w^+ - w^*\| \le \Gamma_2 \|w - w^*\|^2 \tag{2.15}$$

for every $w \in B(w^*, r)$ and $w^+ \in \mathcal{R}^\ell$ for which, for each $i \in \{1, \dots, \ell\}$, either

(i) $|w^{+,(i)} - w^{*,(i)}| \le \Gamma_1 \|v^N(w)\|^2$

or

(ii) $|w^{+,(i)} - (w^{(i)} + v_i^N(w))| \le \Gamma_1 \|v^N(w)\|^2$.

Let $w := [x^{\mathrm{T}}, z^{\mathrm{T}}]^{\mathrm{T}}$, $w_k := [x_k^{\mathrm{T}}, z_k^{\mathrm{T}}]^{\mathrm{T}}$, etc., let

$$\Phi(w, \mu) = \begin{bmatrix} -(g(x) - B(x)^{\mathrm{T}} z) \\ D(x)z - \mu \end{bmatrix}, \tag{2.16}$$

and let $M(w)$ denote the matrix in the left-hand side of (2.6) with W the "exact Hessian," i.e.,

$$W = \nabla^2_{xx} f(x) - \sum_{i=1}^{m} z^{(i)} \nabla^2_{xx} d_i(x).$$

Thus $M(w)$ is the Jacobian of $\Phi(w, \mu)$ with respect to w. (Note that $M(w)$ does not depend on μ.) We will invoke Lemma 2.1 with $F := \Phi(\cdot, 0)$. Observe that

$$\Delta w_k^0 = -M(w_k)^{-1} \Phi(w_k, 0)$$

and

$$\Delta w_k = -M(w_k)^{-1} \Phi(w_k, \mu_k),$$

and, since $\mu_k = O(\|\Delta x_k^0\|^\nu)$ and $M(w_k)^{-1}$ is bounded (in view of Lemma PTH-3.5* in the appendix), that

$$\Delta w_k - \Delta w_k^0 = O(\|\Delta x_k^0\|^\nu). \tag{2.17}$$

Next, we observe that with a simple additional observation, the proof of Lemma 4.4 in [17] establishes that

$$\|\Delta\tilde{x}\| = O(\|\Delta w\|^2). \tag{2.18}$$

Indeed, in connection with the last displayed equation in that proof, since under our strict complementarity assumption $z_k^{(i)}+\Delta z_k^{(i)}$ ($\lambda_{k,i}$ in the notation of [17]) is bounded away from zero for large k, we can write

$$\frac{z_k^{(i)}}{z_k^{(i)}+\Delta z_k^{(i)}} - 1 = O(|\Delta z_k^{(i)}|) = O(\|\Delta w_k\|),$$

and the claim follows.

Now, proceed analogously to the proof of Theorem 3.11 in [21]. Thus, with reference to Lemma 2.1, let $r > 0$ be such that $M(w)$ is nonsingular for all $w \in B(w^*, r)$. (In view of Lemma PTH-3.5* in the appendix, such an r exists.) Since $\{w_k\} \to w^*$ as $k \to \infty$, there exists k_0 such that $w_k \in B(w^*, r)$ for all $k \geq k_0$. Now let us first consider $\{z_k\}$. For $i \in I(x^*)$, in view of strict complementarity, $z_{k+1}^{(i)} = z_k^{(i)} + \Delta z_k^{(i)}$ for k large enough so that, in view of (2.17), condition (ii) in Lemma 2.1 holds for k large enough. Next, for $i \notin I(x^*)$, for each k either again $z_{k+1}^{(i)} = z_k^{(i)}+\Delta z_k^{(i)}$ or (in view of our modified updating formula for z_k) $z_{k+1}^{(i)} = \|\Delta x_k\|^2$. In the latter case, since $z^{*,(i)} = 0$, noting again (2.17), we conclude that condition (i) in Lemma 2.1 holds. Next, consider $\{x^k\}$. Since $\alpha_k = 1$ for k large enough, we have

$$\|x_{k+1} - (x_k + \Delta x_k^0)\| = \|\Delta x_k - \Delta x_k^0 + \Delta\tilde{x}_k\|,$$

which in view of (2.17) and (2.18) implies that condition (ii) again holds. Thus the conditions of Lemma 2.1 hold, and Q-quadratic convergence follows.

3. Overall algorithm. Suppose now that m_{e} is not necessarily zero. Let X denote the feasible set for (P); i.e., let

$$X := \{x \in \mathcal{R}^n : c_j(x) = 0,\ j = 1, \ldots, m_{\mathrm{e}},\ d_j(x) \geq 0,\ j = 1, \ldots, m_{\mathrm{i}}\}. \tag{3.1}$$

Further, let A denote the Jacobian of c, let $C(x) = \mathrm{diag}(c_j(x))$ and, just as above, let B denote the Jacobian of d and let $D(x) = \mathrm{diag}(d_j(x))$.

In [15], Mayne and Polak proposed a scheme to convert (P) to a sequence of inequality constrained optimization problems of the type

$$(\mathrm{P}_\rho) \qquad \begin{array}{lll} \min\limits_{x\in\mathcal{R}^n} & f_\rho(x) & \\ \text{s.t.} & c_j(x) \geq 0, & j = 1, \ldots, m_{\mathrm{e}}, \\ & d_j(x) \geq 0, & j = 1, \ldots, m_{\mathrm{i}}, \end{array}$$

where $f_\rho(x) = f(x)+\rho\sum_{j=1}^{m_{\mathrm{e}}} c_j(x)$, and where $\rho > 0$. Examination of (P_ρ) shows that large values of ρ penalize iterates satisfying $c_j(x) > 0$ for any j, while feasibility for the modified problem ensures that $c_j(x) \geq 0$. Thus, intuitively, for large values of ρ, iterates generated by a feasible-iterate algorithm will tend towards feasibility for the original problem (P). In fact, the penalty function is "exact" in that convergence to a solution of (P) is achieved without the need to drive ρ to infinity. In other words, under mild assumptions, for large enough but finite values of ρ, solutions to (P_ρ) are solutions to (P).

Let $\tilde{X}$ and $\tilde{X}_0$ be the feasible and strictly feasible sets for problems (P_ρ); i.e., let

$$\tilde{X} := \{x \in \mathcal{R}^n : c_j(x) \geq 0,\ j = 1, \ldots, m_{\mathrm{e}}, \quad d_j(x) \geq 0,\ j = 1, \ldots, m_{\mathrm{i}}\}, \tag{3.2}$$

$$\tilde{X}_0 := \{x \in \mathcal{R}^n : c_j(x) > 0,\ j = 1, \ldots, m_{\mathrm{e}}, \quad d_j(x) > 0,\ j = 1, \ldots, m_{\mathrm{i}}\}. \tag{3.3}$$

Also, for $x \in \tilde{X}$, let $I^{\mathrm{e}}(x)$ and $I^{\mathrm{i}}(x)$ be the active index sets corresponding to c and d, i.e.,

$$I^{\mathrm{e}}(x) = \{j : c_j(x) = 0\}; \quad I^{\mathrm{i}}(x) = \{j : d_j(x) = 0\}.$$

Before proceeding, we state some basic assumptions.

ASSUMPTION 1. *X is nonempty.*

ASSUMPTION 2. *f, c_i, $i = 1 \ldots, m_{\mathrm{e}}$, and d_i, $i = 1, \ldots, m_{\mathrm{i}}$, are continuously differentiable.*

ASSUMPTION 3. *For all $x \in \tilde{X}$,* (i) *the set $\{\nabla c_j(x) : j \in I^{\mathrm{e}}(x)\} \cup \{\nabla d_j(x) : j \in I^{\mathrm{i}}(x)\}$ is linearly independent;* (ii) *if $x \notin X$, then no scalars $y^{(j)} \geq 0$, $j \in I^{\mathrm{e}}(x)$, and $z^{(j)} \geq 0$, $j \in I^{\mathrm{i}}(x)$, exist such that*

$$\sum_{j=1}^{m_{\mathrm{e}}} \nabla c_j(x) = \sum_{j \in I^{\mathrm{e}}(x)} y^{(j)} \nabla c_j(x) + \sum_{j \in I^{\mathrm{i}}(x)} z^{(j)} \nabla d_j(x). \tag{3.4}$$

Note that Assumption 1 implies that $\tilde{X}$ is nonempty and, together with Assumptions 2 and 3(i), that $\tilde{X}_0$ is nonempty, $\tilde{X}$ being its closure.[7]

Our regularity assumption, Assumption 3, is considerably milder than linear independence of the gradients of all c_i's and all active d_i's. As observed in [23], the latter assumption is undesirable, in that whenever there are two or more equality constraints and the total number of constraints exceeds n, it is typically violated over entire submanifolds of $\tilde{X} \setminus X$. On the other hand, as stated in the next lemma, Assumption 3(ii) is equivalent to the mere existence at every $x \in \tilde{X} \setminus X$ of a feasible (with respect to $\tilde{X}$) direction of strict descent for the ℓ_1 norm of $c(x)$. (Indeed Assumption 3(ii) simply states that the sum in the left-hand side of (3.4) does not belong to the closed convex cone generated by the listed constraint gradients, and existence of such strict descent direction amounts to strict separation of that sum from this cone.)

LEMMA 3.1. *Suppose Assumptions 2 and 3(i) hold. Then Assumption 3(ii) is equivalent to the following statement (S): for every $x \in \tilde{X} \setminus X$, there exists $v \in \mathcal{R}^n$ such that*

$$\left\langle \sum_{j=1}^{m_{\mathrm{e}}} \nabla c_j(x), v \right\rangle < 0,$$

$$\langle \nabla c_j(x), v \rangle > 0 \quad \forall j \in I^{\mathrm{e}}(x),$$

$$\langle \nabla d_j(x), v \rangle > 0 \quad \forall j \in I^{\mathrm{i}}(x).$$

In [23], a simple optimization problem was exhibited, on which many recently proposed interior-point methods converge to infeasible points at which such a direction

[7] Nonemptiness of $\tilde{X}_0$ follows from the Mangasarian–Fromovitz constraint qualification (and Assumption 2), which in turn follows from Assumption 3(i).

v exists, in effect showing that convergence of these algorithms to KKT points cannot be proved unless a strong assumption is used that rules out such seemingly innocuous problems. On the other hand, it is readily checked that directions v as in Lemma 3.1 do exist at all spurious limit points identified in [23]. Indeed, in the problem from [23], for some a, b, with $b \geq 0$, $c_1(x) = (x^{(1)})^2 - x^{(2)} + a$, $c_2(x) = -x^{(1)} + x^{(3)} + b$,[8] $d_1(x) = x^{(2)}$, and $d_2(x) = x^{(3)}$, and the spurious limit points are points of the form $(\zeta, 0, 0)^{\mathrm{T}}$, with $\zeta < 0$, at which *both* c_1 *and* c_2 *are nonzero*; $v = (1, 1, 1)^{\mathrm{T}}$ meets our conditions at such points. In fact, it is readily verified that Assumption 3(ii) is satisfied whenever $a \geq 0$ and that, when $a < 0$, the only point $x \in \tilde{X} \setminus X$ at which the condition in Assumption 3(ii) is violated is $(-\sqrt{|a|}, 0, 0)^{\mathrm{T}}$, at which $c_1(x) = 0$. In section 5 we will discuss the behavior on this example of the algorithm proposed below.

Before presenting our algorithm, we briefly explore a connection between problems (P) and (P$_\rho$). A point x is a *KKT point* of (P) if there exist $y \in \mathcal{R}^{m_{\mathrm{e}}}$, $z \in \mathcal{R}^{m_{\mathrm{i}}}$ such that

$$g(x) - A(x)^{\mathrm{T}} y - B(x)^{\mathrm{T}} z = 0, \tag{3.5}$$

$$c(x) = 0, \tag{3.6}$$

$$d(x) \geq 0, \tag{3.7}$$

$$z^{(j)} d_j(x) = 0, \ j = 1, \ldots, m_{\mathrm{i}}, \tag{3.8}$$

$$z \geq 0. \tag{3.9}$$

Following [17] we term a point x *stationary* for (P) if there exist $y \in \mathcal{R}^{m_{\mathrm{e}}}$, $z \in \mathcal{R}^{m_{\mathrm{i}}}$ such that (3.5)–(3.8) hold (but possibly not (3.9)). Next, for given ρ, a point $x \in \tilde{X}$ is a KKT point of (P$_\rho$) if there exist $y \in \mathcal{R}^{m_{\mathrm{e}}}$, $z \in \mathcal{R}^{m_{\mathrm{i}}}$ such that

$$g(x) + A(x)^{\mathrm{T}}(\rho \mathrm{e}) - A(x)^{\mathrm{T}} y - B(x)^{\mathrm{T}} z = 0, \tag{3.10}$$

$$c(x) \geq 0, \tag{3.11}$$

$$d(x) \geq 0, \tag{3.12}$$

$$y^{(j)} c_j(x) = 0, \ j = 1, \ldots, m_{\mathrm{e}}, \tag{3.13}$$

$$y \geq 0, \tag{3.14}$$

$$z^{(j)} d_j(x) = 0, \ j = 1, \ldots, m_{\mathrm{i}}, \tag{3.15}$$

$$z \geq 0, \tag{3.16}$$

where $\mathrm{e} \in \mathcal{R}^{m_{\mathrm{e}}}$ is a vector whose components are all 1. A point x is stationary for (P$_\rho$) if there exist $y \in \mathcal{R}^{m_{\mathrm{e}}}$, $z \in \mathcal{R}^{m_{\mathrm{i}}}$ such that (3.10)–(3.13) and (3.15) hold (but possibly not (3.14) and (3.16)). The following proposition, found in [15], is crucial to the development and is repeated here for ease of reference.

PROPOSITION 3.2. *Suppose Assumptions* 1 *and* 2 *hold. Let* ρ *be given. If* x *is stationary for* (P$_\rho$) *with multiplier vectors* y *and* z *and* $c(x) = 0$, *then it is stationary for* (P) *with multiplier vectors* $y - \rho \mathrm{e}$ *and* z. *Furthermore, if* $z \geq 0$, *then* x *is a KKT point for* (P).

Proof. Using the fact that $c(x) = 0$, equations (3.10)–(3.13) and (3.15) imply

$$g(x) - A(x)^{\mathrm{T}}(y - \rho \mathrm{e}) - B(x)^{\mathrm{T}} z = 0, \quad c(x) = 0, \quad d(x) \geq 0, \quad z^{(j)} d_j(x) = 0 \quad \forall j.$$

[8] Our $c_2(x)$ is the negative of that in [23] because in our framework equality constraints must take on positive values at the initial point, while at the initial points of interest (as per Theorem 1 in [23]) $c_2(x)$ as defined in [23] is negative.

Thus x is stationary for (P) with multipliers $y - \rho e \in \mathcal{R}^{m_e}$ and $z \in \mathcal{R}^{m_i}$. The second assertion follows similarly. □

The proposed algorithm is based on solving problem (P_ρ) for fixed values of $\rho > 0$ using the interior-point method outlined in section 2. The key issue will then be to determine how to adjust ρ to force the iterate to asymptotically satisfy $c(x) = 0$.

For problem (P_ρ), the barrier function (2.2) becomes

$$\beta(x, \rho, \mu) = f(x) + \rho \sum_{j=1}^{m_e} c_j(x) - \sum_{j=1}^{m_e} {\mu_e}^{(j)} \ln(c_j(x)) - \sum_{j=1}^{m_i} {\mu_i}^{(j)} \ln(d_j(x)).$$

Its gradient is given by

$$\nabla_x \beta(x, \rho, \mu) = g(x) + A(x)^T(\rho e) - A(x)^T C(x)^{-1} \mu_e - B(x)^T D(x)^{-1} \mu_i. \tag{3.17}$$

Proceeding as in section 2, define

$$y = C(x)^{-1} \mu_e, \quad z = D(x)^{-1} \mu_i,$$

and consider solving the nonlinear system in (x, y, z):

$$g(x) + A(x)^T(\rho e - y) - B(x)^T z = 0, \tag{3.18}$$

$$\mu_e - C(x)y = 0, \tag{3.19}$$

$$\mu_i - D(x)z = 0, \tag{3.20}$$

by means of the (quasi-)Newton iteration

$$(L(x, y, z, W, \rho, \mu_e, \mu_i))$$
$$\begin{bmatrix} -W & A(x)^T & B(x)^T \\ YA(x) & C(x) & 0 \\ ZB(x) & 0 & D(x) \end{bmatrix} \begin{bmatrix} \Delta x \\ \Delta y \\ \Delta z \end{bmatrix} = \begin{bmatrix} g(x) + A(x)^T(\rho e - y) - B(x)^T z \\ \mu_e - C(x)y \\ \mu_i - D(x)z \end{bmatrix},$$

where $Y = \mathrm{diag}(y^{(j)})$, $Z = \mathrm{diag}(z^{(j)})$, and W is equal to, or approximates, the Hessian with respect to x, at (x, y, z), of the Lagrangian associated with (P_ρ).

System $L(x, y, z, W, \rho, \mu_e, \mu_i)$ is solved first with $(\mu_e, \mu_i) = (0, 0)$ and then with (μ_e, μ_i) set analogously to (2.7). Following that, a correction $\Delta\tilde{x}$ is computed by solving the appropriate linear least squares problem, and new iterates x^+, y^+, and z^+ are obtained as in section 2.

There remains the central issue of how ρ is updated. As noted in the introduction, Mayne and Polak [15] adaptively increase ρ to keep it above the magnitude of the most negative equality constraint multiplier estimate. They use a rather expensive estimation scheme, which was later improved upon in [13] in a different context. A simpler update rule is used here, which involves no computational overhead. It is based on the observation that ρ should be increased whenever convergence is detected to a point—a KKT point for (P_ρ), in view of the convergence properties established in [17] and reviewed in section 2—where some equality constraint is violated. Care must be exercised because, if such convergence is erroneously signaled (false alarm), a runaway phenomenon may be triggered, with ρ increasing uncontrollably without a KKT point of (P) being approached. We avoid this by requiring that the following three conditions—all of which are needed in the convergence proof—be satisfied in order for an increase of ρ to be triggered (here γ_1, γ_2, and γ_3 are prescribed positive

constants): (a) $\|\Delta x_k^0\| \le \gamma_1$, indicating the proximity of a stationary point for (P_{ρ_k}); (b) $y_k + \Delta y_k^0 \not\ge \gamma_2 e$; i.e., not all c_js become strongly binding as the limit point is approached; (c) $y_k + \Delta y_k^0 \ge -\gamma_3 e$ and $z_k + \Delta z_k^0 \ge -\gamma_3 e$; i.e., no components of y_k or z_k are diverging to $-\infty$ due to ρ_k being increased too fast (i.e., if ρ_k is growing large, either y_k and z_k become nonnegative or their negative components become negligible compared to ρ_k), the violation of which would indicate that the limit point is not KKT.

We are now ready to state the algorithm.

ALGORITHM A.

Parameters. $\xi \in (0, 1/2)$, $\eta \in (0,1)$, $\gamma_1 > 0$, $\gamma_2 > 0$, $\gamma_3 > 0$, $\nu > 2$, $\theta \in (0,1)$, $w_{\max} > 0$, $\delta > 1$, $\tau \in (2,3)$, $\kappa \in (0,1)$.

Data. $x_0 \in \tilde{X}_0$, $\rho_0 > 0$, $y_0^{(i)} \in (0, w_{\max}]$, $i = 1, \ldots, m_e$, $z_0^{(i)} \in (0, w_{\max}]$, $i = 1, \ldots, m_i$; $W_0 \in \mathcal{R}^{n \times n}$ such that

$$W_0 + \begin{bmatrix} A(x_0)^T & B(x_0)^T \end{bmatrix} \begin{bmatrix} C(x_0)^{-1} Y_0 & \bigcirc \\ \bigcirc & D(x_0)^{-1} Z_0 \end{bmatrix} \begin{bmatrix} A(x_0) \\ B(x_0) \end{bmatrix} \tag{3.21}$$

is positive definite.

Step 0: Initialization. Set $k = 0$.

Step 1: Computation of search arc:

i. Compute $(\Delta x_k^0, \Delta y_k^0, \Delta z_k^0)$ by solving $L(x_k, y_k, z_k, W_k, \rho_k, 0, 0)$. If $\Delta x_k^0 = 0$ and $m_e = 0$, then stop.

ii. Check the following three conditions: (i) $\|\Delta x_k^0\| \le \gamma_1$, (ii) $y_k + \Delta y_k^0 \not\ge \gamma_2 e$, (iii) $y_k + \Delta y_k^0 \ge -\gamma_3 e$ and $z_k + \Delta z_k^0 \ge -\gamma_3 e$. If all three conditions hold, then set $\rho_{k+1} = \delta \rho_k$, $x_{k+1} = x_k$, $y_{k+1} = y_k$, $z_{k+1} = z_k$, $W_{k+1} = W_k$, set $k = k+1$, and go back to Step 1i. Otherwise, proceed to Step 1iii.

iii. Compute $(\Delta x_k^1, \Delta y_k^1, \Delta z_k^1)$ by solving $L(x_k, y_k, z_k, W_k, \rho_k, \|\Delta x_k^0\|^\nu y_k, \|\Delta x_k^0\|^\nu z_k)$.

iv. Set

$$\varphi_k = \begin{cases} 1 & \text{if } \langle \nabla f_{\rho_k}(x_k), \Delta x_k^1 \rangle \le \theta \langle \nabla f_{\rho_k}(x_k), \Delta x_k^0 \rangle, \\ (1-\theta) \frac{\langle \nabla f_{\rho_k}(x_k), \Delta x_k^0 \rangle}{\langle \nabla f_{\rho_k}(x_k), \Delta x_k^0 - \Delta x_k^1 \rangle} & \text{otherwise.} \end{cases}$$

v. Set

$$\begin{aligned} \Delta x_k &= (1-\varphi_k)\Delta x_k^0 + \varphi_k \Delta x_k^1, \\ \Delta y_k &= (1-\varphi_k)\Delta y_k^0 + \varphi_k \Delta y_k^1, \\ \Delta z_k &= (1-\varphi_k)\Delta z_k^0 + \varphi_k \Delta z_k^1. \end{aligned}$$

vi. Set

$$I_k^e = \{j : c_j(x_k) \le y_k^{(j)} + \Delta y_k^{(j)}\}, \quad I_k^i = \{j : d_j(x_k) \le z_k^{(j)} + \Delta z_k^{(j)}\},$$

$$J_k^e = \{j : y_k^{(j)} + \Delta y_k^{(j)} \le -c_j(x_k)\}, \quad J_k^i = \{j : z_k^{(j)} + \Delta z_k^{(j)} \le -d_j(x_k)\}.$$

vii. Set $\Delta \tilde{x}_k$ to be the solution of the linear least squares problem

$$\min \tfrac{1}{2}\langle \Delta \tilde{x}, W_k \Delta \tilde{x} \rangle \text{ s.t. } \begin{array}{ll} c_j(x_k + \Delta x_k) + \langle \nabla c_j(x_k), \Delta \tilde{x}_k \rangle = \psi_k & \forall j \in I_k^e, \\ d_j(x_k + \Delta x_k) + \langle \nabla d_j(x_k), \Delta \tilde{x}_k \rangle = \psi_k & \forall j \in I_k^i, \end{array} \tag{3.22}$$

where

$$\psi_k = \max\left\{\|\Delta x_k\|^\tau, \max_{j\in I_k^{\mathrm{e}}}\left|\frac{\Delta y_k^{(j)}}{y_k^{(j)}+\Delta y_k^{(j)}}\right|^\kappa \|\Delta x_k\|^2, \max_{j\in I_k^{\mathrm{i}}}\left|\frac{\Delta z_k^{(j)}}{z_k^{(j)}+\Delta z_k^{(j)}}\right|^\kappa \|\Delta x_k\|^2\right\}.$$

If $J_k^{\mathrm{e}} \cup J_k^{\mathrm{i}} \neq \emptyset$ or (3.22) is infeasible or unbounded or $\|\Delta\tilde{x}_k\| > \|\Delta x_k\|$, set $\Delta\tilde{x}_k$ to 0.

Step 2. Arc search. Compute α_k, the first number α in the sequence $\{1, \eta, \eta^2, \dots\}$ satisfying

$$\begin{aligned}
&f_{\rho_k}(x_k + \alpha\Delta x_k + \alpha^2\Delta\tilde{x}_k) \le f_{\rho_k}(x_k) + \xi\alpha\langle\nabla f_{\rho_k}(x_k), \Delta x_k\rangle,\\
&c_j(x_k + \alpha\Delta x_k + \alpha^2\Delta\tilde{x}_k) > 0 \quad \forall j,\\
&d_j(x_k + \alpha\Delta x_k + \alpha^2\Delta\tilde{x}_k) > 0 \quad \forall j,\\
&c_j(x_k + \alpha\Delta x_k + \alpha^2\Delta\tilde{x}_k) \ge c_j(x_k) \quad \forall j \in J_k^{\mathrm{e}},\\
&d_j(x_k + \alpha\Delta x_k + \alpha^2\Delta\tilde{x}_k) \ge d_j(x_k) \quad \forall j \in J_k^{\mathrm{i}}.
\end{aligned}$$

Step 3. Updates. Set $x_{k+1} = x_k + \alpha_k\Delta x_k + \alpha_k^2\Delta\tilde{x}_k$. If $J_k^{\mathrm{e}} \cup J_k^{\mathrm{i}} = \emptyset$, set

$$\begin{aligned}
y_{k+1}^{(j)} &= \min\{\max\{\|\Delta x_k\|^2, y_k^{(j)} + \Delta y_k^{(j)}\}, w_{\max}\}, \quad j = 1, \dots, m_{\mathrm{e}},\\
z_{k+1}^{(j)} &= \min\{\max\{\|\Delta x_k\|^2, z_k^{(j)} + \Delta z_k^{(j)}\}, w_{\max}\}, \quad j = 1, \dots, m_{\mathrm{i}};
\end{aligned}$$

otherwise, set $y_{k+1} = y_0$ and $z_{k+1} = z_0$. Set $\rho_{k+1} = \rho_k$ and select W_{k+1} such that

$$W_{k+1} + \begin{bmatrix} A(x_{k+1}) & B(x_{k+1}) \end{bmatrix}^{\mathrm{T}} \begin{bmatrix} C(x_{k+1})^{-1}Y_{k+1} & \bigcirc \\ \bigcirc & D(x_{k+1})^{-1}Z_{k+1} \end{bmatrix} \begin{bmatrix} A(x_{k+1}) \\ B(x_{k+1}) \end{bmatrix}$$

is positive definite. Set $k = k + 1$ and go back to Step 1.

REMARK 1. *The values assigned to y_{k+1} and z_{k+1} in Step* 1ii *are of no consequence as far as the theoretical properties of the algorithm are concerned, provided dual feasibility is preserved. Rather than reusing the previous values as stated in the algorithm, it may be advisable to make use of the just computed corrections Δy_k^0 and Δz_k^0, e.g., by setting*

$$\begin{aligned}
y_{k+1}^{(j)} &= \min\{\max\{y_k^{(j)}, y_k^{(j)} + \Delta y_k^{0,(j)}\}, w_{\max}\}, \quad j = 1, \dots, m_{\mathrm{e}},\\
z_{k+1}^{(j)} &= \min\{\max\{z_k^{(j)}, z_k^{(j)} + \Delta z_k^{0,(j)}\}, w_{\max}\}, \quad j = 1, \dots, m_{\mathrm{i}},
\end{aligned}$$

which still ensures dual feasibility. (A side effect of such a rule is that the components of y_k and z_k are possibly increased but never decreased when ρ_k is increased, which makes some intuitive sense.)

REMARK 2. *Similarly, variations can be considered for the dual variable update rule in Step* 3 *in the case when $J_k^{\mathrm{e}} \cup J_k^{\mathrm{i}} \neq \emptyset$. Indeed the convergence analysis of* [17] *remains unaffected as long as the components of y_{k+1} and z_{k+1} stay bounded away from zero (and bounded) over the set of iterates k at which $J_k^{\mathrm{e}} \cup J_k^{\mathrm{i}} \neq \emptyset$. A possible choice would be*

$$\begin{aligned}
y_{k+1}^{(j)} &= \min\{\max\{w_{\min}, y_k^{(j)} + \Delta y_k^{(j)}\}, w_{\max}\}, \quad j = 1, \dots, m_{\mathrm{e}},\\
z_{k+1}^{(j)} &= \min\{\max\{w_{\min}, z_k^{(j)} + \Delta z_k^{(j)}\}, w_{\max}\}, \quad j = 1, \dots, m_{\mathrm{i}},
\end{aligned}$$

where $w_{\min} \in (0, w_{\max})$ *is prescribed. Unlike the update rule used in the algorithm statement (taken from* [17]*) this rule attempts to make use of some of the multiplier estimates even when* $J_k^{\mathrm{e}} \cup J_k^{\mathrm{i}} \neq \emptyset$.

REMARK 3. *If an initial point* $x_0 \in \tilde{X}_0$ *is not readily available, a point* $x_0 \in \tilde{X}$ *can be constructed as follows:* (i) *Perform a "Phase I" search by maximizing* $\min_j d_j(x)$ *without constraints. This can be done, e.g., by applying Algorithm A to the problem*

$$\max_{(x,\zeta)\in\mathcal{R}^{n+1}} \zeta \text{ s.t. } d_j(x) - \zeta \geq 0 \quad \forall j.$$

A point x_0 *satisfying* $\min_j d_j(x) \geq 0$ *will eventually be obtained (or the constructed sequence* $\{x_k\}$ *will be unbounded), provided* $\min_j d_j(x)$ *has no stationary point with negative value, i.e., provided that, for all* x *such that* $\zeta := \min_j d_j(x) < 0$, *the origin does not belong to the convex hull of* $\{\nabla d_j(x) : d_j(x) = \zeta\}$. (ii) *Redefine* $c_j(x)$ *to take values* $-c_j(x)$ *for every* j *such that the original* $c_j(x)$ *is negative. As a result,* x_0 *will be in* $\tilde{X}$ *for the reformulated problem. If it is on the boundary of* $\tilde{X}$ *rather than in its interior* $\tilde{X}_0$, *it can be readily perturbed into a point in* $\tilde{X}_0$ *(under Assumption* 3(i)*).*

4. Convergence analysis. We first show that Algorithm A is well defined. First of all, the conditions imposed on W_0 and (in Step 3) on W_k in Algorithm A are identical, for every fixed k, to the second condition in Assumption PTH-A6*. Thus the matrix in our (quasi-)Newton iteration is nonsingular, and it is readily checked that if $\Delta x_k^0 = 0$ for some k, then $\nabla f_{\rho_k}(x_k) = 0$; i.e., x_k is an unconstrained KKT point for (P_ρ) (cf. Proposition 3.4 of [17]); and it is readily checked that in such a case, $y_k + \Delta y_k^0$ and $z_k + \Delta z_k^0$ are the associated KKT multiplier vectors, i.e., are both zero. Thus, if finite termination occurs at Step 1i (i.e., $m_{\mathrm{e}} = 0$), then $\nabla f(x_k) = 0$; i.e., x_k is an unconstrained KKT point for (P); and if $\Delta x_k^0 = 0$ but finite termination does not occur (i.e., $m_{\mathrm{e}} > 0$), then conditions (i) through (iii) in Step 1ii are satisfied, and the algorithm loops back to Step 1i. Thus Step 1iii is never executed when Δx_k^0 is zero. It then follows from Proposition 3.3 of [17] that under Assumptions 1, 2, and 3(i), Algorithm A is well defined. (Assumptions A4 through A6 of [17] are not needed in that proposition.)

From now on, we assume that the algorithm never stops, i.e., that an infinite sequence $\{x_k\}$ is constructed. Our next task will be to show that unless $\{x_k\}$ itself is unbounded, ρ_k is increased at most finitely many times. Assumption 3(ii) will be crucial here. An additional assumption, adapted from PTH-A6*, will be needed as well.

ASSUMPTION 4. *Given any index set* K *such that the sequence* $\{x_k\}$ *constructed by Algorithm A is bounded, there exist* $\sigma_1, \sigma_2 > 0$ *such that, for all* $k \in K$,

$$\|W_k\| \leq \sigma_2$$

and

$$\left\langle v, \left(W_k + \sum_{i=1}^{m_{\mathrm{e}}} \frac{y_k^{(i)}}{c_i(x_k)} \nabla c_i(x_k) \nabla c_i(x_k)^{\mathrm{T}} + \sum_{i=1}^{m_{\mathrm{i}}} \frac{z_k^{(i)}}{d_i(x_k)} \nabla d_i(x_k) \nabla d_i(x_k)^{\mathrm{T}} \right) v \right\rangle$$

$$\geq \sigma_1 \|v\|^2 \quad \forall v \in \mathcal{R}^n.$$

LEMMA 4.1. *Suppose Assumptions* 1–4 *hold. If the infinite sequence* $\{x_k\}$ *generated by Algorithm A is bounded, then* ρ_k *is increased at most finitely many times.*

Proof. The proof is by contradiction. Suppose ρ_k is increased infinitely many times; i.e., there exists an infinite index set $\mathcal{K}$ such that $\rho_{k+1} > \rho_k$ for all $k \in \mathcal{K}$. The criteria that trigger ρ_k to increase must thus be satisfied for all $k \in \mathcal{K}$, i.e., with $y'_k = y_k + \Delta y_k^0$ and $z'_k = z_k + \Delta z_k^0$,

$$\|\Delta x_k^0\| \le \gamma_1 \quad \forall k \in \mathcal{K}, \tag{4.1}$$

$$y'_k \not\ge \gamma_2 \mathrm{e} \quad \forall k \in \mathcal{K}, \tag{4.2}$$

$$y'_k \ge -\gamma_3 \mathrm{e} \quad \forall k \in \mathcal{K}, \tag{4.3}$$

$$z'_k \ge -\gamma_3 \mathrm{e} \quad \forall k \in \mathcal{K}. \tag{4.4}$$

As per Step 1i of Algorithm A, we have

$$W_k \Delta x_k^0 + g(x_k) + A(x_k)^{\mathrm{T}}(\rho_k \mathrm{e} - y'_k) - B(x_k)^{\mathrm{T}} z'_k = 0, \tag{4.5}$$

$$Y_k A(x_k) \Delta x_k^0 + C(x_k) y'_k = 0, \tag{4.6}$$

$$Z_k B(x_k) \Delta x_k^0 + D(x_k) z'_k = 0. \tag{4.7}$$

Since $\{\rho_k\}$ tends to infinity, it follows from (4.2) that $\{\|\rho_k \mathrm{e} - y'_k\|_\infty\}$ tends to infinity on $\mathcal{K}$. Consequently, the sequence $\{\alpha_k\}$, with

$$\alpha_k = \max\left\{\|\rho_k \mathrm{e} - y'_k\|_\infty, \|z'_k\|_\infty, 1\right\},$$

tends to infinity on $\mathcal{K}$ as well. Define

$$\hat{y}_k = \alpha_k^{-1}(\rho_k \mathrm{e} - y'_k), \tag{4.8}$$

$$\hat{z}_k = \alpha_k^{-1} z'_k \tag{4.9}$$

for $k \in \mathcal{K}$. By construction $\max\{\|\hat{y}_k\|_\infty, \|\hat{z}_k\|_\infty\} = 1$ for all $k \in \mathcal{K}$, k large enough. Since in addition the sequence $\{x_k\}_{k\in\mathcal{K}}$ is bounded by assumption, there must exist an infinite index set $\mathcal{K}' \subseteq \mathcal{K}$ and vectors $x^* \in \mathcal{R}^n$, $\hat{y}^* \in \mathcal{R}^{m_e}$, and $\hat{z}^* \in \mathcal{R}^{m_i}$, with $\hat{y}^*$ and $\hat{z}^*$ not both zero, such that

$$\lim_{\substack{k\to\infty \\ k\in\mathcal{K}'}} x_k = x^*, \quad \lim_{\substack{k\to\infty \\ k\in\mathcal{K}'}} \hat{y}_k = \hat{y}^*, \quad \lim_{\substack{k\to\infty \\ k\in\mathcal{K}'}} \hat{z}_k = \hat{z}^*.$$

Boundedness of $\{x_k\}$ and the continuity assumptions imply that $\{A(x_k)\}$ and $\{B(x_k)\}$ are bounded. Further, $\{Y_k\}$ and $\{Z_k\}$ are bounded by construction. Dividing both sides of (4.6) by α_k, letting $k \to \infty$, $k \in \mathcal{K}'$, and using (4.1) shows that

$$\alpha_k^{-1} C(x_k) y'_k \to 0 \quad \text{as } k \to \infty,\ k \in \mathcal{K}',$$

implying that, for every $j \notin I^{\mathrm{e}}(x^*)$,

$$\alpha_k^{-1} y_k'^{,(j)} \to 0 \quad \text{as } k \to \infty,\ k \in \mathcal{K}'.$$

Together with (4.8), this implies that $\{\rho_k/\alpha_k\}$ converges to some limit $\omega \ge 0$ as $k \to \infty$, $k \in \mathcal{K}'$, with

$$\hat{y}^{*,(j)} = \omega \quad \forall\ j \notin I^{\mathrm{e}}(x^*).$$

Next it follows from (4.3), (4.4), (4.8), and (4.9) that

$$\hat{y}^{*,(j)} \le \omega \quad \forall j, \quad \hat{z}^* \ge 0. \tag{4.10}$$

Further, dividing (4.7) by α_k and taking the limit as $k \to \infty$, $k \in \mathcal{K}'$, yields

$$D\,(x^*)\,\hat{z}^* = 0.$$

Thus $\hat{z}^{*,(j)} = 0$ for all $j \notin I^{\mathrm{i}}(x^*)$. Finally, in view of (4.1) and Assumption 4, dividing (4.5) by α_k and taking the limit as $k \to \infty$, $k \in \mathcal{K}'$, yields

$$A(x^*)^{\mathrm{T}}\hat{y}^* - B(x^*)^{\mathrm{T}}\hat{z}^* = 0,$$

i.e.,

$$\sum_{j=1}^{m_{\mathrm{e}}} \hat{y}^{*,(j)} \nabla c_j(x^*) - \sum_{j \in I^{\mathrm{i}}(x^*)} \hat{z}^{*,(j)} \nabla d_j(x^*) = 0. \tag{4.11}$$

Since $\hat{y}^*$ and $\hat{z}^*$ are not both zero, (4.11) together with Assumption 3(i) implies that $I^{\mathrm{e}}(x^*) \neq \{1, \ldots, m_{\mathrm{e}}\}$ (i.e., $x^* \notin X$) and $\omega > 0$. Dividing both sides of (4.11) by ω and adding and subtracting $\sum_{j \in I^{\mathrm{e}}(x^*)} \nabla c_j(x^*)$ then yields

$$\sum_{j=1}^{m_{\mathrm{e}}} \nabla c_j(x^*) - \sum_{j \in I^{\mathrm{e}}(x^*)} y^{(j)} \nabla c_j(x^*) - \sum_{j \in I^{\mathrm{i}}(x^*)} z^{(j)} \nabla d_j(x^*) = 0,$$

where we have defined $y^{(j)} = 1 - \frac{\hat{y}^{*,(j)}}{\omega}$ and $z^j = \frac{\hat{z}^{*,(j)}}{\omega}$. In view of (4.10) and since $x^* \notin X$, this contradicts Assumption 3(ii). □

In what follows, $\bar{\rho}$ denotes the final value of ρ_k.

Algorithm A now reduces to the algorithm described in section 2 applied to problem $(\mathrm{P}_{\bar{\rho}})$. It is shown in [17] that under Assumptions 1–4, if the sequence $\{x_k\}$ constructed by Algorithm A is bounded, then all its accumulation points are stationary for $(\mathrm{P}_{\bar{\rho}})$. To conclude that they are KKT points for $(\mathrm{P}_{\bar{\rho}})$, an additional assumption is used. Recall that $\bar{\rho}$ is of the form $\rho_0 \delta^\ell$ for some nonnegative integer ℓ.

Assumption 5. *For $\rho \in \{\rho_0 \delta^\ell :\ \ell$ a nonnegative integer$\}$, all stationary points of (P_ρ) are isolated.*

Thus, under Assumptions 1–5, all accumulation points of $\{x_k\}$ are KKT points for $(\mathrm{P}_{\bar{\rho}})$ (Theorem 3.11 in [17]). Now, since ρ_k eventually stops increasing, at least one of the conditions in Step 1ii of Algorithm A is not eventually always satisfied. For convergence to KKT points of (P) to be guaranteed, the fact that condition (ii) in Step 1ii of Algorithm A must eventually be violated if ρ_k stops increasing is crucial, since this would imply that $c(x_k)$ goes to zero. A glance at the three conditions in that step suggests that this will be the case if the dual variables converge to the KKT multipliers for $(\mathrm{P}_{\bar{\rho}})$ (since in such a case conditions (i) and (iii) will eventually hold). To prove that the latter indeed occurs, one more assumption is used.

Assumption 6. *The sequence $\{x_k\}$ generated by Algorithm A has an accumulation point which is an isolated KKT point for $(\mathrm{P}_{\bar{\rho}})$ and at which strict complementarity holds.*

Proposition 4.2. *Suppose Assumptions 1–6 hold. If the infinite sequence $\{x_k\}$ generated by Algorithm A is bounded, then it converges to a KKT point x^* of $(\mathrm{P}_{\bar{\rho}})$. Moreover, with y^* and z^* the associated KKT multiplier vectors corresponding, respectively, to the "c" and "d" constraints,*

(i) *$\{\Delta x_k\} \to 0$ as $k \to \infty$, $\{y_k + \Delta y_k\} \to y^*$ as $k \to \infty$, and $\{z_k + \Delta z_k\} \to z^*$ as $k \to \infty$;*

(ii) *for k large enough, $J_k^{\mathrm{e}} = \emptyset = J_k^{\mathrm{i}}$, $I_k^{\mathrm{e}} = I^{\mathrm{e}}(x^*)$, and $I_k^{\mathrm{i}} = I^{\mathrm{i}}(x^*)$;*

(iii) *if* $y^{*,(j)} \leq w_{\max}$ *for all* j*, then* $\{y_k\} \to y^*$ *as* $k \to \infty$*; if* $z^{*,(j)} \leq w_{\max}$ *for all* j*, then* $\{z_k\} \to z^*$ *as* $k \to \infty$.

Proof. The proof follows from Proposition 4.2 in [17], noting that our Assumption 6 is the only portion of Assumption A8 of [17] that is needed in the proofs of Lemma 4.1 of [17] and Proposition 4.2 of [17]. ∎

THEOREM 4.3. *Suppose Assumptions* 1–6 *hold. If the infinite sequence* $\{x_k\}$ *generated by Algorithm A is bounded, then it converges to a KKT point* x^* *of* (P). *Moreover, in such a case,* $\{y_k + \Delta y_k - \rho \mathrm{e}\}$ *converges to* $\bar{y}^*$ *and* $\{z_k + \Delta z_k\}$ *converges to* z^**, where* $\bar{y}^*$ *and* z^* *are the multiplier vectors associated with* x^* *for problem* (P).

Proof. We know from Proposition 4.2 that (i) $\{x_k\} \to x^*$, a KKT point for $(\mathrm{P}_{\bar{\rho}})$; (ii) $\{\Delta x_k\} \to 0$; (iii) $\{y_k + \Delta y_k\} \to y^* \geq 0$, the multiplier vector associated with the "c" constraints; and (iv) $\{z_k + \Delta z_k\} \to z^* \geq 0$, the multiplier vector associated with the "d" constraints. Further, in view of strict complementarity, it follows from Lemma PTH-3.1* in the appendix that the matrix in $L(x^*, y^*, z^*, W^*, \bar{\rho}, 0, 0)$ is nonsingular given any accumulation point W^* of $\{W_k\}$. Together with (i), (iii), and (iv) above, and since $L(x^*, y^*, z^*, W^*, \bar{\rho}, 0, 0)$ admits $(0, y^*, z^*)$ as its unique solution, this implies that on every subsequence on which $\{W_k\}$ converges, $\{\Delta x_k^0\}$ goes to 0, $\{y_k + \Delta y_k^0\}$ goes to y^*, and $\{z_k + \Delta z_k^0\}$ goes to z^*. As a consequence (invoking Assumption 4 and a simple contradiction argument), without the need to go down to a subsequence, $\{\Delta x_k^0\} \to 0$, $\{y_k + \Delta y_k^0\} \to y^*$, and $\{z_k + \Delta z_k^0\} \to z^*$. Thus conditions (i) and (iii) in Step 1ii of Algorithm A are all satisfied for k large enough. Since $\rho_k = \bar{\rho}$ for k large enough, it follows from Step 1ii of Algorithm A that condition (ii) must fail for k large enough, i.e., $y_k + \Delta y_k^0 \geq \gamma_2 \mathrm{e}$ for k large enough, implying that $y^* \geq \gamma_2 \mathrm{e}$. Since $\gamma_2 > 0$, it follows from complementary slackness that $c(x^*) = 0$. Since the algorithm generates feasible iterates, we are guaranteed that $d_j(x^*) \geq 0$, $j = 1, \ldots, m_i$. Application of Proposition 3.2 concludes the proof of the first claim. The second claim then follows from Proposition 3.2 and Proposition 4.2(i). ∎

Rate of convergence results are inherited from the results in [17]. We report them here for ease of reference. As above, let $\bar{y}^*$ and z^* be the multipliers associated with KKT point x^* of (P). The Lagrangian associated with (P) is given by

$$\mathcal{L}(x, \bar{y}, z) = f(x) - \langle \bar{y}, c(x) \rangle - \langle z, d(x) \rangle.$$

With the correspondence $\bar{y} = y - \bar{\rho}\mathrm{e}$, it is identical to the Lagrangian associated with $(\mathrm{P}_{\bar{\rho}})$, i.e.,

$$\mathcal{L}_{\bar{\rho}}(x, y, z) = f(x) + \bar{\rho} \sum_{j=1}^{m_\mathrm{e}} c_j(x) - \langle y, c(x) \rangle - \langle z, d(x) \rangle.$$

ASSUMPTION 7. *f, c_j, $j = 1, \ldots, m_\mathrm{e}$, and d_j, $j = 1, \ldots, m_\mathrm{i}$, are three times continuously differentiable. Furthermore, the second order sufficiency condition holds (with strict complementarity under Assumption* 6*) for* (P) *at* x^**; i.e.,* $\nabla^2 \mathcal{L}_{xx}(x^*, \bar{y}^*, z^*)$ *is positive definite on the subspace*

$$\{v \in \mathcal{R}^n \ \langle \nabla c_j(x^*), v \rangle = 0 \ \forall j, \langle \nabla d_j(x^*), v \rangle = 0 \ \forall j \in I^\mathrm{i}(x^*)\}.$$

It is readily checked that the second order sufficiency condition for $(\mathrm{P}_{\bar{\rho}})$ is identical to that for (P).

As a final assumption, superlinear convergence requires that the sequence $\{W_k\}$ asymptotically carry appropriate second order information.

ASSUMPTION 8.

$$\frac{\|N_k(W_k - \nabla^2_{xx}\mathcal{L}(x^*, \bar{y}^*, z^*))N_k\Delta x_k\|}{\|\Delta x_k\|} \to 0 \quad as\ k \to \infty, \tag{4.12}$$

where

$$N_k = I - \hat{G}_k^{\mathrm{T}}\left(\hat{G}_k\hat{G}_k^{\mathrm{T}}\right)^{-1}\hat{G}_k$$

with

$$\hat{G}_k = [\nabla c_j(x_k),\ j = 1, \ldots, m_{\mathrm{e}}, \nabla d_j(x_k),\ j \in I^{\mathrm{i}}(x^*)]^{\mathrm{T}} \in \mathcal{R}^{(m_{\mathrm{e}}+|I(x^*)|)\times n}.$$

THEOREM 4.4. *Suppose Assumptions* 1–8 *hold, and suppose that* $y^{*,(j)} \le w_{\max}$, $j = 1, \ldots, m_{\mathrm{e}}$, *and* $z^{*,(j)} \le w_{\max}$, $j = 1, \ldots, m_{\mathrm{i}}$. *Then the arc search in Step* 2 *of Algorithm A eventually accepts a full step of one, i.e.,* $\alpha_k = 1$ *for all* k *large enough, and* $\{x_k\}$ *converges to* x^* *two-step superlinearly, i.e.,*

$$\lim_{k\to\infty} \frac{\|x_{k+2} - x^*\|}{\|x_k - x^*\|} = 0.$$

Finally, it is readily verified that, under Assumption 7, for k large enough, Assumption 4 holds when W_k is selected to be *equal* to the Hessian of the Lagrangian $\mathcal{L}_{\bar{\rho}}$. In view of the discussion in section 2.3, Q-quadratic convergence follows.

THEOREM 4.5. *Suppose Assumptions* 1–7 *hold, suppose that, at every iteration except possibly finitely many,* W_k *is selected as*

$$W_k = \nabla^2_{xx}\mathcal{L}_{\rho_k}(x_k, y_k, z_k),$$

and suppose that $y^{*,(j)} \le w_{\max}$, $j = 1, \ldots, m_{\mathrm{e}}$, *and* $z^{*,(j)} \le w_{\max}$, $j = 1, \ldots, m_{\mathrm{i}}$. *Then* (x_k, y_k, z_k) *converges to* (x^*, y^*, z^*) *Q-quadratically; equivalently,* $(x_k, y_k - \rho_k \mathrm{e}, z_k)$ *converges to* $(x^*, \bar{y}^*, z^*)$, *Q-quadratically; i.e., there exists a constant* $\Gamma > 0$ *such that*

$$\left\| \begin{bmatrix} x_{k+1} - x^* \\ y_{k+1} - y^* \\ z_{k+1} - z^* \end{bmatrix} \right\| \le \Gamma \left\| \begin{bmatrix} x_k - x^* \\ y_k - y^* \\ z_k - z^* \end{bmatrix} \right\|^2 \quad \forall k; \tag{4.13}$$

equivalently,

$$\left\| \begin{bmatrix} x_{k+1} - x^* \\ y_{k+1} - \rho_{k+1}\mathrm{e} - \bar{y}^* \\ z_{k+1} - z^* \end{bmatrix} \right\| \le \Gamma \left\| \begin{bmatrix} x_k - x^* \\ y_k - \rho_k\mathrm{e} - \bar{y}^* \\ z_k - z^* \end{bmatrix} \right\|^2 \quad \forall k.$$

5. Numerical tests.

5.1. Technical aspects. We tested a MATLAB 6.1 implementation of Algorithm A with the following differences in comparison with the algorithm statement of section 3:

- The suggestion made in Remark 2 was adopted.
- In the update formulae for the multipliers in Step 3, $\|\Delta x_k\|^2$ was changed to $\min\{w_{\min}, \|\Delta x_k\|^2\}$ in both places. The motivation is that $\|\Delta x_k\|$ is meaningful only when it is small. This change does not affect the convergence analysis.

The following parameter values were used: $\xi = 10^{-4}$, $\eta = 0.8$, $\gamma_1 = 1$, $\gamma_2 = 1$, $\gamma_3 = 1$, $\nu = 3$, $\theta = 0.8$, $w_{\min} = 10^{-4}$ (see Remark 2), $w_{\max} = 10^{20}$, $\delta = 2$, $\tau = 2.5$, and $\kappa = 0.5$.

In our tests, we allowed for the initial point x_0 to lie on the boundary of the feasible set $\tilde{X}$. It is readily checked that in such a case, under our assumptions, $L(x_0, y_0, z_0, W_0, \rho_0, 0, 0)$ is still uniquely solvable and, unless $\Delta x_0^0 = 0$, the initial iteration is still well defined and yields a strictly feasible second iterate. (When $\Delta x_0^0 = 0$ and $c(x_0) = 0$, x_0 is stationary for (P). When $\Delta x_0^0 = 0$ but $c(x_0) \neq 0$, x_0 is stationary for (P_{ρ_0}) but not for (P), and unless the sum of the gradients of inactive $c_j(x_0)$'s belongs to the span of the gradients of all active constraints, increasing ρ forces a nonzero Δx_0^0 and the iteration can proceed.) When x_0 is not in the interior of $\tilde{X}$, the condition to be satisfied by W_0 must be modified by replacing in (3.21) the infinite (diagonal) entries of $C(x_0)^{-1}$ and $D(x_0)^{-1}$ by 0, and by requiring positive definiteness of the modified expression merely on the tangent plane to the active constraints, i.e., on

$$\left\{v \in \mathcal{R}^n : \langle \nabla c_i(x_0), v\rangle = 0,\ \langle \nabla d_j(x_0), v\rangle = 0\ \forall i \in I^{\mathrm{e}}(x_0), j \in I^{\mathrm{i}}(x_0)\right\}.$$

In the numerical tests reported below, the initial value x_0 was selected in each case as specified in the source of the test problem. Initial values y_0, z_0, and ρ_0 were selected as follows. Let y_0' and z_0' be the (linear least squares) solution of

$$\min_{y_0', z_0'} \|g(x_0) - A(x_0)y_0' - B(x_0)z_0'\|^2.$$

Then ρ_0 was set to the smallest power of δ that is no less than $\max\{1, \max_j\{\gamma_2 - {y_0'}^{(j)}\}\}$, and, for all j, $y_0^{(j)}$ was set to $y_0^{(j)} + \rho_0$ and $z_0^{(j)}$ to $\max\{0.1, z_0^{(j)}\}$. In all the tests, y_0 and z_0 thus defined satisfied the condition specified in the algorithm that their components should all be no larger than $w_{\max}$.

Next, for $k = 0, 1, \ldots$, W_k was constructed as follows from second order derivative information. Let $\lambda_{\min}$ be the leftmost eigenvalue of the restriction of the matrix

$$\nabla^2_{xx}\mathcal{L}_{\rho_k}(x_k, y_k, z_k) + \sum_{i \in I_k^{\mathrm{e}'}} \frac{y_k^{(i)}}{c_i(x_k)} \nabla c_i(x_k) \nabla c_i(x_k)^{\mathrm{T}} + \sum_{i \in I_k^{\mathrm{i}'}} \frac{z_k^{(i)}}{d_i(x_k)} \nabla d_i(x_k) \nabla d_i(x_k)^{\mathrm{T}},$$

where $I_k^{\mathrm{e}'}$ and $I_k^{\mathrm{i}'}$ are the sets of indices of "c" and "d" constraints with value *larger* than 10^{-10}, to the tangent plane to the constraints left out of the sum, i.e., to the subspace

$$\left\{v \in \mathcal{R}^n : \langle \nabla c_i(x_k), v\rangle = 0,\ \langle \nabla d_j(x_k), v\rangle = 0\ \forall i \notin I_k^{\mathrm{e}'}, j \notin I_k^{\mathrm{i}'}\right\}.$$

Then, set

$$W_k = \nabla^2_{xx}\mathcal{L}_{\rho_k}(x_k, y_k, z_k) + h_k I,$$

where

$$h_k = \begin{cases} 0 & \text{if } \lambda_{\min} > 10^{-5}, \\ -\lambda_{\min} + 10^{-5} & \text{if } |\lambda_{\min}| \leq 10^{-5}, \\ 2|\lambda_{\min}| & \text{otherwise.} \end{cases}$$

Note that under our regularity assumptions (which imply that W_k is bounded whenever x_k is bounded), this ensures that Assumption 4 holds. The motivation for the third alternative is to preserve the order of magnitude of the eigenvalues and condition number.

The stopping criterion (inserted at the end of Step 1i) was as follows, with $\epsilon_{\text{stop}} = 10^{-8}$. First, accounting for the fact that in our tests we allowed the initial point to lie on the boundary of $\tilde{X}$, we stopped with the error message "initial point stationary for (P)" if $\|\Delta x_0^0\| < 0.001\epsilon_{\text{stop}}$ and $\|c(x_0)\|_\infty < \epsilon_{\text{stop}}$. Second, the run was deemed to have terminated successfully if at any iteration k

$$\max\left\{\|c(x_k)\|_\infty, \max_j\left\{-\left(y_k^{(j)} + \Delta y_k^{0(j)}\right)\right\}, \max_j\left\{-\left(z_k^{(j)} + \Delta z_k^{0(j)}\right)\right\}\right\} < \epsilon_{\text{stop}}$$

and either

$$\|\Delta x_k^0\|_\infty < \epsilon_{\text{stop}}$$

or

$$\max\left\{\|\nabla_x \mathcal{L}(x_k, y_k, z_k)\|_\infty, \max_j\left\{z_k^{(j)} d_j(x_k)\right\}\right\} < \epsilon_{\text{stop}}.$$

Iterations at which only Steps 1i and 1ii are executed were not included in our iteration counts. The reason is that the computational cost of these iterations is dramatically less than that of "regular" iterations: no additional function evaluations and no additional matrix factorization—the same factorization is later used at the next regular iteration. All tests were run within the CUTEr testing environment [9], on a SUN Enterprise 250 with two UltraSparc-II 400MHz processors, running Solaris 2.7.

5.2. Numerical results. We first considered two instances of the example from [23] briefly discussed in section 3 (immediately following Lemma 3.1), specifically $(a, b) = (1, 1)$ with $(-3, 1, 1)^{\mathrm{T}}$ as initial guess, and $(a, b) = (-1, 1/2)$ with $(-2, 1, 1)^{\mathrm{T}}$ as initial guess, both of which satisfy the conditions in Theorem 1 of [23]. In both cases we used $f(x) = x^{(1)}$ as objective function (as in the example of section 4 of [23]). Recall that under those conditions, all methods of type "Algorithm I" in [23] construct sequences that converge to points of the form $(\zeta, 0, 0)^{\mathrm{T}}$, with $\zeta < 0$, where both c_1 and c_2 are nonzero. As noted in our earlier discussion, Assumption 3(ii) is satisfied in the first instance, while in the second instance the condition in that assumption is violated only at $\hat{x} := (-1, 0, 0)^{\mathrm{T}}$ (with $c_1(\hat{x}) = 0$). Thus, at $\hat{x}$, there is no direction of strict descent for $c_1(x) + c_2(x)$ (the ℓ_1 norm of $c(x)$ when $x \in \tilde{X}$) that is feasible for $c_1(x) \geq 0$, $d_1(x) \geq 0$, and $d_2(x) \geq 0$.

In the first instance, our Algorithm A was observed to converge to the unique global solution $(1, 2, 0)^{\mathrm{T}}$ in 13 iterations, with a final penalty parameter value $\bar{\rho}$ of 4. In the second instance, Algorithm A failed in that the constructed sequence converged to the infeasible point $\hat{x}$. Interestingly, it can be checked that, at $\hat{x}$, there is not even a descent direction for $\|c(x)\|_1$ that is feasible for the mere bound constraints $d_1(x) \geq 0$ and $d_2(x) \geq 0$.

REMARK 4. *For the second instance of the example of* [23] *just discussed, directions do exist at $\hat{x}$ that are of strict descent for the* Euclidean *norm of $c(x)$ and are feasible for the bound constraints. The existence of such directions allows the algorithm proposed in* [3] *to proceed from such a point. (Also see the related discussion in the penultimate paragraph of* [23]*.)*

We then ran the MATLAB code on all but four of those problems from [12] for which the initial point provided in [12] satisfies all inequality constraints. (While a phase I-type scheme could be used on the other problems—see Remark 3—testing such a scheme is outside the main scope of this paper.) Problems 67, 68, 69, and 87 were left out: the first one because it is irregular,[9] the next two because of numerical difficulties in connection with the use of Chebyshev approximations in function evaluations, and the last one because the objective function in that problem is nonsmooth. In problems 31, 35, 44, 55, 71, and 86, the given x_0 is stationary for problem (P) and in problem 74, the given x_0 is stationary for (P_ρ) for every ρ. Results on the remaining 63 problems are reported in Table 5.1. The first column in the table gives the problem number from [12], the second column the total number of iterations, the third column the final value $\bar{\rho}$ of the penalty parameter, and the last column the final value of the objective function.

TABLE 5.1
Results on test problems from [12].

Prob.	#Itr	$\bar{\rho}$	f_{final}	Prob.	#Itr	$\bar{\rho}$	f_{final}
HS1	24	1	6.5782e-27	HS51	8	4	0.0000e+00
HS3	4	1	8.5023e-09	HS52	4	8	5.3266e+00
HS4	4	1	2.6667e+00	HS53	5	8	4.0930e+00
HS5	6	1	-1.9132e+00	HS54	23	4	-1.6292e-54
HS6	7	2	0.0000e+00	HS56	12	4	-3.4560e+00
HS7	9	2	-1.7321e+00	HS57	15	1	2.8460e-02
HS8	14	1	-1.0000e+00	HS60	7	1	3.2568e-02
HS9	10	1	-5.0000e-01	HS61	44	128	-1.4365e+02
HS12	5	1	-3.0000e+01	HS62	5	1	-2.6273e+04
HS24	14	1	-1.0000e+00	HS63	5	2	9.6172e+02
HS25	62	1	1.8185e-16	HS66	1000†	1	5.1817e-01
HS26	19	2	2.8430e-12	HS70	22	1	1.7981e-01
HS27	14	32	4.0000e-02	HS73	16	1	2.9894e+01
HS28	6	1	0.0000e+00	HS75	28	16	5.1744e+03
HS29	8	1	-2.2627e+01	HS77	13	1	2.4151e-01
HS30	7	1	1.0000e+00	HS78	4	4	-2.9197e+00
HS32	24	4	1.0000e+00	HS79	7	2	7.8777e-02
HS33	29	1	-4.5858e+00	HS80	6	2	5.3950e-02
HS34	30	1	-8.3403e-01	HS81	9	8	5.3950e-02
HS36	10	1	-3.3000e+03	HS84	30	1	-5.2803e+06
HS37	7	1	-3.4560e+03	HS85	296‡	1	-2.2156e+00
HS38	37	1	3.1594e-24	HS93	12	1	1.3508e+02
HS39	19	4	-1.0000e+00	HS99	8	2	-8.3108e+08
HS40	4	2	-2.5000e-01	HS100	9	1	6.8063e+02
HS42	6	4	1.3858e+01	HS107	1000†	8192	5.0545e+03
HS43	9	1	-4.4000e+01	HS110	6	1	-4.5778e+01
HS46	25	2	6.6616e-12	HS111	1000†	1	-4.7760e+01
HS47	25	16	8.0322e-14	HS112	11	1	-4.7761e+01
HS48	6	4	0.0000e+00	HS113	10	1	2.4306e+01
HS49	69	64	3.5161e-12	HS114	39	256	-1.7688e+03
HS50	11	512	4.0725e-17	HS117	25	1	3.2349e+01

On three of the problems (66, 107, and 111) our stopping criterion was not met after 1000 iterations. However, in all three cases the final objective value was equal, with three or more figures of accuracy, to the optimal value given in [12]. (Indeed, four

[9] It is termed irregular in [12]: computation of the cost and constraint functions involves an iterative procedure with variable number of iterations, rendering these functions discontinuous.

figures of accuracy were obtained on problems 66 and 111, after 120 and 887 iterations, respectively; and three figures of accuracy were reached on problem 107 after 95 iterations.) A large number of iterations was needed on problem 85, on which the algorithm failed in the last iteration to produce an acceptable step size due to numerical difficulties. When the algorithm stopped, near $x = (704.41, 68.60, 102.90, 282.03, 37.46)^{\mathrm{T}}$, the value of $\|\Delta x_k^0\|$ was less than $2 \cdot 10^{-8}$, and the objective value obtained was lower than the (locally) optimal value given in [12]. Overall, comparison with published results obtained on the same problems with other interior-point methods suggests that Algorithm A has promise. In particular, on 39 of the 63 problems listed in Table 5.1, our results in terms of number of iterations are better than those reported in [22]. (On three other problems they are identical, and problem 67 is not listed in [22].) More extensive testing on larger size problems is in order for a more definite assessment of the value of the proposed approach. Such testing will require a more elaborate implementation of the algorithm.

6. Concluding remarks. An interior-point algorithm for the solution of general nonconvex constrained optimization problems has been proposed and analyzed. The algorithm involves a novel, simple exact penalty parameter updating rule. Global convergence as well as local superlinear and quadratic convergence have been proved under mild assumptions. In particular, it was pointed out that the proposed algorithm does not suffer a common pitfall recently pointed out in [23]. Promising preliminary numerical results were reported.

While the present paper focused on applying a version of the Mayne–Polak scheme to the algorithm of [17], there should be no major difficulty in similarly extending other feasible interior-point algorithms for inequality constrained problems to handle general constrained problems.

7. Appendix. We discuss the implications of substituting Assumption PTH-A6*, as stated in section 2, for Assumption A6 of [17]. For the reader's ease of reference, throughout this appendix, we use the notation of [17]; Assumption PTH-A6* then reads as follows.

Assumption PTH-A6*. *Given any index set K such that $\{x_k\}_{k\in K}$ is bounded, there exist $\sigma_1, \sigma_2 > 0$ such that, for all $k \in K$,*

$$\|H_k\| \leq \sigma_2$$

and

$$\left\langle d, \left(H_k - \sum_{i=1}^{m} \frac{\mu_{k,i}}{g_i(x_k)} \nabla g_i(x_k) \nabla g_i(x_k)^T \right) d \right\rangle \geq \sigma_1 \|d\|^2 \quad \forall d \in R^n.$$

First of all, under this weaker assumption, a stronger version of Lemma 3.1 of [17] is needed.

Lemma PTH-3.1*. *Let $x \in X$, let $\mu \in R^m$ be such that $\mu_i \geq 0$ for all i and $\mu_i > 0$ for all $i \in I(x)$, and let H be a symmetric matrix satisfying the condition*

$$\left\langle d, \left(H - \sum_{i \notin I(x)} \frac{\mu_i}{g_i(x)} \nabla g_i(x) \nabla g_i(x)^T \right) d \right\rangle > 0 \quad \forall d \in \mathcal{T}(x) \setminus \{0\}, \tag{7.1}$$

where

$$\mathcal{T}(x) = \{d \in \mathcal{R}^n : \langle \nabla g_i(x), d \rangle = 0 \ \forall i \in I(x)\}.$$

Then the matrix $F(x, H, \mu)$ defined by

$$F(x,H,\mu) = \begin{bmatrix} H & \nabla g_1(x) & \dots & \nabla g_m(x) \\ \mu_1 \nabla g_1(x)^{\mathrm{T}} & g_1(x) & & \bigcirc \\ \vdots & & \ddots & \\ \mu_m \nabla g_m(x)^{\mathrm{T}} & \bigcirc & & g_m(x) \end{bmatrix}$$

is nonsingular.

Proof. It is enough to show that the only solution (d, λ) of the homogeneous system

$$Hd + \sum_{i=1}^{m} \lambda_i \nabla g_i(x) = 0, \tag{7.2}$$

$$\mu_i \langle \nabla g_i(x), d\rangle + \lambda_i g_i(x) = 0, \quad i = 1, \dots, m, \tag{7.3}$$

is (0,0). Scalar multiplication of both sides of (7.2) by d yields

$$\langle d, Hd\rangle + \sum_{i=1}^{m} \lambda_i \langle \nabla g_i(x), d\rangle = 0. \tag{7.4}$$

On the other hand, it follows from (7.3) and the assumption on μ that

$$\langle \nabla g_i(x), d\rangle = 0 \quad \forall\ i \in I(x). \tag{7.5}$$

Now, from (7.4) and (7.5), we get

$$\langle d, Hd\rangle + \sum_{i \notin I(x)} \lambda_i \langle \nabla g_i(x), d\rangle = 0. \tag{7.6}$$

Solving (7.3) for λ_i (for $i \notin I(x)$) and substituting in (7.6) yields

$$\langle d, Hd\rangle - \sum_{i \notin I(x)} \langle \nabla g_i(x), d\rangle \frac{\mu_i}{g_i(x)} \langle \nabla g_i(x), d\rangle = 0.$$

In view of (7.5), it follows from (7.1) that $d = 0$. It then follows from (7.3) that $\lambda_i = 0$ for all $i \notin I(x)$. Assumption A5 of [17] together with (7.2) then implies that $(d, \lambda) = (0, 0)$. ∎

Next, the first inequality in equation (3.6) of [17] is unaffected. While the second inequality in that equation still holds as well, it is not of much value now that H_k is no longer assumed to be positive definite. However, we note that with S_k denoting the Schur complement of $G_k := \mathrm{diag}(g_i(x_k))$ in $F(x_k, H_k, \mu_k)$ (see page 794 of [17]), i.e.,

$$S_k := H_k - A_k G_k^{-1} M_k A_k^{\mathrm{T}},$$

where A_k and M_k are defined on page 808 in [17], we get

$$d_k^0 = -S_k^{-1} \nabla f(x_k)$$

yielding

$$\langle \nabla f(x_k), d_k^0 \rangle = -\langle d_k^0, S_k d_k^0 \rangle \leq -\sigma_1 \|d_k^0\|^2, \tag{7.7}$$

where we have invoked Assumption PTH-A6*. Where equation (3.6) (of [17]) is used in the analysis of [17], equation (7.7) must sometimes be used instead.

Propositions 3.3 and 3.4 of [17] then readily follow. The only remaining notable issue is that a stronger version of Lemma 3.5 of [17] is needed, as follows.

Lemma PTH-3.5*. *Let K be an infinite index set such that, for some x^* and μ^*,*

$$\lim_{\substack{k\to\infty \\ k\in K}} x_k = x^* \text{ and } \lim_{\substack{k\to\infty \\ k\in K}} \mu_k = \mu^*.$$

Suppose, moreover, that $\mu_i^* > 0$ if $g_i(x^*) = 0$. Then, given any accumulation point H^* of $\{H_k\}_{k\in K}$, $F(x^*, H^*, \mu^*)$ is nonsingular. Moreover, there exists C such that, for all $k \in K$,

$$\|d_k - d_k^0\| \leq C \|d_k^0\|^\nu.$$

Proof. Let $K' \subseteq K$ be an infinite index set such that $H_k \to H^*$ as $k \to \infty$, $k \in K'$. We first show that (x^*, H^*, μ^*) satisfies the assumptions of Lemma PTH-3.1*. Thus let $v \neq 0$ be such that

$$\langle \nabla g_i(x^*), v \rangle = 0 \quad \forall i \in I(x^*).$$

In view of our linear independence assumption, there exists[10] a sequence $\{v_k\}_{k\in K'}$ converging to v on K' such that for all $k \in K'$

$$\langle \nabla g_i(x_k), v_k \rangle = 0 \quad \forall i \in I(x^*).$$

It then follows from Assumption PTH-A6* (by adding zero terms) that for all $k \in K'$

$$\left\langle v_k, \left(H_k - \sum_{i\notin I(x^*)} \frac{\mu_{k,i}}{g_i(x_k)} \nabla g_i(x_k) \nabla g_i(x_k)^T \right) v_k \right\rangle \geq \sigma_1 \|v_k\|^2.$$

Letting $k \to \infty$, $k \in K'$ shows that

$$\left\langle v, \left(H^* - \sum_{i\notin I(x^*)} \frac{\mu_i^*}{g_i(x^*)} \nabla g_i(x^*) \nabla g_i(x^*)^T \right) v \right\rangle \geq \sigma_1 \|v\|^2 > 0.$$

Thus the assumptions of Lemma PTH-3.1* are satisfied. It follows that $F(x^*, H^*, \mu^*)$ is nonsingular. Since $F(x_k, H_k, \mu_k)$ is nonsingular for all k, boundedness of $\{H_k\}$ and our continuity assumptions imply that $F(x_k, H_k, \mu_k)^{-1}$ is uniformly bounded on K. The remainder of the proof is as in [17]. ∎

With these strengthened results, the remainder of the analysis in [17] is essentially unaffected by the weakening of the assumption on H_k. Specifically, Lemma 3.6 of [17] (where the "old" Assumption A6 of [17] is invoked) still follows, using the stronger Lemmas PTH-3.1* and PTH-3.5*. While the "boundedness of H_k" portion

[10] For instance, v_k can be selected to be the orthogonal projection of v on the orthogonal complement of the span of the $\nabla g_i(x_k)$'s.

of Assumption A6 is used at many other places in the analysis of [17], the "positive definiteness" portion of that assumption (which is the only portion that is relaxed in Assumption PTH-A6*) is not used anywhere else. The strengthened Lemmas PTH-3.1* and PTH-3.5* are needed in the proof of Lemma 4.1 of [17]: Lemma 3.1 of [17] is implicitly used in the last sentence of that proof to conclude that the limit system (4.1)–(4.2) of [17] is invertible.

Finally, Lemma 4.4 of [17] still holds under the milder Assumption PTH-A6* (and so do Proposition 4.5 of [17] and Theorem 4.6 of [17]), but again the strengthened Lemmas PTH-3.1* and PTH-3.5* are needed in its proof. In particular, for k large enough, the second order sufficiency condition still holds at the solution of (LS3) of [17], and thus solving (LS3) is still equivalent to solving the stated linear system of equations (in the proof of Lemma 4.4). (The notation $\|d\|^2_{H_k}$ used in (LS3) is now inappropriate though, and should be replaced with $\langle d, H_k d\rangle$.) Furthermore, it follows from Lemma PTH-3.5* and the fact that $\mu_{k,i}$ tends to zero for $i \notin I(x^*)$ that, for k large enough, this linear system still has a unique solution; i.e., (LS3) still has a well-defined (unique) minimizer.

Acknowledgments. The authors wish to thank Jean-Charles Gilbert and Paul Armand for discussions in connection with an early version of Assumption 3(ii); as a result the current Assumption 3(ii) is significantly less restrictive. They also wish to thank Bill Woessner for his help with the numerical experiments, as well as two anonymous referees and the associate editor for their feedback that helped improve the paper.

REFERENCES

[1] S. Bakhtiari and A.L. Tits, *A simple primal-dual feasible interior-point method for nonlinear programming with monotone descent*, Comput. Optim. Appl., 25 (2003), pp. 17–38.

[2] J.V. Burke, *A robust trust region method for constrained nonlinear programming problems*, SIAM J. Optim., 2 (1992), pp. 325–347.

[3] R.H. Byrd, J.C. Gilbert, and J. Nocedal, *A trust region method based on interior point techniques for nonlinear programming*, Math. Program., 89 (2000), pp. 149–185.

[4] R.H. Byrd, M.E. Hribar, and J. Nocedal, *An interior point algorithm for large-scale nonlinear programming*, SIAM J. Optim., 9 (1999), pp. 877–900.

[5] A.S. El-Bakry, R.A. Tapia, T. Tsuchiya, and Y. Zhang, *On the formulation and theory of the Newton interior-point method for nonlinear programming*, J. Optim. Theory Appl., 89 (1996), pp. 507–541.

[6] A.V. Fiacco and G.P. McCormick, *Nonlinear Programming: Sequential Unconstrained Minimization Techniques*, Wiley, New York, 1968.

[7] A. Forsgren and P.E. Gill, *Primal-dual interior methods for nonconvex nonlinear programming*, SIAM J. Optim., 8 (1998), pp. 1132–1152.

[8] D.M. Gay, M.L. Overton, and M.H. Wright, *A primal-dual interior method for nonconvex nonlinear programming*, in Advances in Nonlinear Programming, Y. Yuan, ed., Kluwer Academic Publishers, Dordrecht, The Netherlands, 1998, pp. 31–56.

[9] N.I.M. Gould, D. Orban, and Ph.L. Toint, *CUTEr (and SifDec), a Constrained and Unconstrained Testing Environment, Revisited*, Technical report TR/PA/01/04, CERFACS, Toulouse, France, 2001.

[10] J.N. Herskovits, *Développement d'une Méthode Numérique pour l'Optimization Non-Linéaire*, Ph.D. thesis, Université Paris IX - Dauphine, Paris, France, 1982.

[11] J.N. Herskovits, *A two-stage feasible directions algorithm for nonlinear constrained optimization*, Math. Program., 36 (1986), pp. 19–38.

[12] W. Hock and K. Schittkowski, *Test Examples for Nonlinear Programming Codes*, Lecture Notes in Econom. and Math. Systems 187, Springer-Verlag, Berlin, New York, 1981.

[13] C.T. Lawrence and A.L. Tits, *Nonlinear equality constraints in feasible sequential quadratic programming*, Optim. Methods Softw., 6 (1996), pp. 265–282.

[14] N. Maratos, *Exact Penalty Function Algorithms for Finite Dimensional and Optimization Problems*, Ph.D. thesis, Imperial College of Science and Technology, University of London, London, UK, 1978.

[15] D.Q. Mayne and E. Polak, *Feasible direction algorithms for optimization problems with equality and inequality constraints*, Math. Program., 11 (1976), pp. 67–80.

[16] D.Q. Mayne and E. Polak, *A superlinearly convergent algorithm for constrained optimization problems*, Math. Program. Stud., 16 (1982), pp. 45–61.

[17] E.R. Panier, A.L. Tits, and J.N. Herskovits, *A QP-free, globally convergent, locally superlinearly convergent algorithm for inequality constrained optimization*, SIAM J. Control Optim., 26 (1988), pp. 788–811.

[18] E. Polak and A.L. Tits, *On globally stabilized quasi-Newton methods for inequality constrained optimization problems*, in Proceedings of the 10th IFIP Conference on System Modeling and Optimization, New York, 1981, Lecture Notes in Control and Inform. Sci. 38, R.F. Drenick and E.F. Kozin, eds., Springer-Verlag, Berlin, 1982, pp. 539–547.

[19] M. Sahba, *Globally convergent algorithm for nonlinearly constrained optimization problems*, J. Optim. Theory Appl., 52 (1987), pp. 291–309.

[20] S. Segenreich, N. Zouain, and J.N. Herskovits, *An optimality criteria method based on slack variables concept for large structural optimization*, in Proceedings of the Symposium on Applications of Computer Methods in Engineering, Los Angeles, 1977, pp. 563–572.

[21] A.L. Tits and J.L. Zhou, *A simple, quadratically convergent algorithm for linear and convex quadratic programming*, in Large Scale Optimization: State of the Art, W.W. Hager, D.W. Hearn, and P.M. Pardalos, eds., Kluwer Academic Publishers, Dordrecht, The Netherlands, 1994, pp. 411–427.

[22] R.J. Vanderbei and D.F. Shanno, *An interior-point algorithm for nonconvex nonlinear programming*, Comput. Optim. Appl., 13 (1999), pp. 231–252.

[23] A. Wächter and L.T. Biegler, *Failure of global convergence for a class of interior point methods for nonlinear programming*, Math. Program., 88 (2000), pp. 565–574.

[24] A. Wächter and L.T. Biegler, *Global and Local Convergence of Line Search Filter Methods for Nonlinear Programming*, Technical report CAPD B-01-09, Carnegie Mellon University, Pittsburgh, PA, 2001.

[25] H. Yabe and H. Yamashita, *Q-superlinear convergence of primal-dual interior point quasi-newton methods for constrained optimization*, J. Oper. Res. Soc. Japan, 40 (1997), pp. 415–436.

[26] H. Yamashita, *A globally convergent primal-dual interior point method for constrained optimization*, Optim. Methods Softw., 10 (1998), pp. 443–469.

[27] H. Yamashita, H. Yabe, and T. Tanabe, *A Globally and Superlinearly Convergent Primal-Dual Interior Point Trust Region Method for Large Scale Constrained Optimization*, Technical report, Mathematical Systems, Inc., Tokyo, Japan, 1998.

[28] Y. Yuan, *On the convergence of a new trust region algorithm*, Numer. Math., 70 (1995), pp. 515–539.

SIAM J. OPTIM.
Vol. 14, No. 1, pp. 200–222

A DECOMPOSITION METHOD BASED ON SQP FOR A CLASS OF MULTISTAGE STOCHASTIC NONLINEAR PROGRAMS*

XINWEI LIU† AND GONGYUN ZHAO‡

Abstract. Multistage stochastic programming problems arise in many practical situations, such as production and manpower planning, portfolio selections, and so on. In general, the deterministic equivalents of these problems can be very large and may not be solvable directly by general-purpose optimization approaches. Sequential quadratic programming (SQP) methods are very effective for solving medium-size nonlinear programming. By using the scenario analysis technique, a decomposition method based on SQP for solving a class of multistage stochastic nonlinear programs is proposed, which generates the search direction by solving parallelly a set of quadratic programming subproblems with much less size than the original problem at each iteration. Conjugate gradient methods can be introduced to derive the estimates of the dual multiplier associated with the nonanticipativity constraints. By selecting the step-size to reduce an exact penalty function sufficiently, the algorithm terminates finitely at an approximate optimal solution to the problem with any desirable accuracy. Some preliminary numerical results are reported.

Key words. multistage stochastic nonlinear programs, sequential quadratic programming, scenario analysis, decomposition

AMS subject classifications. 90C15, 90C06, 49M27, 90C55

DOI. 10.1137/S1052623402361447

1. Introduction.

Stochastic programming studies optimization problems with uncertain data. Multistage stochastic programming problems arise in many practical situations, such as production and manpower planning, portfolio selections, and so on. Consider the following multistage stochastic program with recourse:

$$\min_{x\in X}\ \hat{c}_0(x) + E_{\xi_1} Q_1(x,\xi_1), \tag{1.1}$$

where $X = \{x | c_0(x) = 0\} \subseteq \Re^{n_0}$, the recourse function is

$$Q_1(x,\hat{\xi}_1) = \min_{y_1}\ q_1(x,y_1,\hat{\xi}_1) + E_{\xi_2} Q_2(x,y_1,\hat{\xi}_1,\xi_2) \text{ subject to (s.t.) } c_1(x,y_1,\hat{\xi}_1) = 0, \tag{1.2}$$

and for $t = 2,\dots,T-1$, recursively we have

$$\begin{aligned} Q_t(x,y_1,\dots,y_{t-1},\hat{\xi}_1,\dots,\hat{\xi}_t) = \min_{y_t}\ & q_t(x,y_1,\dots,y_{t-1},y_t,\hat{\xi}_1,\dots,\hat{\xi}_t) \\ & + E_{\xi_{t+1}} Q_{t+1}(x,y_1,\dots,y_t,\hat{\xi}_1,\dots,\hat{\xi}_y,\xi_{t+1}) \qquad (1.3) \\ \text{s.t. } & c_t(x,y_1,\dots,y_t,\hat{\xi}_1,\dots,\hat{\xi}_t) = 0, \qquad (1.4) \end{aligned}$$

*Received by the editors January 23, 2002; accepted for publication (in revised form) March 6, 2003; published electronically July 18, 2003. This research was partially supported by grant R-146-000-032-112 of the National University of Singapore.

http://www.siam.org/journals/siopt/14-1/36144.html

†Singapore-MIT Alliance, National University of Singapore, E4-04-10, 4 Engineering Drive 3, Singapore 117576 (smaliuxw@nus.edu.sg). This work was done when this author was a research fellow of the Department of Mathematics of the National University of Singapore. This author is on leave from the Hebei University of Technology, Tianjin, China. His research was partially supported by a Hebei doctoral fund.

‡Department of Mathematics, National University of Singapore, Singapore 119260 (matzgy@nus.edu.sg).

$Q_T = 0$. Vector $x \in \Re^{n_0}$ is the deterministic vector, and $\hat{\xi}_i$ is the realization of the random vector ξ_i. Vector $y_i \in \Re^{n_i}$ is the decision vector in the ith stage, which is generated recursively by $x, y_1, \ldots, y_{i-1}$ and $\hat{\xi}_1, \ldots, \hat{\xi}_i$; hence y_i is a function on $(x, y_1, \ldots, y_{i-1}, \hat{\xi}_1, \ldots, \hat{\xi}_i)$ actually. Functions $\hat{c}_0$ and c_0 are real-valued functions on $\Re^{n_0}$, and c_t is random since it is related to $\hat{\xi}_1, \ldots, \hat{\xi}_t$. For the discrete random vector $\xi = (\xi_1, \ldots, \xi_{T-1})$, if c_t has finite realizations $c_{ti}(i = 1, \ldots, S_t)$, then all these c_{ti} form the constraint functions on stage t. The details on the formulation of multistage stochastic programs can be found, e.g., in Kall and Wallace [17].

Scenario analysis was introduced to deal with multistage stochastic programs by Rockafellar and Wets in 1987, by which the program is specified into a finite number of scenarios for the considered time period. Let $\xi = (\xi_1, \ldots, \xi_{T-1})$, and assume that (Ω, Θ, P) is the associated probability space. Suppose that we have S scenarios $\xi^{(s)} = (\xi_1^{(s)}, \xi_2^{(s)}, \ldots, \xi_{T-1}^{(s)})$, $s = 1, \ldots, S$, and have a fixed and known probability distribution $\{(\xi^{(s)}, p_s) | s = 1, 2, \ldots, S\}$. Then program (1.1)–(1.4) can be reformulated as the following nonlinear programming problem:

$$\min \sum_{s=1}^{S} f_s(z^{(s)}) \tag{1.5}$$

$$\text{s.t. } h_s(z^{(s)}) = 0, \ s = 1, 2, \ldots, S, \tag{1.6}$$

$$\sum_{s=1}^{S} A_s z^{(s)} = 0, \tag{1.7}$$

where $z^{(s)} = (x^{(s)}, y_1^{(s)}, \ldots, y_{T-1}^{(s)}) \in \Re^n$, $n = \sum_{i=0}^{T-1} n_i$,

$$f_s(z^{(s)}) = p_s \left(\hat{c}_0(x^{(s)}) + \sum_{t=1}^{T-1} q_t(x^{(s)}, y_1^{(s)}, \ldots, y_t^{(s)}, \xi_1^{(s)}, \ldots, \xi_t^{(s)}) \right), \tag{1.8}$$

$$\begin{aligned} h_s(z^{(s)}) = (&c_0(x^{(s)}), c_1(x^{(s)}, y_1^{(s)}, \xi_1^{(s)}), \ldots, \\ &c_{T-1}(x^{(s)}, y_1^{(s)}, \ldots, y_{T-1}^{(s)}, \xi_1^{(s)}, \ldots, \xi_{T-1}^{(s)})). \end{aligned} \tag{1.9}$$

Constraints (1.7) are the so-called nonanticipativity constraints, which reflect the fact that scenarios sharing a common history up to any moment of time must have a common decision at that moment. Readers can refer to Rockafellar and Wets [25] for more details on this reformulation.

In this paper, we consider solving the program (1.5)–(1.7). It is assumed that $f_s : \Re^n \to \Re$ and $h_s : \Re^n \to \Re^m$ are twice continuously differentiable functions, $z^{(s)} \in \Re^n$, $h_{si} : \Re^n \to \Re (i = 1, \ldots, m)$ and $h_s(z^{(s)}) = (h_{s1}(z^{(s)}), \ldots, h_{sm}(z^{(s)}))$. Matrix $A_s \in \Re^{m_0 \times n} (s = 1, \ldots, S)$, $A = [A_1 \ A_2 \ \cdots \ A_S]$ is an $m_0 \times nS$ matrix with full row rank and has a special structure, which will be further identified for the concrete examples in section 5.

When the scenario number S is large, program (1.5)–(1.7) can be very large and may not be solvable directly by general-purpose optimization approaches. One of the important selections for solving the stochastic programming is to develop efficient decomposition techniques; see, e.g., Ruszczyński [28, 29]. Moreover, the parallelization of computers provides the feasibility for implementing the decomposition methods.

There are many references contributed to the decomposition methods in linear and nonlinear programming in the literature; e.g., see Lasdon [18], Feinberg [10], Han [14], Ruszczyński [26, 27], et al. Most of them are related to the well-known

decomposition principle of Dantzig and Wolfe [9], and to the duality theory based on the Lagrangian and augmented Lagrangian functions.

The L-shaped decomposition method is efficient for solving multistage stochastic linear programs. In each cycle, sets of feasibility cuts and optimality cuts are generated recursively, and a sequence of decreasing feasible regions is derived. Some other methods for multistage stochastic linear programs can be found, e.g., in Birge [2], Birge and Louveaux [3], and the references therein. More recently, Zhao [33, 34] proposed logarithmic barrier methods for solving multistage stochastic linear programs. Since all these methods are based on the special structures and properties of stochastic linear programs, it is difficult to generalize them to solve the stochastic nonlinear programs.

Based on scenario analysis technique, Rockafellar and Wets [25] proposed the progressive hedging algorithm (PHA) for multistage stochastic programming, which is an iterative method. Mulvey and Vladimirou [20] applied the PHA to the stochastic generalized networks and achieved satisfactory numerical results. The additional works on the PHA include Chun and Robinson [8], Helgason and Wallace [16], et al. One of the difficulties in implementing the PHA is the selection of a suitable penalty parameter. Chun and Robinson [8] showed that the PHA is not the best candidate for the loosely coupled scenario analysis problems, and the bundle-based decomposition method in Robinson [24] is more competitive than the PHA. A new iterative method based on scenario analysis was proposed recently by Zhao [35], which relaxes the nonanticipativity constraints by the Lagrangian dual approach and combines with the use of logarithmic barrier methods.

Sequential quadratic programming (SQP) is an iterative method and is very effective for solving medium-size nonlinear programming; e.g., see Fukushima [12], Powell and Yuan [22], Boggs and Tolle [6], Liu and Yuan [19], et al. Recently, it has been applied to solve the complementarity problems, the variational inequality problems, and the nonsmooth problems; e.g., see Fukushima [13], Pang, Han, and Rangaraj [21], Han, Pang, and Rangaraj [15], and Qi [23]. In this paper, SQP is applied to the program (1.5)–(1.7). By combining with the Lagrangian dual approach, we present a decomposition method based on SQP. Conjugate gradient methods can be introduced to derive the estimates of the dual multiplier associated with the nonanticipativity constraints (1.7), and the search direction is generated by solving parallelly a set of quadratic programming (QP) subproblems with much less size than the original problem at each iteration. The global convergence of the algorithm is analyzed. The algorithm is also used to solve some stochastic nonlinear programs, and the preliminary numerical results are reported. Our method can be taken as one of the examples of SQP for solving large-scale structural nonlinear optimization.

Our method has some similarity to the very recently published papers [4, 5, 31] in that all of them need to solve some kinds of QP subproblems or the Karush–Kuhn–Tucker (KKT) systems equivalently. Steinbach [31] and Blomvall and Lindberg [4, 5] considered the solution of multistage stochastic convex programming with linear and bound constraints by interior point methods. Some decomposition techniques based on KKT systems and the approximate QP of the barrier subproblem, which were induced by scenario tree formulation, were developed so that the sparsity in local constraints was not destroyed. However, the method in this paper is to exploit the special structure of the scenario formulation with explicit nonanticipativity constraints. Some induced difficulties including the rank deficiency of the Jacobian should be coped with. A different decomposition method for QP subproblems is presented, by which

QP splits apart into separate subproblems according to the scenarios.

This paper is organized as follows. In section 2, some discussions for the algorithm are presented, including the decomposition of a large-scale QP and the boundedness of the dual multiplier corresponding to the nonanticipativity constraints. The algorithm is then proposed in section 3. In section 4 we analyze the global convergence of the algorithm. We consider the extension of the algorithm to the stochastic nonlinear programs with inequality constraints in section 5. Some preliminary results are reported in section 6.

Although all notations can be identified easily from the context, we list some of the notations used in the paper for the reader's convenience: A letter with the superscript (s) or subscript s is associated with the sth scenario; for example, $z^{(s)}$ is the decision vector and f_s is the function associated with the sth scenario. A letter with the subscript k corresponds to the iteration k. For vectors $z^{(1)}, \ldots, z^{(S)}$, $(z^{(1)}, \ldots, z^{(S)}) = [z^{(1)\top} \ \cdots \ z^{(S)\top}]^\top$. We also use symbols

$$\nabla_s \cdot = \frac{\partial \cdot}{\partial z^{(s)}}, \ \ \nabla \cdot (z) = \frac{\partial \cdot}{\partial z}, \ \ \nabla \cdot (\mu) = \frac{\partial \cdot}{\partial \mu}. \tag{1.10}$$

Thus, $\nabla \cdot (z) = [\nabla_1 \cdot^\top \ \nabla_2 \cdot^\top \ \cdots \ \nabla_S \cdot^\top]^\top$. For simplicity of statement, when we use some functions, we may omit the corresponding variables; for example, $\nabla_s h_s$ and h_{ks} represent $\nabla_s h_s(z^{(s)})$ and $h_s(z_k^{(s)})$, respectively. If it is not specified, all vectors throughout the paper are column vectors, and norm $\|\cdot\|$ is the Euclidean norm.

2. Basic discussions. Suppose that $\lambda^{(s)} \in \Re^m (s = 1, \ldots, S)$ and $\mu \in \Re^{m_0}$ are the multipliers corresponding to the constraints (1.6) and (1.7), respectively. Let $\lambda = (\lambda^{(1)}, \ldots, \lambda^{(S)})$. The Lagrangian of the program (1.5)–(1.7) is

$$L(z, \lambda, \mu) = \sum_{s=1}^{S} L_s(z^{(s)}, \lambda^{(s)}, \mu), \tag{2.1}$$

where $L_s(z^{(s)}, \lambda^{(s)}, \mu) = f_s(z^{(s)}) + \lambda^{(s)\top} h_s(z^{(s)}) + \mu^\top A_s z^{(s)}$. Thus, the Lagrangian Hessian has a block diagonal structure with the s-diagonal block being

$$\nabla_s^2 L_s(z^{(s)}, \lambda^{(s)}, \mu) = \nabla_s^2 f_s(z^{(s)}) + \sum_{i=1}^{m} \lambda_i^{(s)} \nabla_s^2 h_{si}(z^{(s)}). \tag{2.2}$$

SQP was developed by Wilson, Han, and Powell for solving nonlinear programming problems. At current iteration z, a QP is solved to generate the search direction d. The new iteration point z_+ is derived by the formula

$$z_+ = z + \alpha d, \tag{2.3}$$

where $\alpha \in (0, 1]$ is the step-size decided by some line search procedure. SQP for the program (1.5)–(1.7) needs to solve the following QP at each iteration:

$$\min \sum_{s=1}^{S} \left(\nabla_s f_s^\top d^{(s)} + \frac{1}{2} d^{(s)\top} H_s d^{(s)} \right) \tag{2.4}$$

$$\text{s.t. } h_s + \nabla_s h_s^\top d^{(s)} = 0, \ \ s = 1, \ldots, S, \tag{2.5}$$

$$\sum_{s=1}^{S} A_s (z^{(s)} + d^{(s)}) = 0, \tag{2.6}$$

where H_s is an approximation to the Lagrangian Hessian block $\nabla_s^2 L_s(z^{(s)}, \lambda^{(s)}, \mu)$ for $s = 1, \ldots, S$ and is supposed to be positive definite.

It is well known that the bottleneck in applying SQP to the large-scale nonlinear programs is in the effective solution of all QPs. Evidently, problem (2.4)–(2.6) has the same size as the original problem (1.5)–(1.7). When the number S of scenarios is large, the number nS of variables and the number $(mS + m_0)$ of constraints are large correspondingly, which may bring about severe difficulties in solving QP (2.4)–(2.6) directly due to the necessity of huge memory (at least $O(S^3)$). These difficulties may also happen in applying direct SQP to the scenario tree formulation (see [4, 5, 31]).

To overcome these kinds of difficulties, we use a Lagrangian dual to exploit the separable structure of (2.4)–(2.6). What is more special and subtle, constraints (2.5) and (2.6) may be inconsistent for any given z, since the coefficient vectors of (2.5)–(2.6) may be linearly dependent even if the coefficient vectors of (2.5) and the coefficient vectors of (2.6), respectively, are linearly independent. The problem becomes even more involved when the Lagrangian dual approach is used.

We relax the constraint (2.6) and obtain the Lagrangian dual of (2.4)–(2.6) as follows:

$$\max_{\mu} \quad \varphi(\mu), \tag{2.7}$$

where

$$\varphi(\mu) = \min_{(d^{(1)},\ldots,d^{(S)})} \sum_{s=1}^{S} \left(\nabla_s f_s^\top d^{(s)} + \frac{1}{2} d^{(s)\top} H_s d^{(s)} + \mu^\top A_s (z^{(s)} + d^{(s)}) \right) \tag{2.8}$$

$$\text{s.t.} \quad h_s + \nabla_s h_s^\top d^{(s)} = 0, \ \ s = 1, \ldots, S. \tag{2.9}$$

It is easy to verify that $\varphi(\mu)$ is a concave function. We have the following properties on $\varphi(\mu)$.

LEMMA 2.1. *If, for $s = 1, \ldots, S$, $\nabla_s h_s$ have full column rank and H_s are positive definite, then $\varphi(\mu)$ is continuously differentiable, and for any $\mu \in \Re^{m_0}$ we have*

(i)

$$\nabla \varphi(\mu) = \sum_{s=1}^{S} A_s (z^{(s)} + d(\mu)^{(s)}), \tag{2.10}$$

where $d(\mu) = (d(\mu)^{(1)}, \ldots, d(\mu)^{(S)})$ is the unique solution of program (2.8)–(2.9)*;*

(ii)

$$\nabla^2 \varphi(\mu) = -\sum_{s=1}^{S} A_s (H_s^{-1} - H_s^{-1} \nabla_s h_s (\nabla_s h_s^\top H_s^{-1} \nabla_s h_s)^{-1} \nabla_s h_s^\top H_s^{-1}) A_s^\top. \tag{2.11}$$

Proof. Under the conditions of the lemma, for any given $\mu \in \Re^{m_0}$, problem (2.8)–(2.9) has the unique solution $d(\mu)$. By the KKT conditions of (2.8)–(2.9), there is a $\lambda(\mu) \in \Re^{mS}$ such that, for $s = 1, 2, \ldots, S$,

$$H_s d(\mu)^{(s)} + \nabla_s h_s \lambda(\mu)^{(s)} = -\nabla_s f_s - A_s^\top \mu, \tag{2.12}$$

$$\nabla_s h_s^\top d(\mu)^{(s)} = -h_s. \tag{2.13}$$

Since H_s is positive definite and $\nabla_s h_s$ has full column rank, the Jacobian of (2.12)–(2.13) is invertible. Thus, $d(\mu)$ is a linear function on μ. Hence, $\varphi(\mu)$ is differentiable by (2.8).

(i) For convenience of statement, let $f = (f_1, \ldots, f_S)$, $A = [A_1 \ A_2 \ \cdots \ A_S]$. Then, for any $\tilde{\mu}$,

$$\begin{aligned}\varphi(\tilde{\mu}) &\le \nabla f(z)^\top d(\mu) + \frac{1}{2} d(\mu)^\top H d(\mu) + \tilde{\mu}^\top A(z + d(\mu)) \\ &= \nabla f(z)^\top d(\mu) + \frac{1}{2} d(\mu)^\top H d(\mu) + \mu^\top A(z + d(\mu)) + (\tilde{\mu} - \mu)^\top A(z + d(\mu)) \\ &= \varphi(\mu) + (\tilde{\mu} - \mu)^\top A(z + d(\mu)), \qquad (2.14)\end{aligned}$$

which implies the result.

(ii) Differentiating (2.12)–(2.13) w.r.t. μ, and doing some calculations, we have

$$\nabla d(\mu)^{(s)} = (H_s^{-1} \nabla_s h_s (\nabla_s h_s^\top H_s^{-1} \nabla_s h_s)^{-1} \nabla_s h_s^\top H_s^{-1} - H_s^{-1}) A_s^\top. \qquad (2.15)$$

By (i),

$$\nabla^2 \varphi(\mu) = \sum_{s=1}^{S} A_s \nabla d(\mu)^{(s)}. \qquad (2.16)$$

Thus, the result follows from (2.15). □

It follows from Lemma 2.1(ii) that $\varphi(\mu)$ is a quadratic function, provided the conditions of Lemma 2.1 hold. Moreover, by Lemma 2.1(i),

$$\nabla \varphi(0) = \sum_{s=1}^{S} A_s (z^{(s)} + \hat{d}^{(s)}), \qquad (2.17)$$

where $(\hat{d}^{(1)}, \ldots, \hat{d}^{(S)})$ is the solution of the problem

$$\min_{(d^{(1)}, \ldots, d^{(S)})} \sum_{s=1}^{S} \left(\nabla_s f_s^\top d^{(s)} + \frac{1}{2} d^{(s)\top} H_s d^{(s)} \right) \qquad (2.18)$$

$$\text{s.t.} \quad h_s + \nabla_s h_s^\top d^{(s)} = 0, \ \ s = 1, \ldots, S. \qquad (2.19)$$

It is easy to note that problem (2.18)–(2.19) is a case of problem (2.8)–(2.9) with $\mu = 0$.

LEMMA 2.2. *Under the conditions of Lemma* 2.1*, suppose that there exists a* $\mu_+ \in \Re^{m_0}$ *which maximizes the concave quadratic function*

$$q(\mu) = \frac{1}{2} \mu^\top \nabla^2 \varphi(0) \mu + \nabla \varphi(0)^\top \mu, \qquad (2.20)$$

where $\nabla^2 \varphi(0)$ *is the same as* (2.11)*. Let* $d_+ = d(\mu_+)$*. Then*

(i) $\sum_{s=1}^{S} A_s (z^{(s)} + d_+^{(s)}) = 0$*;*

(ii) d_+ *is the optimal solution of program* (2.4)–(2.6)*.*

Proof. (i) It follows from Lemma 2.1(i) that

$$\sum_{s=1}^{S} A_s (z^{(s)} + d_+^{(s)}) = \nabla^2 \varphi(0) \mu_+ + \nabla \varphi(0). \qquad (2.21)$$

Thus, the result (i) follows immediately from the supposition of the lemma.

(ii) d_+ is the solution of (2.8)–(2.9); thus there is a $\lambda_+ \in \Re^{mS}$ such that (d_+, λ_+) satisfies the system of equations (2.12)–(2.13). Hence, the result follows from (i). ∎

By (2.20) and (2.8)–(2.9), we have

$$\varphi(\mu) = q(\mu) + \varphi(0), \tag{2.22}$$

where $\varphi(0) = \sum_{s=1}^{S}(\nabla_s f_s^\top \hat{d}^{(s)} + \frac{1}{2}\hat{d}^{(s)\top} H_s \hat{d}^{(s)})$.

The following result gives a sufficient condition on the boundedness of μ_+ in Lemma 2.2.

LEMMA 2.3. *Under the conditions of Lemma* 2.1, $\nabla^2\varphi(0)$ *given by* (2.11) *is negative semidefinite. Furthermore, if the matrix* $[\nabla h_1(z)\ \nabla h_2(z)\ \cdots\ \nabla h_S(z)\ A^\top]$ *has full column rank, then* $\nabla^2\varphi(\mu)$ *is negative definite.*

Proof. The first part of the lemma is straightforward.

With some linear algebraic manipulation, one can show that for any matrices V and U, if $[V\ U]$ has full column rank, then

$$\tilde{B} = U^\top(I - V(V^\top V)^{-1}V^\top)U \tag{2.23}$$

is positive definite.

Since H_s is symmetric positive definite, letting

$$U = \operatorname{diag}(H_1^{-\frac{1}{2}}, H_2^{-\frac{1}{2}}, \ldots, H_S^{-\frac{1}{2}})A^\top, \tag{2.24}$$

$$V = \operatorname{diag}(H_1^{-\frac{1}{2}}, H_2^{-\frac{1}{2}}, \ldots, H_S^{-\frac{1}{2}})\operatorname{diag}(\nabla_1 h_1, \nabla_2 h_2, \ldots, \nabla_S h_S), \tag{2.25}$$

it follows from the supposition that $[V\ U]$ has full column rank. Thus, the lemma follows from (2.11) and the positive definiteness of $\tilde{B}$. ∎

Unfortunately, by the observations in section 5, the condition in Lemma 2.3 that $[\nabla h_1(z)\ \nabla h_2(z)\ \cdots\ \nabla h_S(z)\ A^\top]$ has full column rank may not hold in many cases for multistage stochastic nonlinear programs. Thus, functions $\varphi(\mu)$ and $q(\mu)$ may not be strictly concave, in which case μ_+ may tend to infinity if the linearized constraints (2.5) and the constraints (2.6) are inconsistent. The next lemma shows that if the current iterate z meets the nonanticipativity constraints (1.7), then (2.5) and (2.6) are consistent under suitable conditions. Thus, the existence and boundedness of μ_+ are guaranteed by the duality theory of convex programming.

LEMMA 2.4. *Let* $W = \{z | Az = 0\}$ *and* $\bar{n}$ *be the dimension of* W. *For any* $\bar{z} \in \Re^{\bar{n}}$, *suppose that all* ∇c_{ti} $(t = 0, \ldots, T-1; i = 1, \ldots, S_t)$ *are linearly independent, where* c_{ti} *is defined in* (1.1)–(1.4). *Then*

(i) *the linearized constraints* (2.5) *and* (2.6) *are consistent at any* $z \in W$;

(ii) *there exists a* μ_+ *which maximizes* (2.20) *at* $z \in W$. *Moreover,* μ_+ *is bounded if* z *is bounded.*

Proof. (i) Let the columns of E comprise a basis of the subspace W, and let F be a matrix such that $FE = I$. Then $E \in \Re^{nS\times\bar{n}}$, $F \in \Re^{\bar{n}\times nS}$, $W = \{E\bar{z} | \bar{z} \in \Re^{\bar{n}}\}$. Define $\theta(z) = Fz$; it is easy to verify that θ is a bijection, and $\theta^{-1}(\bar{z}) = E\bar{z}$. Evidently, $\nabla\theta^{-1}(\bar{z})^\top = E$ for any $\bar{z} \in \Re^{\bar{n}}$.

Let $h(z) = (h_1(z^{(1)}), \ldots, h_S(z^{(S)}))$. For any $\bar{z} \in \Re^{\bar{n}}$, $\theta^{-1}(\bar{z}) \in W$. Define $\tilde{h}(\bar{z}) = h(\theta^{-1}(\bar{z}))$ for any $\bar{z} \in \Re^{\bar{n}}$. Then

$$\nabla\tilde{h}(\bar{z}) = \nabla\theta^{-1}(\bar{z})\nabla h(\theta^{-1}(\bar{z})) = E^\top \nabla h(\theta^{-1}(\bar{z})). \tag{2.26}$$

For any $z \in W$ let $\bar{z} = \theta(z) \in R^{\tilde{n}}$. We consider the equation

$$\tilde{h}(\bar{z}) + \nabla \tilde{h}(\bar{z})^\top \bar{d} = 0. \tag{2.27}$$

Let $c = (c_0, c_{11}, \ldots, c_{T-1S_{T-1}})$ be the collection of all constraints in (1.1)–(1.4). By (1.9) and the definitions of h and $\tilde{h}$, (2.27) is equivalent to

$$c(\bar{z}) + \nabla c(\bar{z})^\top \bar{d} = 0. \tag{2.28}$$

(Actually, (2.28) can be obtained by deleting those repetitious constraints in (2.27).) It follows from the assumption that $\nabla c(\bar{z})$ is of full column rank that there is a $\bar{d} \in \Re^{\bar{n}}$ such that (2.28) holds. Thus $\bar{d}$ is also a solution of (2.27).

Let $d = E\bar{d}$. By using (2.26), we can write (2.27) as

$$h(z) + \nabla h(z)^\top d = 0. \tag{2.29}$$

Moreover, $E\bar{d} \in W$ since $\bar{d} \in \Re^{\tilde{n}}$, and we have

$$Ad = 0. \tag{2.30}$$

The result follows directly from (2.29) and (2.30).

(ii) Since (2.7) is the dual of (2.4)–(2.6), and (2.5)–(2.6) has feasible solution, by the weak duality theorem, (2.7) is bounded. Furthermore, because $\varphi(\mu)$ is a convex quadratic function, the boundedness of the unconstrained problem (2.7) implies the existence of optimal solutions of (2.7). By (2.22), we have the result. □

It is easy to note that the condition in Lemma 2.4 is based on problem (1.1)–(1.4). This is natural since our aim is to solve problem (1.1)–(1.4). Under the condition of Lemma 2.4, by (1.9), we must have $\nabla_s h_s$ to be of full column rank for all $s = 1, \ldots, S$. However, the following example demonstrates the converse may not be true.

EXAMPLE 2.5. *Consider a two-stage problem with* $c_{ti}(t = 0, 1; S_0 = 1, S_1 = 2)$ *defined by*

$$x_1 - x_2 = 0, \tag{2.31}$$

$$x_1 + y^{(1)} - 1 = 0, \tag{2.32}$$

$$-x_1 + x_2 + y^{(1)} = 0, \tag{2.33}$$

$$x_1 + y^{(2)} - 2 = 0, \tag{2.34}$$

$$-x_1 + x_2 + y^{(2)} = 0, \tag{2.35}$$

where (x_1, x_2) *correspond to the first stage, and* $y^{(1)}$, $y^{(2)}$ *correspond to two different realizations, respectively. By notation* (1.9), *we have* $h_1(z^{(1)}) = 0$ *and* $h_2(z^{(2)}) = 0$, *which are below* (2.36)–(2.38) *and* (2.39)–(2.41), *respectively:*

$$z_1^{(1)} - z_2^{(1)} = 0, \tag{2.36}$$

$$z_1^{(1)} + z_3^{(1)} - 1 = 0, \tag{2.37}$$

$$-z_1^{(1)} + z_2^{(1)} + z_3^{(1)} = 0, \tag{2.38}$$

$$z_1^{(2)} - z_2^{(2)} = 0, \tag{2.39}$$

$$z_1^{(2)} + z_3^{(2)} - 2 = 0, \tag{2.40}$$

$$-z_1^{(2)} + z_2^{(2)} + z_3^{(2)} = 0. \tag{2.41}$$

It is easy to verify that the Jacobian of (2.36)–(2.41) *is of full column rank, but the Jacobian of* (2.31)–(2.35) *is not, which induces that the result of Lemma* 2.4 *does not hold.*

Fortunately, under the condition of Lemma 2.4, owing to the simplicity of the nonanticipativity constraints, we can easily select an initial point $z_0 \in W$. By Lemma 2.2, first maximizing (2.20) and then solving (2.8)–(2.9) to generate the search direction, we can ensure that the nonanticipativity constraints (1.7) hold at the new iterate. Thus, the algorithm can proceed to the new iteration. The details for the algorithm are stated in the next section.

3. The algorithm. Based on the previous discussions, we present our algorithm for solving problem (1.5)–(1.7) in this section. The efficiency of solving problem (1.5)–(1.7) relies on the efficiency of solving the QP (2.4)–(2.6). Here we first brief the idea of solving the QP

$$\min \sum_{s=1}^{S} \phi_s(d^{(s)}) \tag{3.1}$$

$$\text{s.t. } R_s d^{(s)} = r_s,\ s = 1, \ldots, S, \tag{3.2}$$

$$Ad = a, \tag{3.3}$$

where $\phi_s (s = 1, \ldots, S)$ are convex quadratic functions and $Ad = a$ comes from the nonanticipativity constraint. $A \in \Re^{m_0 \times nS}$, where m_0 is dependent on S. However, A has a simple structure.

S is typically very large for multistage stochastic programs. Thus, without making use of special structures, problem (3.1)–(3.3) is in general intractable.

Our decomposition-based method solves the Lagrangian dual of (3.1)–(3.3):

$$\max_{\mu} \varphi(\mu), \tag{3.4}$$

where

$$\varphi(\mu) = \sum_{s=1}^{S} \min_{d^{(s)}} \{\phi(d^{(s)}) + \mu^\top A_s d^{(s)} \mid R_s d^{(s)} = r_s\} - \mu^\top a. \tag{3.5}$$

The function φ is concave. Typically, finding the optimal dual solution μ^* needs infinitely many iterations. Because (3.1)–(3.3) is a quadratic program with linear equality constraints, it is observed that φ defined by (3.5) is a quadratic function (see Lemmas 2.1 and 2.2). Thus, we can find μ^* in one iteration. This can be seen in Notes 2 and 3 of the algorithm below. Step 1 of Algorithm 3.4 starts with an arbitrary μ^0, e.g., $\mu^0 = 0$, and solves (3.5), obtaining $d(\mu^0)$. We can compute $\nabla\varphi(\mu)$ and $\nabla^2\varphi(\mu^0)$ with $d(\mu^0)$ (see the formulas in Lemma 2.1), and represent

$$\varphi(\mu) = \varphi(\mu^0) + \nabla\varphi(\mu^0)^T(\mu - \mu^0) + (\mu - \mu^0)^T\nabla^2\varphi(\mu^0)(\mu - \mu^0). \tag{3.6}$$

Step 2 maximizes this quadratic function (an unconstrained convex quadratic program) obtaining the optimal dual solution μ^*. Step 3 solves (3.5) with $\mu = \mu^*$, obtaining the optimal solution d^* to the QP (3.1)–(3.3).

If we represent the original stochastic program with scenario tree formulation, then the QP employed by the SQP method looks as follows:

$$\min \psi(d) \tag{3.7}$$

$$\text{s.t. } Pd = p, \tag{3.8}$$

where ψ is a convex quadratic function and the size of P is $O(nS)$. Since it is a multistage stochastic program, P has a complicated structure.

One can use, e.g., the projected gradient method to solve problem (3.7)–(3.8), and the structure of P may be exploited in the process of projection on the subspace defined by P. This, however, involves very complicated algebraic operations which, besides heavy computational load, incur great difficulty in writing computer programs to implement them.

Our decomposition-based method solves $2S$ small-scale (m by n) quadratic programs (in Steps 1 and 3 of Algorithm 3.4) and an unconstrained convex quadratic program of dimension m_0 (in Step 2). Obviously, the computational load of performing each of these three steps is much lower than that directly solving the large-scale constrained quadratic program (3.7)–(3.8).

In summary:

(i) Exploiting the structure of the multistage stochastic program with a decomposition method is much easier than with the direct SQP. Thus, writing computer programs for our method is easier than for a direct SQP.

(ii) A decomposition-based method is more suitable to parallel computation.

(iii) The computational load and memory occupation of our method are not more than the direct SQP, even if we assume the direct SQP can properly exploit the sparsity and structure of coefficients of the program.

Before presenting our algorithm, we state some preliminary definitions and properties of the SQP method.

We define the l_1 exact penalty function

$$M(z,\rho)=\sum_{s=1}^{S}\left(f_s(z^{(s)})+\rho\|h_s(z^{(s)})\|_1\right) \tag{3.9}$$

as the merit function, where $\rho>0$ is the penalty parameter. The merit function is used to force the global convergence of the algorithm.

The following lemma is known so that we omit the proof.

LEMMA 3.1. *Let $N(z)=\sum_{s=1}^{S}\|h_s(z^{(s)})\|_1$. Suppose that $h_s(s=1,\dots,S)$ are continuously differentiable; then for all $z=(z^{(1)},\dots,z^{(S)})\in\Re^{nS}$ and $d=(d^{(1)},\dots,d^{(S)})\in\Re^{nS}$, the directional derivative of function N along d, defined by*

$$N'(z;d)=\lim_{\alpha\downarrow 0}\frac{N(z+\alpha d)-N(z)}{\alpha}, \tag{3.10}$$

exists, and we have

$$N'(z;d)\le\sum_{s=1}^{S}\left(\|h_s(z^{(s)})+\nabla_s h_s(z^{(s)})^{\top}d^{(s)}\|_1-\|h_s(z^{(s)})\|_1\right). \tag{3.11}$$

It follows from Lemma 3.1 that

$$\begin{aligned}M'(z,\rho;d)\le\sum_{s=1}^{S}[&\nabla_s f_s(z^{(s)})^{\top}d^{(s)}\\&+\rho(\|h_s(z^{(s)})+\nabla_s h_s(z^{(s)})^{\top}d^{(s)}\|_1-\|h_s(z^{(s)})\|_1)].\end{aligned} \tag{3.12}$$

The inequality (3.12) implies that if the right-hand side of (3.11) is negative, then the penalty parameter ρ can be increased such that $M'(z,\rho;d)<0$. Thus, the d such

that the right-hand side of (3.11) is negative can be a descent direction of the merit function $M(z,\rho)$ for large ρ. On the other hand, if the right-hand side of (3.11) is zero, then $M^{'}(z,\rho;d)<0$ if $\nabla_s f_s(z^{(s)})^\top d^{(s)}<0$ for $s=1,\ldots,S$.

The next result has little difference from the common one for general SQP methods (e.g., see Fukushima [12]).

LEMMA 3.2. *If* $(z^*,\lambda^*,\mu^*)\in\Re^{nS\times mS\times m_0}$ *is a KKT triple of problem* (1.5)–(1.7), $\rho^*>\max(\|\lambda^*\|_\infty;\|\mu^*\|_\infty)$, *then* $M^{'}(z^*,\rho^*;d)\geq 0$ *for all* $d\in\{d\in\Re^{nS}:Ad=0\}$.

An appropriate stopping criterion should be designed to guarantee that the algorithm terminates finitely at the "desirable" point of the problem. For convenience of statement, we need the following definition.

DEFINITION 3.3. *For any* $\epsilon>0$, *we call* $(z^{(1)},\ldots,z^{(S)})\in\Re^{nS}$ *an* ϵ*-optimal solution to the program* (1.5)–(1.7) *if there is a* $(\lambda^{(1)},\ldots,\lambda^{(S)})\in\Re^{mS}$ *and a* $\mu\in\Re^{m_0}$ *such that*

$$\max\{\|\nabla_s f_s(z^{(s)})+\nabla_s h_s(z^{(s)})\lambda^{(s)}+A_s^\top\mu\|,\ s=1,2,\ldots,S\}\leq\epsilon, \tag{3.13}$$

$$\max\{\|h_s(z^{(s)})\|_1,\ s=1,2,\ldots,S\}\leq\epsilon, \tag{3.14}$$

$$\left\|\sum_{s=1}^{S}A_s z^{(s)}\right\|_1\leq\epsilon. \tag{3.15}$$

If $\epsilon=0$, *then* $z=(z^{(1)},\ldots,z^{(S)})$ *is a KKT point of program* (1.5)–(1.7).

We select H_{ks}, for $s=1,2,\ldots,S$, to be the approximation of the Hessian of the Lagrangian

$$L_s(z^{(s)},\lambda^{(s)})=f_s(z^{(s)})+\lambda^{(s)\top}h_s(z^{(s)}) \tag{3.16}$$

at the iteration point $(z_k^{(s)},\lambda_k^{(s)})$ (which is the same as (2.2)), where $\lambda_k^{(s)}$ is an estimate of the multiplier associated with h_s:

$$B_k=\sum_{s=1}^{S}A_s(H_{ks}^{-1}-H_{ks}^{-1}\nabla_s h_{ks}(\nabla_s h_{ks}^\top H_{ks}^{-1}\nabla_s h_{ks})^{-1}\nabla_s h_{ks}^\top H_{ks}^{-1})A_s^\top, \tag{3.17}$$

which is the value of $-\nabla^2\varphi(0)$ at z_k.

ALGORITHM 3.4 (the decomposition method for program (1.5)–(1.7)).

Step 0. *Given* $(z_0^{(1)},\ldots,z_0^{(S)})\in\Re^{nS}$ *such that* (1.7) *holds,* $H_{0s}\in\Re^{n\times n}(s=1,2,\ldots,S)$, $\rho_0>0$, *and positive constants* $\delta<\frac{1}{2}$, ϵ, $\beta<1$, *and* $\sigma_0>0$. *Evaluate* $f_s(z_0^{(s)})$, $h_s(z_0^{(s)})$, $\nabla_s f_s(z_0^{(s)})$, $\nabla_s h_s(z_0^{(s)})$ *for* $s=1,2,\ldots,S$ *and* B_0. *Let* $\mu_0=0$, $k=0$;

Step 1. *For* $s=1,2,\ldots,S$, *solve the subproblems*

$$\min\ \phi_k^{(s)}(d)=\nabla_s f_s(z_k^{(s)})^\top d+\frac{1}{2}d^\top H_{ks}d \tag{3.18}$$

$$s.t.\ h_s(z_k^{(s)})+\nabla_s h_s(z_k^{(s)})^\top d=0. \tag{3.19}$$

Let $\hat{d}_k^{(s)}$, $s=1,\ldots,S$ *be the solutions. Set* $\nu_0=\sigma_k$, $\mu_0=\mu_k$, $j=0$;

Step 2. *Compute* $\mu_{j+1}=\mu_j+d_\mu$, *where* d_μ *is the solution to the unconstrained quadratic programming subproblem*

$$\max_{d_\mu}\ \bar{q}_k(d_\mu)=(A\hat{d}_k-B_k\mu_j)^\top d_\mu-\frac{1}{2}d_\mu^\top(B_k+\nu_j I)d_\mu \tag{3.20}$$

and can be derived by conjugate gradient methods. If

$$B_k\mu_{j+1} = A\hat{d}_k, \tag{3.21}$$

then $\mu_{k+1} = \mu_{j+1}$, $\sigma_{k+1} = \nu_j$ *and go to Step* 3*; Else compute* $r = Ared_j/Pred_j$, *where* $Ared_j = q_k(\mu_{j+1}) - q_k(\mu_j)$ *and* $Pred_j = \bar{q}_k(d_\mu) - \bar{q}_k(0)$. *The scalar* ν_j *is updated as follows:*

$$\nu_{j+1} = \begin{cases} 0.5\nu_j & \text{if } r > 0.75; \\ 4\nu_j & \text{if } r < 0.25; \\ \nu_j & \text{otherwise.} \end{cases} \tag{3.22}$$

Let $j = j + 1$ *and go to Step* 2*;*

Step 3. *For* $s = 1, 2, \ldots, S$, *solve the subproblems*

$$\min\ \psi_k^{(s)}(d) = (\nabla_s f_s(z_k^{(s)}) + A_s^\top \mu_{k+1})^\top d + \frac{1}{2} d^\top H_{ks} d \tag{3.23}$$

$$s.t.\ h_s(z_k^{(s)}) + \nabla_s h_s(z_k^{(s)})^\top d = 0 \tag{3.24}$$

to generate $d_k^{(s)} (s = 1, 2, \ldots, S)$*;*

Step 4. *Check if the stopping criterion*

$$\left| M(z_k, \rho_k) - \sum_{s=1}^{S} f_s(z_k^{(s)}) - \sum_{s=1}^{S} \psi_k^{(s)}(d_k^{(s)}) \right| < \epsilon \tag{3.25}$$

is satisfied. If yes, stop; Otherwise, go to Step 5*;*

Step 5. *Update the penalty parameter* ρ. *If*

$$\sum_{s=1}^{S} \left(\nabla_s f_s(z_k^{(s)})^\top d_k^{(s)} + \frac{1}{2} d_k^{(s)\top} H_{ks} d_k^{(s)} - \rho_k \| h_s(z_k^{(s)}) \|_1 \right) \le 0, \tag{3.26}$$

let $\rho_{k+1} = \rho_k$*; Otherwise,*

$$\rho_{k+1} = \max \left\{ \frac{\sum_{s=1}^{S} \left(\nabla_s f_s(z_k^{(s)})^\top d_k^{(s)} + \frac{1}{2} d_k^{(s)\top} H_{ks} d_k^{(s)} \right)}{\sum_{s=1}^{S} \| h_s(z_k^{(s)}) \|_1},\ 2\rho_k \right\}; \tag{3.27}$$

Step 6. *Select the least positive integer* ℓ *such that*

$$\begin{aligned} &M(z_k + \beta^\ell d_k, \rho_{k+1}) - M(z_k, \rho_{k+1}) \\ &\le \delta \beta^\ell \sum_{s=1}^{S} \left(\nabla_s f_s(z_k^{(s)})^\top d_k^{(s)} - \rho_{k+1} \| h_s(z_k^{(s)}) \|_1 \right). \end{aligned} \tag{3.28}$$

Let $\alpha_k = \beta^\ell$ *and* $z_{k+1}^{(s)} = z_k^{(s)} + \alpha_k d_k^{(s)} (s = 1, \ldots, S)$*;*

Step 7. *Update* H_{ks} *to* $H_{(k+1)s}$, *and calculate* $f_s(z_{k+1}^{(s)})$, $h_s(z_{k+1}^{(s)})$, $\nabla_s f_s(z_{k+1}^{(s)})$, $\nabla_s h_s(z_{k+1}^{(s)})$ *for* $s = 1, \ldots, S$ *and* B_{k+1}. *Set* $k = k + 1$ *and go to Step* 1.

Note 1. One of the key difficulties for an iterative method is how to generate the search direction, by which the new approximate to the solution is generated. In

Algorithm 3.4, we generate the search direction by solving a set of QP subproblems (3.18)–(3.24), where (3.18)–(3.19) is the decomposition of the problem (2.18)–(2.19). It is noted that (3.18)–(3.19) and (3.23)–(3.24) are QP subproblems with n variables and m constraints. Thus, they can be solved by standard algorithms for QP. On the other hand, for $s = 1, 2, \ldots, S$, (3.18)–(3.19) and (3.23)–(3.24) can be solved parallelly.

Note 2. Problem (3.20) is a strictly concave unconstrained quadratic minimization problem with m_0 variables, which may be very large. The Newton method can find the optimal solution in one iteration. Large memory, however, is required for the inverse of matrix $B_k + \nu_j I$. We suggest using conjugate gradient methods with exact line search procedure, which do not need the information on $(B_k + \nu_j I)^{-1}$, and the optimal solution will be derived in a finite number of iterations (e.g., see Bazaraa, Sherali, and Shetty [1]). It is well known that the conjugate gradient method can be more efficient than the Newton method in dealing with the large-scale unconstrained convex optimization problems.

Note 3. Step 2 is designed for maximizing function $q_k(\mu)$, which is defined by (2.20). Since the Hessian of $q_k(\mu)$ may be singular, our algorithm for maximizing (2.20) is similar to the well-known Levenberg–Marquardt method for the linear least squares problems. However, if we use some heuristics to process the matrix $[\nabla h_1(z_k^{(1)}) \ \cdots \ \nabla h_S(z_k^{(S)}) \ A^\top]$ such that the condition of Lemma 2.3 holds (the existence of solution is guaranteed by Lemma 2.4) at each iterate k, then we need only solve one strictly convex unconstrained quadratic optimization, which can be done easily.

Note 4. The approximate Hessian H_{ks} is updated by Powell's damped BFGS formulae:

$$H_{(k+1)s} = H_{ks} - \frac{H_{ks} u_k^{(s)} u_k^{(s)\top} H_{ks}}{u_k^{(s)\top} H_{ks} u_k^{(s)}} + \frac{v_k^{(s)} v_k^{(s)\top}}{u_k^{(s)\top} v_k^{(s)}}, \tag{3.29}$$

where

$$v_k^{(s)} = \begin{cases} \hat{v}_k^s, & \hat{v}_k^{(s)\top} u_k^{(s)} \geq 0.2 u_k^{(s)\top} H_{ks} u_k^{(s)}, \\ \theta_k \hat{v}_k^{(s)} + (1-\theta_k) H_{ks} u_k^{(s)} & \text{otherwise}, \end{cases}$$

$\hat{v}_k^{(s)} = \nabla_s L_s(z_{k+1}^{(s)}, \lambda_{k+1}^{(s)}) - \nabla_s L_s(z_k^{(s)}, \lambda_{k+1}^{(s)})$, $u_k^{(s)} = z_{k+1}^{(s)} - z_k^{(s)}$, and

$$\theta_k = 0.8 u_k^{(s)\top} H_{ks} u_k^{(s)} / (u_k^{(s)\top} H_{ks} u_k^{(s)} - u_k^{(s)\top} \hat{v}_k^{(s)}).$$

It can be proved that $H_{(k+1)s}$ is positive definite if H_{ks} is positive definite.

It may be helpful for us to understand the algorithm and its differences from the direct SQP based on the scenario tree formulation (see [4, 5, 31]) by a two-stage stochastic program example. Suppose there are 5 variables and 3 constraints for the first stage, and 10 variables and 6 constraints for the second stage. If the number S of scenarios is 2000, then we need to solve 2000 QPs with 15 variables and 9 constraints for each QP (which are small-size QPs and can be solved very easily parallelly or sequentially; for any s, we need only save the coefficients corresponding to s which will be replaced by the coefficients corresponding to $s+1$) in Step 1 and Step 3 of Algorithm 3.4, respectively. The program (3.20) in Step 2 of Algorithm 3.4 is an unconstrained convex optimization problem with 5×(2000-1) variables. Comparatively,

all QPs derived from the direct SQP applied to the scenario tree formulation (see [4, 5, 31]) have 20005 variables and 12003 constraints, which is much larger than the QPs in Algorithm 3.4. Moreover, the scales of QPs increase in S times as the numbers of variables and constraints in the second stage increase, whereas the scale of program (3.20) remains unchanged. In some sense, the algorithm in this paper diminishes the bottleneck of SQP in solving some classes of stochastic programs. We think our algorithm can be an alternative to the existing algorithms in solving the very difficult stochastic nonlinear programs.

4. Global convergence. In this section, we prove that the algorithm will terminate finitely at a KKT point of problem (1.5)–(1.7), or an ϵ-optimal solution with any desirable accuracy ϵ will be derived after a finite number of iterations. If $\epsilon = 0$, Algorithm 3.4 will converge to a KKT point of the problem.

ASSUMPTION 4.1.

(1) *For $s = 1, 2, \ldots, S$, f_s and h_s are twice continuously differentiable functions on $\Re^n$, respectively.*

(2) *The sequence $\{z_k\}$ is bounded.*

(3) *For all $k \geq 0$ and $s = 1, \ldots, S$, $\nabla_s h_{ks}$ has full column rank.*

(4) *There exist positive constants δ_1 and δ_2 such that $\delta_1 < \delta_2$ and $\delta_1 \|p\|^2 \leq p^\top H_{ks} p \leq \delta_2 \|p\|^2$ for all $p \in \Re^n$ and all $k \geq 0$, and $s = 1, 2, \ldots, S$.*

(5) *The sequence $\{\mu_k\}$ is bounded.*

In Assumption 4.1, we assume that the sequence $\{\mu_k\}$ exists, in which case we do not use any condition on (1.1)–(1.4). The conditions (1) and (2) are common in an analysis on global convergence of the algorithm for nonlinear smoothing optimization, and (4) is general for convergence of SQP methods. Although (3) may not hold for some problem, it is not restrictive and critical for the algorithm, and some technique for general nonlinear programs (e.g., see [19]) can be introduced to deal with it. Assumption 4.1 (5) is necessary for the global convergence of our algorithm, which is not special in the class of multiplier methods.

LEMMA 4.2. *There holds*

$$\sum_{s=1}^{S} A_s d_k^{(s)} = 0 \tag{4.1}$$

for all $k \geq 0$.

Proof. The result follows from (3.21), Lemma 2.2 (i), and from the fact that $\sum_{s=1}^{S} A_s z_0^{(s)} = 0$. □

If $(d_k^{(1)}, \ldots, d_k^{(S)}) = 0$ for some iterate k, then it follows from the KKT conditions of (3.23)–(3.24) that z_k is a KKT point of problem (1.5)–(1.7).

The next lemma shows that the penalty parameter will remain constant after a finite number of iterations.

LEMMA 4.3. *Under Assumption 4.1, there is a constant $\bar{\rho} > 0$ such that $\rho_k = \bar{\rho}$ for all sufficiently large k.*

Proof. Since $d_k^{(s)}$ solves problem (3.23)–(3.24), there exists a $\lambda_k^{(s)} \in \Re^m$ such that

$$\nabla_s f_{ks} + A_s^\top \mu_{k+1} + H_{ks} d_k^{(s)} + \nabla_s h_{ks} \lambda_k^{(s)} = 0, \tag{4.2}$$

$$h_{ks} + \nabla_s h_{ks}^\top d_k^{(s)} = 0. \tag{4.3}$$

Thus, by (4.1),

$$\begin{aligned}
&\sum_{s=1}^{S}\left(\nabla_s f_{ks}^{\top} d_k^{(s)} + \frac{1}{2} d_k^{(s)\top} H_{ks} d_k^{(s)} - \rho_k \|h_{ks}\|_1\right) \\
&= \sum_{s=1}^{S}\left(-\mu_{k+1}^{\top} A_s d_k^{(s)} + \lambda_k^{(s)\top} h_{ks} - \frac{1}{2} d_k^{(s)\top} H_{ks} d_k^{(s)} - \rho_k \|h_{ks}\|_1\right) \qquad (4.4)\\
&\le \sum_{s=1}^{S}\left(\|\lambda_k^{(s)}\|_\infty - \rho_k\right)\|h_{ks}\|_1.
\end{aligned}$$

Under Assumption 4.1, by (4.2)–(4.3) and doing some calculations, we have

$$\lambda_k^{(s)} = (\nabla_s h_{ks}^{\top} H_{ks}^{-1} \nabla_s h_{ks})^{-1}(h_{ks} - \nabla_s h_{ks}^{\top} H_{ks}^{-1}(\nabla_s f_{ks} + A_s^{\top}\mu_{k+1})). \qquad (4.5)$$

Thus, by Assumption 4.1, there is a constant $\gamma > 0$ such that

$$\max\{\|\lambda_k^{(s)}\|_\infty, s = 1, 2, \ldots, S \text{ and all } k \ge 0\} \le \gamma. \qquad (4.6)$$

Hence, by Step 5 of the algorithm and (4.4), there exists a $\bar{\rho} \ge \gamma$ and integer $k_0 > 0$ such that $\rho_k \ge \bar{\rho}$ for all $k \ge k_0$. □

By (4.2), $d_k^{(s)} = -H_{ks}^{-1}(\nabla_s f_{ks} + A_s^{\top}\mu_{k+1} + \nabla_s h_{ks}\lambda_k^{(s)})$. Thus, it follows from (4.6) and Assumption 4.1 that $\{(d_k^{(1)}, \ldots, d_k^{(S)})\}$ is bounded.

Without loss of generality, we suppose that $\rho_k = \bar{\rho}$ for all $k \ge 0$. Let

$$\Pi_k(d) = \sum_{s=1}^{S}\left[\nabla_s f_{ks}^{\top} d^{(s)} + \bar{\rho}(\|h_{ks} + \nabla_s h_{ks}^{\top} d^{(s)}\|_1 - \|h_{ks}\|_1)\right]. \qquad (4.7)$$

Then Π_k is a convex function on d, $\Pi_k(0) = 0$, and

$$\Pi_k(d_k) = \sum_{s=1}^{S}\left(\nabla_s f_{ks}^{\top} d_k^{(s)} - \bar{\rho}\|h_{ks}\|_1\right). \qquad (4.8)$$

For $s = 1, 2, \ldots, S$, f_s and h_s are twice continuously differentiable functions, so ∇f_s and ∇h_s are Lipschitz continuous on a given bounded set $\mathcal{Q}$. In particular, there exists a positive constant a_0 such that

$$\|\nabla_s f_s(\hat{z}^{(s)}) - \nabla_s f_s(\tilde{z}^{(s)})\| \le a_0\|\hat{z}^{(s)} - \tilde{z}^{(s)}\|, \qquad (4.9)$$

$$\|\nabla_s h_s(\hat{z}^{(s)}) - \nabla_s h_s(\tilde{z}^{(s)})\| \le a_0\|\hat{z}^{(s)} - \tilde{z}^{(s)}\| \qquad (4.10)$$

for all $\hat{z} \in \mathcal{Q}$ and $\tilde{z} \in \mathcal{Q}$.

LEMMA 4.4. *Under Assumption* 4.1*, for $\alpha \ge 0$ and $k \ge 0$, there is a constant C_1 such that*

$$M(z_k + \alpha d_k, \bar{\rho}) - M(z_k, \bar{\rho}) - \Pi_k(\alpha d_k) \le \frac{C_1}{2}\alpha^2\|d_k\|^2. \qquad (4.11)$$

Proof. By (4.9),

$$\begin{aligned}
&f_s(z_k^{(s)} + \alpha d_k^{(s)}) - f_{ks} - \alpha\nabla_s f_{ks}^{\top} d_k^{(s)} \\
&= \int_0^{\alpha}(\nabla_s f_s(z_k^{(s)} + t d_k^{(s)}) - \nabla_s f_{ks})^{\top} d_k^{(s)} dt \qquad (4.12)\\
&\le \frac{1}{2} a_0 \alpha^2 \|d_k^{(s)}\|^2.
\end{aligned}$$

Similarly, by (4.10), we have that $\|h_s(z_k^{(s)}+\alpha d_k^{(s)})-h_{ks}-\alpha\nabla_s h_{ks}^\top d_k^{(s)}\| \le \frac{1}{2}a_0\alpha^2\|d_k^{(s)}\|^2$. Thus, by the properties of the norm,

$$M(z_k+\alpha d_k,\bar\rho)-M(z_k,\bar\rho)-\Pi_k(\alpha d_k)\le\frac{1}{2}\alpha^2\sum_{s=1}^{S}(a_0+\bar\rho a_0C_0)\|d_k^{(s)}\|^2, \tag{4.13}$$

where $C_0>0$ is a constant. Let $C_1=a_0(1+\bar\rho C_0)$. The desired result follows immediately. ∎

The following result shows that the line search procedure in Algorithm 3.4 is well-defined.

LEMMA 4.5. *Under Assumption* 4.1*, there holds*

$$\alpha_k>\beta\hat\alpha, \tag{4.14}$$

where $\hat\alpha=\min\{1,\ (1-\delta)\delta_1/C_1\}$.

Proof. By $\Pi_k(0)=0$ and the convexity of Π_k, for $\alpha\in[0,1]$, we have that

$$\Pi_k(\alpha d_k)-\delta\alpha\Pi_k(d_k)\le(1-\delta)\alpha\Pi_k(d_k). \tag{4.15}$$

It follows from (4.8), (3.26), and Assumption 4.1(3) that

$$\Pi_k(d_k)\le-\frac{1}{2}\sum_{s=1}^{S}d_k^{(s)\top}H_{ks}d_k^{(s)}\le-\frac{1}{2}\delta_1\|d_k\|^2. \tag{4.16}$$

Thus, by (4.11), (4.15), and (4.16),

$$M(z_k+\alpha d_k,\bar\rho)-M(z_k,\bar\rho)-\delta\alpha\Pi_k(d_k)\le\frac{1}{2}\alpha(C_1\alpha-(1-\delta)\delta_1)\|d_k\|^2, \tag{4.17}$$

which implies that $\alpha_k/\beta>\hat\alpha$ by (3.28). ∎

The following result shows the global convergence of our algorithm.

THEOREM 4.6. *Suppose that Assumption* 4.1 *holds. For any given* $\epsilon>0$*, Algorithm* 3.4 *will terminate finitely at a KKT point or an* ϵ*-optimal solution of problem* (1.5)–(1.7)*. If* $\epsilon=0$ *and* $\{z_k\}$ *is an infinite sequence, then any cluster point of* $\{z_k\}$ *is a KKT point of problem* (1.5)–(1.7)*.*

Proof. If for some positive integer $\tilde k$, $d_{\tilde k}=0$, then Algorithm 3.4 terminates finitely at the KKT point $z_{\tilde k}$.

Suppose that the algorithm will not terminate finitely; i.e., there is an infinite sequence $\{z_k\}$ which does not satisfy (3.25) for any given ϵ'.

Let $\epsilon'=2\bar\rho\epsilon$. By Lemma 4.3, there is an integer $k_0>0$ such that for $k\ge k_0$,

$$\rho_k=\bar\rho. \tag{4.18}$$

It follows from (3.28), (4.14), and (4.16) that for $k\ge k_0$,

$$M(z_{k+1},\bar\rho)-M(z_k,\bar\rho)\le\delta\alpha_k\Pi_k(d_k)<-\frac{1}{2}\delta\beta\hat\alpha\delta_1\|d_k\|^2<0, \tag{4.19}$$

which implies that $\{M(z_k,\bar\rho)\}$ is a monotonically decreasing sequence. Thus, by Assumption 4.1, $\{M(z_k,\bar\rho)\}$ is convergent, which results in

$$\|d_k\|^2\to0. \tag{4.20}$$

It follows that there is a constant $C_0 > 0$ and a positive integer k_1 such that for $k \geq k_1$,

$$\begin{aligned}\sum_{s=1}^{S} \|h_s(z_k^{(s)})\|_1 &\leq \sum_{s=1}^{S} C_0 \|h_s(z_k^{(s)})\| \\ &= \sum_{s=1}^{S} C_0 \left(\|h_s(z_k^{(s)})\| - \|h_s(z_k^{(s)}) + \nabla_s h_s(z_k^{(s)})^\top d_k^{(s)}\| \right) \\ &\leq \sum_{s=1}^{S} C_0 \|\nabla_s h_s(z_k^{(s)})\| \|d_k^{(s)}\| \qquad (4.21) \\ &\leq \frac{1}{2(\bar{\rho}+\gamma)} \epsilon^{'} \leq \epsilon,\end{aligned}$$

where γ is defined in (4.6).

Moreover, by (4.1),

$$\begin{aligned}\left\| \sum_{s=1}^{S} A_s z_k^{(s)} \right\|_1 &= \left\| \sum_{s=1}^{S} A_s (z_{k-1}^{(s)} + \alpha_{k-1} d_{k-1}^{(s)}) \right\|_1 \\ &= \left\| \sum_{s=1}^{S} A_s z_{k-1}^{(s)} \right\|_1 = \left\| \sum_{s=1}^{S} A_s z_0^{(s)} \right\|_1 = 0. \qquad (4.22)\end{aligned}$$

Furthermore, by Assumption 4.1, there is an integer $k_2 > 0$ such that for $k \geq k_2$,

$$\sum_{s=1}^{S} \frac{1}{2} d_k^{(s)\top} H_{ks} d_k^{(s)} \leq \frac{\epsilon^{'}}{4} \qquad (4.23)$$

and $\|H_{ks} d_k^{(s)}\| \leq \epsilon$ for $s = 1, 2, \ldots, S$. Combining with (4.21) and (4.22), we see that for all $k \geq \max\{k_0, k_1, k_2\}$, z_k is an ϵ-optimal solution of problem (1.5)–(1.7) by Definition 3.3 and (4.2), and

$$\begin{aligned}&\left| M(z_k, \bar{\rho}) - \sum_{s=1}^{S} f_s(z_k^{(s)}) - \sum_{s=1}^{S} \psi_k^{(s)}(d_k^{(s)}) \right| \\ &\leq (\bar{\rho} + \max_{s=1,\cdot,S} \|\lambda_k^{(s)}\|_\infty) \sum_{s=1}^{S} \|h_s(z_k^{(s)})\|_1 + \sum_{s=1}^{S} \frac{1}{2} d_k^{(s)\top} H_{ks} d_k^{(s)} \qquad (4.24) \\ &\leq \frac{3}{4} \epsilon^{'} < \epsilon^{'},\end{aligned}$$

which is a contradiction. The contradiction implies the first part of the result.

Now we prove the last half of the result. Suppose that $z_k \to z^*$ for $k \in K$. Then $\| \sum_{s=1}^{S} A_s z^{*(s)} \|_1 = 0$. The boundednesses of $\{\lambda_k\}$ and $\{\mu_k\}$ imply that there is $K^{'} \subset K$ such that $\lambda_k \to \lambda^*$ and $\mu_k \to \mu^*$ for $k \to \infty$ and $k \in K^{'}$. Let $\epsilon^{'} \to 0$ and $k \to \infty$ for $k \in K^{'}$. Taking the limit on the two sides of the second inequality of (4.24), we have

$$\sum_{s=1}^{S} \|h_s(z^{*(s)})\|_1 = 0, \text{ and } d_k^{(s)} \to d^{*(s)} = 0 \ (s = 1, \ldots, S). \qquad (4.25)$$

Thus, the desired result follows immediately from (4.2). □

5. Extension to the stochastic nonlinear programs with inequality constraints. According to the discussion in the introduction, the stochastic nonlinear programs with inequality constraints can be reformulated as the following nonlinear programming problem with equality and inequality constraints (including the nonanticipativity constraints):

$$\min \sum_{s=1}^{S} f_s(z^{(s)}) \tag{5.1}$$

$$\text{s.t. } h_s(z^{(s)}) = 0,\ s = 1, \ldots, S, \tag{5.2}$$

$$g_s(z^{(s)}) \le 0,\ s = 1, \ldots, S, \tag{5.3}$$

$$\sum_{s=1}^{S} A_s z^{(s)} = 0, \tag{5.4}$$

where $g_s : \Re^n \to \Re^{m_1} (s = 1, \ldots, S)$. By introducing slack vectors $w^{(s)} \in \Re^{m_1} (s = 1, \ldots, S)$, and transforming the nonnegative constraints on $w^{(s)}$ to the barrier terms in the objective function, we derive the following program:

$$\min \sum_{s=1}^{S} \left[f_s(z^{(s)}) - \beta \sum_{i=1}^{m_1} \ln w_i^{(s)} \right] \tag{5.5}$$

$$\text{s.t. } h_s(z^{(s)}) = 0,\ s = 1, \ldots, S, \tag{5.6}$$

$$g_s(z^{(s)}) + w^{(s)} = 0,\ s = 1, \ldots, S, \tag{5.7}$$

$$\sum_{s=1}^{S} A_s z^{(s)} = 0, \tag{5.8}$$

where $\beta > 0$ is a parameter.

It is easy to note that program (5.5)–(5.8) has precisely the form of program (1.5)–(1.7) if we set $z^{(s)} := (z^{(s)}, w^{(s)})$, $f_s(z^{(s)}) := f_s(z^{(s)}) - \beta \sum_{i=1}^{m_1} \ln w_i^{(s)}$, $h_s(z^{(s)}) := (h_s(z^{(s)}), g_s(z^{(s)}) + w^{(s)})$, and $A_s := [A_s\ 0]$ in program (1.5)–(1.7). By [11], under suitable conditions, as $\beta \to 0$, the sequence of the solution of problem (5.5)–(5.8) converges to the solution of program (5.1)–(5.4). Thus, the decomposition method in this paper may be extended to solve stochastic nonlinear programs with inequality constraints. The difficulties include how to avoid that some slack variables are reduced too fast, which may result in failures of many interior point methods for *nonconvex* nonlinear programming in converging to any stationary point of a simple and very regular problem (see [32]). If any of functions $\hat{c}_0$, c_0, q_t, and $c_t (t = 1, \ldots, T-1)$ in problem (1.1)–(1.4) is *nonconvex*, the difficulty could be reflected in generating the estimates of the dual multiplier corresponding to the nonanticipativity constraints. Although Byrd, Gilbert, and Nocedal [7] presented a trust region method to circumvent the convergence difficulties for nonlinear programming, we have no idea how to deal with the trust region constraints in the decomposition. We think the other possible extension is to further develop the active set technique in the decomposition.

6. Preliminary numerical results. A MATLAB subroutine was programmed to test Algorithm 3.4 and run under version 5.3. All QP subproblems in Algorithm 3.4 were solved by *quadprog.m*, an M-file in MATLAB toolbox. Four test problems are originated from the modifications on the standard nonlinear programming problem 263 in Schittkowski [30].

We chose termination tolerance scalar $\epsilon = 10^{-6}$. Moreover, in our implementation, instead of (3.21), we use

$$\|B_k\mu_{j+1} - A\hat{d}_k\| \le \epsilon_0 \ \text{ or } \ \|d_\mu\| \le \epsilon_0, \tag{6.1}$$

where $\epsilon_0 = 10^{-8}$ is a small tolerance scalar. The initial parameters are $\delta = 0.01$, $\beta = 0.8$, $\sigma_0 = 1$, and $\rho_0 = 1$. The initial approximation to the Lagrangian Hessian H_{0s} is selected to be the $n \times n$ identity matrix. The starting point is $z_0^{(s)} = (10, 10, 10, 10, 10, 10)$ for $s = 1, \ldots, S$.

In order to see the performance of Algorithm 3.4 clearly, we just use two independent random variables ξ_1 and ξ_2, and suppose that each may be 1 and 4 with the same probability $\frac{1}{2}$. Thus, we have $S = 4$.

The first test problem (TP1) is

$$\min \ -24z_1 + 2z_2^2 + 4z_3^2 + 4z_4^2 + EQ(z_1, z_2, z_3, z_4, \xi_1, \xi_2), \tag{6.2}$$

where

$$Q(z_1, z_2, z_3, z_4, \hat{\xi}_1, \hat{\xi}_2) = \min -12z_5 + 4z_6^2 \tag{6.3}$$

$$\text{s.t. } -z_1^3 + z_2 - z_3^2 + z_5^2 = \hat{\xi}_1, \tag{6.4}$$

$$z_1^2 - z_2 - z_4^2 + z_6^2 = \hat{\xi}_2 \tag{6.5}$$

and $(\hat{\xi}_1, \hat{\xi}_2)$ is the realization of $\xi = (\xi_1, \xi_2)$. We have the same case if it is not specified later.

TP1 is a two-stage stochastic program, where (z_1, z_2, z_3, z_4) is the decision vector in the first stage and (z_5, z_6) is the decision vector in the second stage. Thus, the coefficient matrix of the nonanticipativity constraints can be written as follows:

$$A = \begin{bmatrix} I_4 & 0_{4\times 2} & -I_4 & 0_{4\times 2} & & & & \\ & & I_4 & 0_{4\times 2} & -I_4 & 0_{4\times 2} & & \\ & & & & I_4 & 0_{4\times 2} & -I_4 & 0_{4\times 2} \end{bmatrix}, \tag{6.6}$$

where I_4 is the 4×4 identity matrix. By scenario analysis in section 1, TP1 is reformulated as a nonlinear program with 24 variables, 8 separable constraints, and 12 nonanticipativity constraints,

$$f_s(z^{(s)}) = -6z_1^{(s)} + \frac{1}{2}{z_2^{(s)}}^2 + {z_3^{(s)}}^2 + {z_4^{(s)}}^2 - 3z_5^{(s)} + {z_6^{(s)}}^2. \tag{6.7}$$

The separable constraints corresponding to the sth scenario are

$$\begin{cases} -{z_1^{(s)}}^3 + z_2^{(s)} - {z_3^{(s)}}^2 + {z_5^{(s)}}^2 = \xi_1^{(s)}, \\ {z_1^{(s)}}^2 - z_2^{(s)} - {z_4^{(s)}}^2 + {z_6^{(s)}}^2 = \xi_2^{(s)}. \end{cases} \tag{6.8}$$

It is easy to prove that the assumption on full column rank in Lemma 2.3 holds at z_0 for this problem. However, this assumption does not hold at any iterate for the other test problems.

The solution to the decision vector in the first stage is $(10.1701, 0.9538, 0.0, 10.0736)$. The dual multiplier μ corresponding to the nonanticipativity constraints is $(-33.8223, 1.6633, 0.0, 33.5114, 6.8685, -0.3367, 0.0, -6.7831, -40.6703, 2.0, 0.0, 40.2945)$. The optimal value of the objective function is -219.8369. The numerical results of program (6.2)–(6.5) are listed in Table 1.

Our second test example (TP2) is

$$\min\ -24z_1 + 2z_2^2 + 4z_3^2 + 4z_4^2 + E_{\xi_1} Q_1(z_1, z_2, z_3, z_4, \xi_1) \tag{6.9}$$

$$\text{s.t. } -z_1^3 + z_2 - z_3^2 = 0, \tag{6.10}$$

$$z_1^2 - z_2 - z_4^2 = 0, \tag{6.11}$$

where

$$Q_1(z_1, z_2, z_3, z_4, \hat{\xi}_1) = \min\ -12z_5 + E_{\xi_2} Q_2(z_1, z_2, z_3, z_4, z_5, \hat{\xi}_1, \xi_2) \tag{6.12}$$

$$\text{s.t. } -z_1^3 + z_2 + z_5^2 = \hat{\xi}_1 \tag{6.13}$$

and

$$Q_2(z_1, z_2, z_3, z_4, z_5, \hat{\xi}_1, \hat{\xi}_2) = \min\ 4z_6^2 \tag{6.14}$$

$$\text{s.t. } z_1^2 - z_2 - z_5^2 + 2z_6^2 = \hat{\xi}_2. \tag{6.15}$$

It is a three-stage program, where (z_1, z_2, z_3, z_4) is the decision vector in the first stage and (6.10)–(6.11) are the constraints in the first stage. z_5, ξ_1, and constraint (6.13) correspond to the second stage, while z_6, ξ_2, and (6.15) correspond to the third stage. In this case,

$$A = \begin{bmatrix} I_4 & 0 & 0 & -I_4 & 0 & 0 & & & & & & \\ & & & I_4 & 0 & 0 & -I_4 & 0 & 0 & & & \\ & & & & & & I_4 & 0 & 0 & -I_4 & 0 & 0 \\ & 1 & 0 & 0 & -1 & 0 & & & & & & \\ & & & & & & & 1 & 0 & 0 & -1 & 0 \end{bmatrix}, \tag{6.16}$$

where sizes of block 0's can be identified easily.

By scenario analysis, similar to (6.8), constraints (6.10)–(6.11) are reformulated as 8 constraints, and it is easy to verify that the set of their gradients are linearly dependent with the transposes of the first 12 row vectors of A and the rank is 14. In fact, for this problem and the next two test problems, that the Jacobian of (2.5)–(2.6) does not have full column rank at any iterate can be judged directly from the fact that the number of constraints is larger than the number of variables in the reformulation of the problem.

The solution corresponding to the first stage is $(1.0, 1.0, 0.0, 0.0)$, and the dual multiplier $\mu^* = (0.0291, 0.0127, 0.0, 0.0, 0.1116, 0.0672, 0.0354, 0.0, 0.0, 0.0298, 0.0159, 0.0, 0.0, 0.0964)$. The optimal value of the objective function is -30.

The third test example (TP3) is

$$\min\ -24z_1 + 2z_2^2 + E_{\xi_1} Q_1(z_1, z_2, \xi_1) \tag{6.17}$$

$$\text{s.t. } z_1 + z_2^2 = 2, \tag{6.18}$$

where

$$Q_1(z_1, z_2, \hat{\xi}_1) = \min\ 4z_3^2 + 4z_4^2 + E_{\xi_2} Q_2(z_1, z_2, z_3, z_4, \hat{\xi}_1, \xi_2) \tag{6.19}$$

$$\text{s.t. } -z_1^3 + z_2 + z_3^2 = \hat{\xi}_1, \tag{6.20}$$

$$z_1^2 - z_2 + z_4^2 = \hat{\xi}_1 \tag{6.21}$$

TABLE 1

Problem	TP1	TP2	TP3	TP4
nS	24	24	24	24
mS	8	16	20	24
m_0	12	14	10	10
NI	16	16	10	9
NF	27	23	11	10
NG	17	17	11	10
R-KT	4.9749e-05	4.6236e-04	7.1064e-06	8.8052e-13
R-CN	2.7528e-08	2.1505e-07	1.1953e-07	2.5381e-09
R-NA	6.2045e-09	7.4422e-09	5.9241e-09	0
R-ND	3.3235e-05	4.2445e-05	7.1280e-07	6.3453e-10

and

$$Q_2(z_1, z_2, z_3, z_4, \hat{\xi}_1, \hat{\xi}_2) = \min \quad -12z_5 + 4z_6^2 \tag{6.22}$$

$$\text{s.t. } -z_1^3 + z_2 - z_3^2 + z_5^2 = \hat{\xi}_2, \tag{6.23}$$

$$z_1^2 - z_2 - z_4^2 - z_5^2 + 2z_6^2 = \hat{\xi}_2. \tag{6.24}$$

In this problem, z_1 and z_2 are the decision variables corresponding to the first stage. z_3 and z_4 correspond to the second stage, and z_5, z_6 belong to the third stage. Thus,

$$A = \begin{bmatrix} I_2 & 0 & 0 & -I_2 & 0 & 0 & & & & & & \\ & & & I_2 & 0 & 0 & -I_2 & 0 & 0 & & & \\ & & & & & & I_2 & 0 & 0 & -I_2 & 0 & 0 \\ & I_2 & 0 & 0 & -I_2 & 0 & & & & & & \\ & & & & & & & I_2 & 0 & 0 & -I_2 & 0 \end{bmatrix}. \tag{6.25}$$

The solution is $z_1^* = 1.3451$, $z_2^* = 0.8093$. $\mu^* = (-1.6951, 1.5582, -5.7042, 0.0, -4.2031, 15.4878, -1.0663, 1.2661, 0.9857, -6.1159)$. The optimal value of the objective function is -20.1446.

The objective function in our fourth test problem (TP4) is $-4z_1$, which does not relate to the variables in stages 2 and 3; thus it is a stochastic program with no recourse. A constraint

$$z_1^2 - z_2 = 0 \tag{6.26}$$

is added to (6.18) to form the constraints of the first stage of TP4. Since there are the same number of stages, variables, and scenarios in each stage as TP3, the coefficient matrix A is the same as (6.25). The solution to this problem is $z_1^* = 1.0$, $z_2^* = 1.0$. Vector μ^* is a 10-dimensional zero vector. The optimal value of the objective function is -4. The numerical results of problems TP2, TP3, and TP4 are also summarized in Table 1.

In Table 1, nS is the number of all variables of the reformulation problem, and mS and m_0 are the numbers of separable constraints and the nonanticipativity constraints, respectively. NI, NF, and NG represent, respectively, the numbers of iterations, function evaluations, and gradient calculations. R-KT is the l_2 norm of the gradient of the Lagrangian, and R-CN, R-NA, R-ND represent the l_2 norms of the residuals of the separable constraints and the nonanticipativity constraints, and the search direction, respectively.

The numerical results in Table 1 show us that Algorithm 3.4 has solved the test problems TP1–TP4 successfully, and the approximate KKT points for these problems have been derived. Although there are some algorithms for stochastic nonlinear programs, as we note, this is the first paper for stochastic nonconvex programs. It is not enough to draw a conclusion for our algorithm by these numerical experiments, and further computation should be done for larger scale stochastic programs with larger scenario numbers S and larger m and n; we think that the discussion in this paper may give us some clue to develop better algorithms for solving stochastic nonlinear programs.

Acknowledgments. We would like to thank the associate editor and anonymous referees for their valuable comments, which improved the paper greatly.

REFERENCES

[1] M.S. Bazaraa, H.D. Sherali, and C.M. Shetty, *Nonlinear Programming, Theory and Algorithms*, 2nd ed., John Wiley and Sons, New York, 1993.
[2] J.R. Birge, *Current Trends in Stochastic Programming Computation and Applications*, Technical Report 95-15, Department of Industrial and Operations Engineering, University of Michigan, Ann Arbor, MI, 1995.
[3] J.R. Birge and F. Louveaux, *Introduction to Stochastic Programming*, Springer-Verlag, New York, 1997.
[4] J. Blomvall and P.O. Lindberg, *A Riccati-based primal interior point solver for multistage stochastic programming*, European J. Oper. Res., 143 (2002), pp. 452–461.
[5] J. Blomvall and P.O. Lindberg, *A Riccati-based primal interior point solver for multistage stochastic programming - extensions*, Optim. Methods Softw., 17 (2002), pp. 383–407.
[6] P.T. Boggs and J.W. Tolle, *Sequential Quadratic Programming*, in Acta Numerica 1995, Cambridge University Press, Cambridge, UK, 1995, pp. 1–51.
[7] R.H. Byrd, J.C. Gilbert, and J. Nocedal, *A trust region method based on interior point techniques for nonlinear programming*, Math. Program., 89 (2000), pp. 149–185.
[8] B.J. Chun and S.M. Robinson, *Scenario analysis via bundle decomposition*, Ann. Oper. Res., 56 (1995), pp. 39–63.
[9] G.B. Dantzig and P. Wolfe, *Decomposition principle for linear programs*, Oper. Res., 8 (1960), pp. 101–111.
[10] B. Feinberg, *Coercion functions and decentralized linear programming*, Math. Oper. Res., 14 (1989), pp. 177–187.
[11] A.V. Fiacco and G.P. McCormick, *Nonlinear Programming: Sequential Unconstrained Minimization Techniques*, John Wiley and Sons, New York, 1968, reprinted, SIAM, Philadelphia, 1990.
[12] M. Fukushima, *A successive quadratic programming algorithm with global and superlinear convergence properties*, Math. Programming, 35 (1986), pp. 253–264.
[13] M. Fukushima, *A successive quadratic programming method for a class of constrained nonsmooth optimization problems*, Math. Programming, 49 (1991), pp. 231–251.
[14] S.P. Han, *A decomposition method and its application to convex programming*, Math. Oper. Res., 14 (1989), pp. 237–248.
[15] S.P. Han, J.S. Pang, and N. Rangaraj, *Globally convergent Newton methods for nonsmooth equations*, Math. Oper. Res., 17 (1992), pp. 586–607.
[16] T. Helgason and S.W. Wallace, *Approximate scenario solutions in the progressive hedging algorithm*, Ann. Oper. Res., 31 (1991), pp. 425–444.
[17] P. Kall and S.W. Wallace, *Stochastic Programming*, John Wiley and Sons, New York, 1994.
[18] L.S. Lasdon, *Optimization Theory for Large Systems*, Macmillan, New York, 1970.
[19] X.-W. Liu and Y.-X. Yuan, *A robust algorithm for optimization with general equality and inequality constraints*, SIAM J. Sci. Comput., 22 (2000), pp. 517–534.
[20] J.M. Mulvey and H. Vladimirou, *Applying the progressive hedging algorithm to stochastic generalized networks*, Ann. Oper. Res., 31 (1991), pp. 399–424.
[21] J.-S. Pang, S.-P. Han, and N. Rangaraj, *Minimization of locally Lipschitzian functions*, SIAM J. Optim., 1 (1991), pp. 57–82.
[22] M.J.D. Powell and Y. Yuan, *A recursive quadratic programming algorithm that use differentiable exact penalty function*, Math. Programming, 35 (1986), pp. 265–278.

[23] L. Qi, *Superlinearly convergent approximate Newton methods for LC^1 optimization problems*, Math. Programming, 64 (1994), pp. 277–294.
[24] S.M. Robinson, *Extended scenario analysis*, Ann. Oper. Res., 31 (1991), pp. 385–398.
[25] R.T. Rockafellar and R. J.-B. Wets, *Scenarios and policy aggregation in optimization under uncertainty*, Math. Oper. Res., 16 (1991), pp. 119–147.
[26] A. Ruszczyński, *Parallel decomposition of multistage stochastic programming problems*, Math. Programming, 58 (1993), pp. 201–228.
[27] A. Ruszczyński, *On convergence of an augmented Lagrangian decomposition method for sparse convex optimization*, Math. Oper. Res., 20 (1995), pp. 634–656.
[28] A. Ruszczyński, *Decomposition methods in stochastic programming*, Math. Programming, 79 (1997), pp. 333–353.
[29] A. Ruszczyński, *Some advances in decomposition methods for stochastic linear programming*, Ann. Oper. Res., 85 (1999), pp. 153–172.
[30] K. Schittkowski, *More Test Examples for Nonlinear Programming Codes*, Lecture Notes in Econom. and Math. Systems 282, Springer-Verlag, Berlin, 1987.
[31] M.C. Steinbach, *Hierarchical sparsity in multistage convex stochastic programs*, in Stochastic Optimization: Algorithms and Applications, S.P. Uryasev and P.M. Pardalos, eds., Kluwer Academic Publishers, Dordrecht, The Netherlands, 2001, pp. 385–410.
[32] A. Wächter and L.T. Biegler, *Failure of global convergence for a class of interior point methods for nonlinear programming*, Math. Program., 88 (2000), pp. 565–574.
[33] G. Zhao, *Interior-point methods with decomposition for solving large-scale linear programs*, J. Optim. Theory Appl., 102 (1999), pp. 169–192.
[34] G. Zhao, *A log-barrier method with Benders decomposition for solving two-stage stochastic programs*, Math. Program., 90 (2001), pp. 507–536.
[35] G. Zhao, *A Lagrangian Dual Method with Self-Concordant Barrier for Multi-stage Stochastic Convex Nonlinear Programming*, Working paper, Math Department, National University of Singapore, Singapore, 1998.

SIAM J. OPTIM.
Vol. 14, No. 1, pp. 223–244

LARGE SCALE PARAMETER ESTIMATION USING SPARSE NONLINEAR PROGRAMMING METHODS*

JOHN T. BETTS† AND WILLIAM P. HUFFMAN†

Abstract. Traditional methods for solving "inverse problems" usually combine a standard initial-value numerical integration technique with a Gauss–Newton method for optimization. This paper presents an approach that uses neither of the traditional processes. Initial-value integration is replaced by a discretization of the relevant differential-algebraic equations. We then exploit sparse finite difference approximations to the Hessian matrix which permits us to construct a quadratically convergent algorithm for solving parameter estimation problems. Computational results with the approach are presented in a series of examples.

Key words. parameter estimation, sparse nonlinear programming, inverse problems

AMS subject classifications. 49J15, 90C30, 90C90, 70M20, 70F15, 65L10, 65L80

DOI. 10.1137/S1052623401399216

1. Introduction. The behavior of many physical processes can be described mathematically by ordinary differential or differential-algebraic equations. Commonly a finite number of parameters appear in the description of the system dynamics. A parameter estimation problem arises when it is necessary to compute values for these parameters based on observations of the system dynamics. Methods for solving these so-called inverse problems have been used for many years [10]. In fact, most techniques in use today are based on ideas proposed by Gauss more than 100 years ago, which he used to solve orbit determination problems.

One approach to solving estimation problems is to parameterize the dynamic variables using values at mesh points on the interval. A consequence of this discretization is that the original problem is *transcribed* into a finite dimensional nonlinear programming problem. Since the discrete variables directly optimize the approximate problem this approach is referred to as the *direct transcription* method. Furthermore, this nonlinear programming problem has two important properties that can be exploited. First, it is possible to efficiently compute the (Hessian) matrix of second derivatives, thereby overcoming one of the major limitations of the Gauss algorithm. Second, the Hessian and Jacobian matrices are sparse, and as a consequence very efficient linear algebra techniques can be utilized. In this paper we describe a quadratically convergent algorithm for solving parameter estimation problems.

2. The parameter estimation problem. Typically the dynamics of the system are defined for $t_I \le t \le t_F$ by a set of ordinary differential equations written in explicit form, which are referred to as the *state equations*

$$\dot{\mathbf{y}} = \mathbf{f}[\mathbf{y}(t), \mathbf{u}(t), \mathbf{p}, t], \tag{2.1}$$

where $\mathbf{y}$ is the n_y dimension state vector, and $\mathbf{u}$ is an n_u dimension vector of algebraic variables. In addition the solution must satisfy *algebraic path constraints* of the form

$$\mathbf{g}_L \le \mathbf{g}[\mathbf{y}(t), \mathbf{u}(t), \mathbf{p}, t] \le \mathbf{g}_U, \tag{2.2}$$

*Received by the editors December 5, 2001; accepted for publication (in revised form) March 7, 2003; published electronically July 18, 2003.
http://www.siam.org/journals/siopt/14-1/39921.html
†Mathematics and Engineering Analysis, The Boeing Company, P.O. Box 3707, MS 7L-21, Seattle, WA 98124-2207 (john.t.betts@boeing.com, william.p.huffman@boeing.com).

where $\mathbf{g}$ is a vector of size n_g, with elements of the form

$$g[\mathbf{y}(t), \mathbf{u}(t), \mathbf{p}, t] = \boldsymbol{\alpha}^{\mathsf{T}}\mathbf{v} + \boldsymbol{\beta}^{\mathsf{T}}\mathbf{a}[\mathbf{v}, t], \tag{2.3}$$

where

$$\mathbf{v} = \begin{bmatrix} \mathbf{y}(t) \\ \mathbf{u}(t) \\ \mathbf{p} \end{bmatrix} \tag{2.4}$$

and

$$\mathbf{a}[\mathbf{v}, t] = \begin{bmatrix} a_0(\mathbf{y}, \mathbf{u}, \mathbf{p}, t) \\ a_1(\mathbf{y}, \mathbf{u}, \mathbf{p}, t) \\ \vdots \\ a_{n_a}(\mathbf{y}, \mathbf{u}, \mathbf{p}, t) \end{bmatrix}. \tag{2.5}$$

The constraint definition can include *analytic* terms involving $\boldsymbol{\alpha}^{\mathsf{T}}\mathbf{v}$, where the $(n_y + n_u + n_p)$ vector $\boldsymbol{\alpha}$ is constant, as well as linear combinations of the n_a *auxiliary functions* $a_k(\mathbf{v}, t)$ for $k = 0, \ldots, n_a$, where the coefficients β_k are nonzero constants. By convention, a path constraint with a single nonlinear term $a_0(\mathbf{y}, \mathbf{u}, \mathbf{p}, t)$ has no auxiliary functions ($n_a = 0$). Observe that each individual path constraint may have a different number of auxiliary functions and analytic terms. In addition to the general constraints (2.2) it is computationally useful to include simple linear bounds on the state variables

$$\mathbf{y}_L \le \mathbf{y}(t) \le \mathbf{y}_U, \tag{2.6}$$

the algebraic variables

$$\mathbf{u}_L \le \mathbf{u}(t) \le \mathbf{u}_U, \tag{2.7}$$

and the n_p parameters

$$\mathbf{p}_L \le \mathbf{p} \le \mathbf{p}_U. \tag{2.8}$$

Note that an equality constraint can be imposed if the upper and lower bounds are equal; e.g., $(g_L)_k = (g_U)_k$ for some k. Boundary conditions at the initial time t_I and/or final time t_F are defined by

$$\boldsymbol{\psi}_L \le \boldsymbol{\psi}[\mathbf{y}(t_I), \mathbf{u}(t_I), t_I, \mathbf{y}(t_F), \mathbf{u}(t_F), t_F, \mathbf{p}] \le \boldsymbol{\psi}_U. \tag{2.9}$$

The basic parameter estimation problem is to determine the n_p-dimensional vector $\mathbf{p}$ to minimize the performance index

$$F = \frac{1}{2}\mathbf{r}^{\mathsf{T}}\mathbf{r} = \frac{1}{2}\sum_{k=1}^{\ell} r_k^2, \tag{2.10}$$

where $\mathbf{r}$ is the ℓ-dimensional *residual* vector. Components of the residual vector can be of two forms. State residuals are of the form

$$r_k = w_k \left[y_{i(k)}(\theta_k) - \hat{y}_{i(k)} \right], \tag{2.11}$$

where $y_{i(k)}(\theta_k)$ is the value of state variable $i(k)$ computed at time θ_k and $\hat{y}_{i(k)}$ is the observed value at the same point. Algebraic residuals are of the form

$$r_k = w_k \left[u_{i(k)}(\theta_k) - \hat{u}_{i(k)} \right], \tag{2.12}$$

where $u_{i(k)}(\theta_k)$ is the value of algebraic variable $i(k)$ computed at time θ_k and $\hat{u}_{i(k)}$ is the observed value at the same point. The residual weights are typically positive, i.e., $w_k > 0$. It is required that data evaluation points satisfy

$$t_I \leq \theta_k \leq t_F. \tag{2.13}$$

Often the evaluation points are arranged monotonically, that is, $\theta_k \leq \theta_{k+1}$. It is also common to have many residuals evaluated at the same time, e.g., $\theta_k = \theta_{k+1}$. Although neither of these assumptions is necessary for our approach, we do require that the initial and final times t_I and t_F be fixed.

It is worth noting that more complicated problem descriptions can be accommodated by the formulation given. For example, suppose it is required to minimize the expression

$$F = \frac{1}{2} \sum_{k=1}^{N} [\mathbf{h}(\mathbf{y}(\theta_k), \mathbf{u}(\theta_k), \mathbf{p}, \theta_k) - \widehat{\mathbf{h}}_k]^{\mathsf{T}} \mathbf{\Lambda} [\mathbf{h}(\mathbf{y}(\theta_k), \mathbf{u}(\theta_k), \mathbf{p}, \theta_k) - \widehat{\mathbf{h}}_k], \tag{2.14}$$

where $\widehat{\mathbf{h}}_k$ are the observed values of the function $\mathbf{h}$ at the times θ_k and $\mathbf{\Lambda}$ is the inverse covariance matrix of these quantities. Since the positive definite matrix can be factored as $\mathbf{\Lambda} = \mathbf{Q}^{\mathsf{T}}\mathbf{Q}$ we can define a new set of algebraic variables

$$\mathbf{z}(t) = \mathbf{Q}\mathbf{h}(\mathbf{y}(t), \mathbf{u}(t), \mathbf{p}, t) \tag{2.15}$$

and transform the observed data

$$\widehat{\mathbf{z}}_k = \mathbf{Q}\widehat{\mathbf{h}}_k. \tag{2.16}$$

The *maximum likelihood* objective function (2.14) then becomes

$$F = \frac{1}{2} \sum_{k=1}^{N} [\mathbf{z}_k - \widehat{\mathbf{z}}_k]^{\mathsf{T}} [\mathbf{z}_k - \widehat{\mathbf{z}}_k], \tag{2.17}$$

where the residuals have the form given by (2.12) and the transformation (2.15) can be treated as an equality path constraint as in (2.2).

This example suggests that in general the discrete data may involve complicated expressions of the "real" state and algebraic variables $\mathbf{y}(t), \mathbf{u}(t)$ and the parameters $\mathbf{p}$. When this occurs the problem can be restated in terms of an augmented system. In the most common situation the *observation*

$$\mathbf{z}(t) = \mathbf{h}[\mathbf{y}(t), \mathbf{u}(t), t] \tag{2.18}$$

is treated as an (additional) algebraic constraint and it is natural to augment the "real" algebraic variable $\mathbf{u}(t)$ to include the additional algebraic variable $\mathbf{z}(t)$. On the other hand, if the observation is given by

$$\mathbf{z}(t) = \mathbf{h}[\mathbf{y}(t), t], \tag{2.19}$$

then it is possible to augment the "real" *state* variable $\mathbf{y}(t)$ to include the additional state $\mathbf{z}(t)$. In this case the state equations (2.1) must be augmented to include

$$\dot{\mathbf{z}}(t) = \mathbf{h}_y \dot{\mathbf{y}} + \dot{\mathbf{h}} = \mathbf{h}_y \mathbf{f} + \dot{\mathbf{h}}, \tag{2.20}$$

where the vector $\mathbf{h}_y \doteq (\partial h/\partial y_1, \ldots, \partial h/\partial y_n)$ is considered a row vector.

For the sake of simplicity we have not introduced problems with multiple "phases." Nevertheless, our software implementation [4] does not have these restrictions.

3. Transcription formulation. The basic approach for solving the optimal control problem by transcription has been presented in detail elsewhere [1]. For completeness we give a brief outline of the method. All approaches divide the time domain into n_s intervals

$$t_I = t_1 < t_2 < \cdots < t_M = t_F, \tag{3.1}$$

where the points are referred to as node, mesh, or grid points. Define the number of mesh points as $M \equiv n_s + 1$. Note that the grid points do not necessarily coincide with the data evaluation points given by the values θ_k. Let us introduce the notation $\mathbf{y}_j \equiv \mathbf{y}(t_j)$ to indicate the value of the state variable at a grid point. In like fashion denote the algebraic variable at a grid point by $\mathbf{u}_j \equiv \mathbf{u}(t_j)$. In addition some discretization schemes require values for the algebraic variable at the midpoint of an interval, and we denote this quantity by $\overline{\mathbf{u}}_j \equiv \mathbf{u}(\bar{t})$ with $\bar{t} = \frac{1}{2}(t_j + t_{j-1})$. Two primary discretization schemes will be considered, namely trapezoidal and Hermite–Simpson. Each scheme produces a distinct set of nonlinear programming (NLP) variables and constraints.

For the trapezoidal discretization, the NLP variables are

$$\mathbf{x}^{\mathsf{T}} = [\mathbf{y}_1, \mathbf{u}_1, \mathbf{y}_2, \mathbf{u}_2, \ldots, \mathbf{y}_M, \mathbf{u}_M, \mathbf{p}]. \tag{3.2}$$

The state equations (2.1) are approximately satisfied by solving the *defect* constraints

$$\boldsymbol{\zeta}_j = \mathbf{y}_{j+1} - \mathbf{y}_j - \frac{h_j}{2}[\mathbf{f}_{j+1} + \mathbf{f}_j] = 0 \tag{3.3}$$

for $j = 1, \ldots, n_s$. The step size is denoted by $h_j \equiv t_{j+1} - t_j$, and the right-hand side of the differential equations (2.1) are given by $\mathbf{f}_j \equiv \mathbf{f}[\mathbf{y}(t_j), \mathbf{u}(t_j), \mathbf{p}, t_j]$.

For the Hermite–Simpson discretization scheme, the NLP variables are

$$\mathbf{x}^{\mathsf{T}} = [\mathbf{y}_1, \mathbf{u}_1, \overline{\mathbf{u}}_2, \mathbf{y}_2, \mathbf{u}_2, \overline{\mathbf{u}}_3, \ldots, \mathbf{y}_M, \mathbf{u}_M, \mathbf{p}]. \tag{3.4}$$

The defects for this discretization are given by

$$\boldsymbol{\zeta}_j = \mathbf{y}_{j+1} - \mathbf{y}_j - \frac{h_j}{6}\left[\mathbf{f}_{j+1} + 4\overline{\mathbf{f}}_{j+1} + \mathbf{f}_j\right], \tag{3.5}$$

where

$$\overline{\mathbf{f}}_{j+1} = \mathbf{f}[\overline{\mathbf{y}}_{j+1}, \overline{\mathbf{u}}_{j+1}, \bar{t}] \tag{3.6}$$

with

$$\overline{\mathbf{y}}_{j+1} = \frac{1}{2}\left[\mathbf{y}_j + \mathbf{y}_{j+1}\right] + \frac{h_j}{8}[\mathbf{f}_j - \mathbf{f}_{j+1}] \tag{3.7}$$

for $j = 1, \ldots, n_s$.

For the Hermite–Simpson discretization scheme written in separated form, the NLP variables are

$$\mathbf{x}^{\mathsf{T}} = (\mathbf{y}_1, \mathbf{u}_1, \overline{\mathbf{y}}_2, \overline{\mathbf{u}}_2, \mathbf{y}_2, \mathbf{u}_2, \ldots, \overline{\mathbf{y}}_M, \overline{\mathbf{u}}_M, \mathbf{y}_M, \mathbf{u}_M, \mathbf{p}). \tag{3.8}$$

For this discretization, the defect constraints $\boldsymbol{\zeta}_j = \mathbf{0}$ are given by

$$\mathbf{0} = \overline{\mathbf{y}}_{j+1} - \frac{1}{2}(\mathbf{y}_{j+1} + \mathbf{y}_j) - \frac{h_j}{8}(\mathbf{f}_j - \mathbf{f}_{j+1}), \tag{3.9}$$

$$\mathbf{0} = \mathbf{y}_{j+1} - \mathbf{y}_j - \frac{h_j}{6}\left[\mathbf{f}_{j+1} + 4\overline{\mathbf{f}}_{j+1} + \mathbf{f}_j\right]. \tag{3.10}$$

When there are no algebraic constraints the discretization error is $\mathcal{O}(h^p)$, where $p = 2$ for the trapezoidal scheme and $p = 4$ for the Hermite–Simpson methods. However, when path constraints are active the order can be reduced, and the ultimate choice of the discretization scheme is determined by a number of conflicting criteria. For a more complete discussion of the mesh refinement procedure the reader should consult [1].

As a result of the transcription, dynamic constraints (2.1)–(2.2) are replaced by the NLP constraints

$$\mathbf{c}_L \le \mathbf{c}(\mathbf{x}) \le \mathbf{c}_U, \tag{3.11}$$

where the m-vector

$$\mathbf{c}(\mathbf{x}) = \left[\boldsymbol{\zeta}_1, \boldsymbol{\zeta}_2, \ldots, \boldsymbol{\zeta}_{M-1}, \boldsymbol{\psi}_I, \boldsymbol{\psi}_F, \mathbf{g}_1, \mathbf{g}_2, \ldots, \mathbf{g}_M\right]^{\top} \tag{3.12}$$

with

$$\mathbf{c}_L = [\mathbf{0}, \ldots, \mathbf{0}, \mathbf{g}_L, \ldots, \mathbf{g}_L]^{\top} \tag{3.13}$$

and a corresponding definition of $\mathbf{c}_U$. The first $n_y n_s$ equality constraints require that the defect vectors from each of the n_s segments be zero, thereby approximately satisfying the differential equations (2.1). The boundary conditions are enforced directly by the equality constraints on $\boldsymbol{\psi}$, and the nonlinear path constraints are imposed at the grid points. Note that nonlinear equality path constraints are enforced by setting $\mathbf{c}_L = \mathbf{c}_U$. In a similar fashion the state and algebraic variable bounds (2.6) and (2.7) become simple bounds on the NLP variables. The path constraints and variable bounds are always imposed at the grid points, and for the Hermite–Simpson discretization the path constraints and variable bounds are also imposed at the interval midpoints.

4. Parameter estimation algorithm. There are three primary operations that are performed when solving a parameter estimation problem using a transcription method. Briefly the approach is as follows:

Direct transcription. Transcribe the parameter estimation problem into a NLP problem by discretization.

Sparse nonlinear program. Solve the sparse NLP using sequential quadratic programming.

Mesh refinement. Assess the accuracy of the approximation (i.e., the finite dimensional problem), and if necessary refine the discretization, and then repeat the optimization steps.

The NLP problem can be stated as follows: Find the n-vector $\mathbf{x}$ defined by (3.2), (3.4), or (3.8) which minimizes the objective function (2.10) subject to the constraints (3.11). This large, sparse NLP can be solved efficiently using a sequential quadratic programming (SQP) method as described in [1, 2, 3]. Optimal control problems have been solved using similar techniques (cf. [6, 9, 11]). Although it is not necessary to employ an SQP method as we do in this paper, it is very important to exploit the nonlinear least squares form of the objective function for efficiency in the NLP algorithm. Specifically, the $\ell \times n$ *residual Jacobian* matrix $\mathbf{R}$ is defined by

$$\mathbf{R}^{\mathsf{T}} = [\nabla \mathbf{r}_1, \ldots, \nabla \mathbf{r}_\ell] \tag{4.1}$$

and the gradient vector is

$$\nabla F = \mathbf{R}^{\mathsf{T}} \mathbf{r} = \sum_{i=1}^{\ell} r_i \nabla \mathbf{r}_i. \tag{4.2}$$

And finally, the Hessian of the Lagrangian is given by

$$\mathbf{H}_L(\mathbf{x}, \boldsymbol{\lambda}) = \sum_{i=1}^{\ell} r_i \nabla^2 r_i - \sum_{i=1}^{m} \lambda_i \nabla^2 c_i + \mathbf{R}^{\mathsf{T}} \mathbf{R} \tag{4.3}$$

$$\equiv \mathbf{V} + \mathbf{R}^{\mathsf{T}} \mathbf{R}. \tag{4.4}$$

The matrix $\mathbf{R}^{\mathsf{T}}\mathbf{R}$ is referred to as the *normal matrix* and the matrix $\mathbf{V}$ is referred to as the *residual Hessian*. Our NLP algorithm uses full second order information constructed using the sparse finite difference technique discussed in section 6. Consequently it converges quadratically, even for problems with nonlinear residuals and/or nonzero sum of squares. In contrast, the widely used Gauss method does not utilize the residual Hessian ($\mathbf{V} = \mathbf{0}$), and for this reason it converges at a linear rate unless either $F(\mathbf{x}^*) = 0$ or the residuals are linear functions of $\mathbf{x}$.

5. Computing the residuals. In order to evaluate the residuals (2.11) it is necessary to compute the value of the state variable at the data evaluation time θ_k as illustrated in Figure 1. This quantity can be constructed from the Hermite interpolating polynomial. Thus for any particular residual k there is an interval $t_j \le \theta_k \le t_{j+1}$, and a particular state $\nu = i(k)$. Then the value of the state needed in the residual calculation is

$$\begin{aligned} y_\nu(\theta_k) &= (1 - 3\delta^2 + 2\delta^3) y_{\nu j} + (3\delta^2 - 2\delta^3) y_{\nu,j+1} \\ &\quad + (h_j\delta - 2h_j\delta^2 + h_j\delta^3) f_{\nu j} + (-h_j\delta^2 + h_j\delta^3) f_{\nu,j+1}, \end{aligned} \tag{5.1}$$

where $h_j = t_{j+1} - t_j$ is the length of the discretization interval and $\delta = (\theta_k - t_j)/h_j$ defines the location of the evaluation time relative to the beginning of the interval. In this expression, $y_{\nu j}$ is the value of state variable ν at grid point j and $f_{\nu j}$ is the corresponding value of the right-hand side (2.1) at the same grid point. The value of the algebraic variable required in (2.12) can be constructed from a quadratic interpolant when the Hermite–Simpson discretization is used, i.e., according to

$$u_\nu(\theta_k) = (1 - \delta)(1 - 2\delta) u_{\nu j} + 4\delta(1 - \delta)\bar{u}_{\nu,j+1} - \delta(1 - 2\delta) u_{\nu,j+1}, \tag{5.2}$$

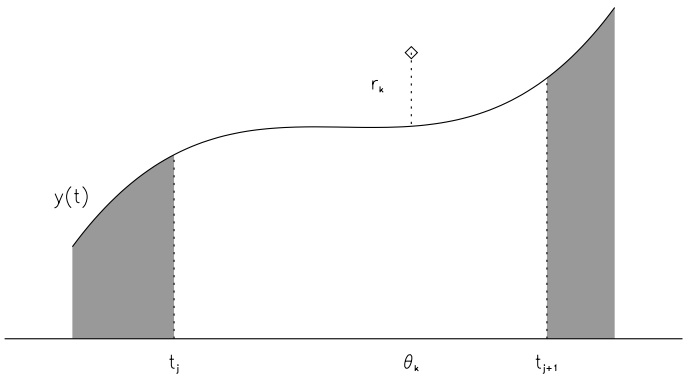


FIG. 1. *Residual evaluation.*

where $\delta = (\theta_k - t_j)/h_j$. Similarly, when a trapezoidal discretization is used, linear interpolation between the grid points yields

$$u_\nu(\theta_k) = (1-\delta)u_{\nu j} + \delta u_{\nu,j+1}. \tag{5.3}$$

It is important to observe that the residuals are computed by interpolation and do not have any direct effect on the location of the discretization grid points. This is often referred to as *dense output* in methods for numerical integration (cf. [8]). It is worth emphasizing another property of the interpolation scheme. In each of the expressions (5.1), (5.2), and (5.3) the interpolated value is written as a linear combination of the NLP variables and the right-hand side functions $\mathbf{f}$ at the grid points. In particular the quantities h_j are fixed by the mesh refinement procedure, and the quantities δ are fixed by the location of the discrete data points within a mesh. Thus, within a particular mesh refinement step, the coefficients defining the interpolants are *constant* during the NLP optimization iterations. For example, the term $(-h_j\delta^2 + h_j\delta^3)$ in (5.1) remains unchanged by the NLP variables. This will be exploited when constructing derivatives as described in section 6.

6. Computing derivatives. First and second derivatives are constructed by exploiting the sparse finite differencing techniques described in [1, sects. 2.10.3 and 4.6.8]. The key notion is to write the complete set of transcribed NLP functions as

$$\begin{bmatrix} \mathbf{c}(\mathbf{x}) \\ \mathbf{r}(\mathbf{x}) \end{bmatrix} = \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{q}(\mathbf{x}) + \boldsymbol{\zeta}, \tag{6.1}$$

where $\mathbf{A}$ and $\mathbf{B}$ are matrices (constant during the NLP) and $\mathbf{q}$ involves the nonlinear functions at grid points. The vector $\boldsymbol{\zeta}$ is typically zero for defect constraints. Similar information for the nonlinear boundary functions $\boldsymbol{\psi}$ can also be incorporated. We then

construct finite difference estimates for the first derivatives of the set of υ functions $q_i(\mathbf{x})$ with respect to the n variables $\mathbf{x}$ in the $\upsilon \times n$ matrix

$$\mathbf{D} \equiv \begin{bmatrix} (\nabla q_1)^\top \\ (\nabla q_2)^\top \\ \vdots \\ (\nabla q_\upsilon)^\top \end{bmatrix} = \frac{\partial \mathbf{q}}{\partial \mathbf{x}}. \tag{6.2}$$

The efficiency of the differencing technique depends on the sparsity of the matrix $\mathbf{D}$ (cf. [1, sect. 2.2.1]). The columns of $\mathbf{D}$ can be partitioned into subsets called *index sets* such that each subset has at most one nonzero element per *row*. Derivatives are constructed by perturbing all variables in an index set at the same time, and consequently the number of perturbations needed to construct $\mathbf{D}$ can be much smaller than the number of variables n. In our software, we construct this problem dependent *sparsity template* information by random sampling of the user functions. From the sparsity template information, it is possible to construct the sparsity for the matrix $\mathbf{D}$ and compute the finite difference index sets. The first derivative information needed to solve the NLP can then be computed from

$$\begin{bmatrix} \mathbf{G} \\ \mathbf{R} \end{bmatrix} = \mathbf{A} + \mathbf{BD}, \tag{6.3}$$

where $\mathbf{G}$ is the Jacobian of the constraints and $\mathbf{R}$ is the residual Jacobian.

If we define

$$\boldsymbol{\omega}^\mathsf{T} = (-\lambda_1, \ldots, -\lambda_m, r_1, \ldots, r_\ell), \tag{6.4}$$

where λ_k are the Lagrange multipliers with $\upsilon = m + \ell$, then we can also utilize sparse differencing to compute second derivatives of the function

$$\Omega(\mathbf{x}) = \sum_{i=1}^{\upsilon} \omega_i q_i(\mathbf{x}) = -\sum_{i=1}^{m} \lambda_i c_i(\mathbf{x}) + \sum_{i=1}^{\ell} [r_i] r_i(\mathbf{x}). \tag{6.5}$$

Note that elements of $\boldsymbol{\omega}$ are not perturbed during the finite difference operation. To emphasize this, we have written the second term above as $[r_i]r_i(\mathbf{x})$ since the quantities $[r_i]$ do *not* change during the perturbations. Then, it follows that the residual Hessian in (4.4) is given by

$$\mathbf{V} \equiv \nabla^2 \Omega(\mathbf{x}) = \sum_{i=1}^{\upsilon} \omega_i \nabla^2 q_i(\mathbf{x}). \tag{6.6}$$

It is also easy to demonstrate that the sparsity pattern for the NLP Hessian is a subset of the sparsity for the matrix $(\mathbf{BD})^\mathsf{T}(\mathbf{BD})$, which can be constructed from the known sparsity of $\mathbf{D}$.

Let us now present the details of the decomposition (6.1) for the various terms. To express a state residual given by (2.11) and (5.1) in the decomposed form we write

$$\begin{aligned}
\mathbf{r}_k(\mathbf{x}) &= \mathbf{A}_{k+m}\mathbf{x} + \mathbf{B}_{k+m}\mathbf{q}(\mathbf{x}) + \boldsymbol{\zeta}_{k+m} \\
&= \begin{bmatrix} \mathbf{0} & w_k(1-3\delta^2+2\delta^3) & \mathbf{0} & w_k(3\delta^2-2\delta^3) & \mathbf{0} \end{bmatrix} \begin{bmatrix} \cdot \\ y_{\nu j} \\ \cdot \\ y_{\nu,j+1} \\ \cdot \end{bmatrix} \\
&\quad + \begin{bmatrix} \mathbf{0} & w_k(h_j\delta - 2h_j\delta^2 + h_j\delta^3) & \mathbf{0} & w_k(-h_j\delta^2 + h_j\delta^3) & \mathbf{0} \end{bmatrix} \begin{bmatrix} \cdot \\ f_{\nu j} \\ \cdot \\ f_{\nu,j+1} \\ \cdot \end{bmatrix} \\
&\quad - w_k\hat{y}_{\nu j}. \qquad (6.7)
\end{aligned}$$

Observe that there are only two nonzero values in row $(k+m)$ of the matrices $\mathbf{A}$ and $\mathbf{B}$ denoted by $\mathbf{A}_{k+m}$ and $\mathbf{B}_{k+m}$, respectively. The nonlinear portions of the residual have been isolated in the vector $\mathbf{q}$. Furthermore, the problem dependent sparsity of the nonlinear quantities can be exploited because of separability; i.e., there is no interdependence between grid points. Finally, it should be clear that the algebraic residuals (2.12) can also be written in the separable form required by (6.1) using either the quadratic (5.2) or the linear (5.3) interpolant.

The user can exploit the benefits of sparsity by utilizing the separable form for the algebraic equations. Specifically an algebraic constraint function $g[\mathbf{y}(t), \mathbf{u}(t), \mathbf{p}, t]$ as given in (2.3) can be expressed as

$$\begin{aligned}
\mathbf{c}_k(\mathbf{x}) &= \mathbf{A}_k\mathbf{x} + \mathbf{B}_k\mathbf{q}(\mathbf{x}) + \boldsymbol{\zeta}_k \\
&= \begin{bmatrix} \mathbf{0} & \boldsymbol{\alpha}^{\mathsf{T}} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \cdot \\ \mathbf{y}_j \\ \mathbf{u}_j \\ \mathbf{p} \\ \cdot \end{bmatrix} + \begin{bmatrix} \mathbf{0} & \beta_0 & , \dots , & \beta_{n_a} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \cdot \\ a_0(t_j) \\ \cdots \\ a_{n_a}(t_j) \\ \cdot \end{bmatrix}. \qquad (6.8)
\end{aligned}$$

Here, $\mathbf{A}_k = [\mathbf{0}, \boldsymbol{\alpha}^{\mathsf{T}}, \mathbf{0}]$, $\mathbf{B}_k = [\mathbf{0}, \beta_0, \dots, \beta_{n_a}, \mathbf{0}]$, and $\boldsymbol{\zeta}_k = 0$. In contrast to the decomposition of the residual specified by (6.7), which can be performed algorithmically, this formulation must be given by the user. Nevertheless, the efficiency improvements are similar. Again, the key notion is to define the vector $\mathbf{q}$ which is differentiated so that the nonlinearities are isolated and involve quantities at a single grid point.

Ultimately the software implementation must compute derivatives of user supplied quantities via sparse finite differences. However, the user can reduce the cost of finite differencing by exploiting separability in the functions. To illustrate this point consider three different, yet mathematically equivalent, formulations of the same problem. Suppose the dynamic system has one state variable y, one control variable u, and one parameter p that satisfy the DAE system:

$$\begin{aligned}
\dot{y} &= f(u), \\
0 &= g(y, u, p), \qquad (6.9)
\end{aligned}$$

where $g(y, u, p) \equiv b_0(y) + u + b_1(p)$. There is some flexibility in how to group the terms in the path constraint when constructing the expression $g(y, u, p) = \boldsymbol{\alpha}^{\mathsf{T}}\mathbf{v} + \boldsymbol{\beta}^{\mathsf{T}}\mathbf{a}[\mathbf{v}, t]$.

One approach is to ignore separability and simply compute the terms enclosed between "{" and "}" together, i.e., compute $g(y,u,p) = \{b_0(y)+u+b_1(p)\}$. With this approach we define the quantities in the path constraint function (2.3) as follows:

$$\begin{aligned} \boldsymbol{\alpha}^{\mathsf{T}} &= (0,0,0), \\ n_a &= 0, \\ a_0(y,u,p) &= b_0(y) + u + b_1(p), \\ \boldsymbol{\beta}^{\mathsf{T}} &= (1). \end{aligned} \tag{6.10}$$

For this formulation the user must compute the functions f and a_0, and the matrix $\mathbf{D}$ will contain repeated blocks with the sparsity template

$$struct \begin{pmatrix} \frac{\partial f}{\partial y} & \frac{\partial f}{\partial u} & \frac{\partial f}{\partial p} \\ \frac{\partial a_0}{\partial y} & \frac{\partial a_0}{\partial u} & \frac{\partial a_0}{\partial p} \end{pmatrix} = \begin{bmatrix} 0 & \times & 0 \\ \times & \times & \times \end{bmatrix}.$$

Since the rows of the matrix $\mathbf{D}$ corresponding to the path constraint will have three nonzero elements, this formulation will require three index sets and hence three perturbations to compute a finite difference approximation for $\mathbf{D}$.

A second alternative is to compute the first two terms together and explicitly identify an auxiliary function, i.e., $g(y,u,p) = \{b_0(y)+u\} + \{b_1(p)\}$. Here we define

$$\begin{aligned} \boldsymbol{\alpha}^{\mathsf{T}} &= (0,0,0), \\ n_a &= 1, \\ a_0(y,u,p) &= b_0(y) + u, \\ a_1(y,u,p) &= b_1(p), \\ \boldsymbol{\beta}^{\mathsf{T}} &= (1,1). \end{aligned} \tag{6.11}$$

Since the user must compute the functions f, a_0, and a_1 individually the corresponding sparsity template will have the form

$$struct \begin{pmatrix} \frac{\partial f}{\partial y} & \frac{\partial f}{\partial u} & \frac{\partial f}{\partial p} \\ \frac{\partial a_0}{\partial y} & \frac{\partial a_0}{\partial u} & \frac{\partial a_0}{\partial p} \\ \frac{\partial a_1}{\partial y} & \frac{\partial a_1}{\partial u} & \frac{\partial a_1}{\partial p} \end{pmatrix} = \begin{bmatrix} 0 & \times & 0 \\ \times & \times & 0 \\ 0 & 0 & \times \end{bmatrix}.$$

Using this formulation the finite difference derivatives can be computed using two perturbations.

A third alternative is to define

$$\begin{aligned} \boldsymbol{\alpha}^{\mathsf{T}} &= (0,1,0), \\ n_a &= 1, \\ a_0(y,u,p) &= b_0(y), \\ a_1(y,u,p) &= b_1(p), \\ \boldsymbol{\beta}^{\mathsf{T}} &= (1,1). \end{aligned} \tag{6.12}$$

Here we explicitly identify both the analytic term and the auxiliary function. Since

the sparsity template is

$$struct\begin{pmatrix} \frac{\partial f}{\partial y} & \frac{\partial f}{\partial u} & \frac{\partial f}{\partial p} \\ \frac{\partial a_0}{\partial y} & \frac{\partial a_0}{\partial u} & \frac{\partial a_0}{\partial p} \\ \frac{\partial a_1}{\partial y} & \frac{\partial a_1}{\partial u} & \frac{\partial a_1}{\partial p} \end{pmatrix} = \begin{bmatrix} 0 & \times & 0 \\ \times & 0 & 0 \\ 0 & 0 & \times \end{bmatrix},$$

this formulation requires only one perturbation to compute the finite difference approximation for $\mathbf{D}$.

This example illustrates the need for a more general software interface with the user. Typically when solving a semiexplicit DAE such as (6.9), the user must provide a subroutine to compute the "right-hand side" functions $f(y,u,p,t)$ and $g(y,u,p,t)$ for given values of the arguments (y,u,p,t). However, to fully exploit sparsity our software implementation requires the user to compute the augmented set of right-hand side functions $f(y,u,p,t)$ and $a_k(y,u,p,t)$ for $k = 0,\ldots,n_a$. Nevertheless, a twofold computational benefit is observed by exploiting separability. First, the Hessian matrix is usually more sparse since it is determined by the structure of $\mathbf{(BD)}^{\mathsf{T}}(\mathbf{BD})$. This leads to computational savings when solving the linear systems required by the NLP algorithm. Second, since gradient information can be computed with fewer perturbations, it is not necessary to call the user function routines as many times, leading to additional computational savings.

There are a number of aspects of the approach that deserve emphasis. First, because the grid points (3.1) do not necessarily coincide with the data evaluation points (2.13), the sparsity pattern of the matrices $\mathbf{R}$ and $\mathbf{V}$ do not have a simple block form. Second, the grid points are placed to efficiently control the discretization error by the mesh refinement procedure. However, the data points at θ_k do not have any direct relation to the grid points at t_j. In essence the numerical integration of the differential equations is not controlled by the observation data. This also has an impact on the sparsity of the residual Jacobian and Hessian as illustrated in Figure 2. In this illustration, when the mesh includes the points at t_j and t_{j+1}, the partial derivative of the residual $r_k = w_k\,[y_\nu(\theta_k) - \hat{y}_{\nu j}]$ with respect to the state at the left grid point is nonzero, i.e.,

$$\frac{\partial r_k}{\partial y(t_j)} \neq 0.$$

However, when the mesh is refined by adding a new grid point at $t_r = \frac{1}{2}(t_{j+1} + t_j)$, we find that

$$\frac{\partial r_k}{\partial y(t_r)} \neq 0 \quad \text{but} \quad \frac{\partial r_k}{\partial y(t_j)} = 0.$$

Thus, mesh refinement alters the sparsity pattern of the residual Jacobian and Hessian matrices. Although it is more complicated to implement the construction of the sparsity pattern, there is no apparent impact on the solution times.

7. Computational experience.

7.1. Test examples. A collection of test examples from the literature have been used to check the performance and behavior of the algorithm. One example described by Bock [7] as a "notorious test problem" was originally introduced by Bulirsch [5].

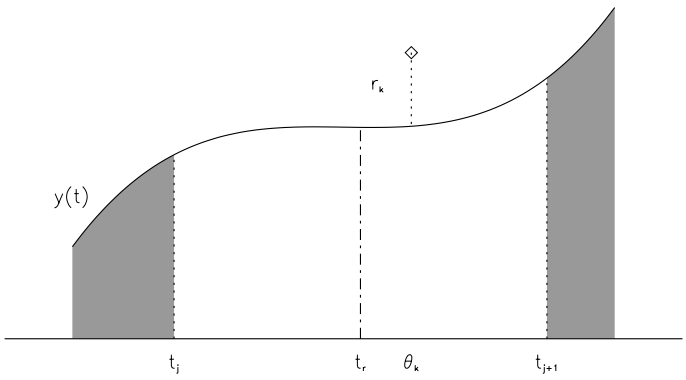


FIG. 2. *Mesh refinement alters sparsity.*

The differential equations are

$$\dot{y}_1 = y_2, \tag{7.1}$$

$$\dot{y}_2 = \mu^2 y_1 - (\mu^2 + p^2)\sin(pt) \tag{7.2}$$

with $y_1(0) = 0$, $y_2(0) = \pi$, $\mu = 60$, and $0 \le t \le 1$. It is easily verified that if the parameter $p = \pi$, then the corresponding analytic solution to (7.2) is given by

$$y_1 = \sin(\pi t), \tag{7.3}$$

$$y_2 = \pi \cos(\pi t). \tag{7.4}$$

Data for this problem can be constructed by evaluating the true solution at the data points θ_k and then adding normally distributed random variables with mean zero, and standard deviation $\sigma = .05$. It is easy to demonstrate that the optimal value of the objective function $F^* \approx \ell\sigma^2$, where ℓ is the total number of residuals. This deceptively simple example is extremely difficult to solve using any type of shooting method, because the differential equations are unstable. In contrast, the parameter estimation process using direct transcription is very well behaved. Furthermore, we can use the example to demonstrate two major features of the new algorithm, namely,

- the grid distribution is not determined by the data location, and
- the algorithm converges quadratically for nonzero, nonlinear residuals.

Consider three different cases:

1. for $k = 1, 10$ select $\theta_k = .1k$;
2. for $k = 1, 2000$ select θ_k as a uniformly distributed random variable in the region $0 \le \theta_k \le 1$; and
3. for $k = 1, 10$ select $\theta_k = .1k$; for $k = 11, 2000$ select θ_k as a normally distributed random variable with mean = .4, and standard deviation = .1.

The first case has a relatively small number of residuals, and as such a small, albeit, nonzero objective at the solution. The second case has a large amount of data spread

TABLE 1
Algorithm performance summary.

Case	1	2	3
No. mesh it.	5	4	5
No. grid pt.	92	73	91
No. NLP it.	23	20	23
F^*	.030372674	2.6563759	2.4887179
$\lvert p^* - \pi \rvert$	3.6×10^{-8}	4.7×10^{-8}	3.6×10^{-8}

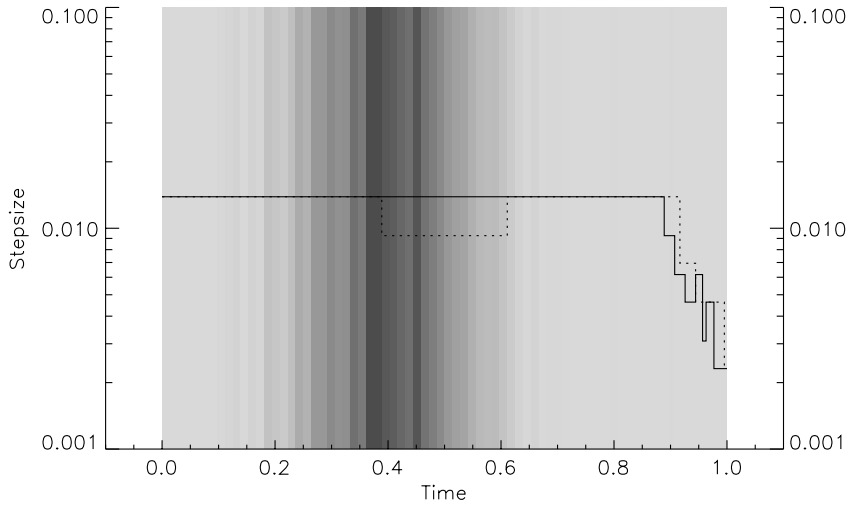


FIG. 3. *Stepsize history (case* 1: *dashed; case* 3: *solid).*

over the entire domain, whereas the third case has lots of data clustered in only one portion of the time domain near $t = .4$. Table 1 summarizes the performance of the algorithm. Notice that the number of mesh refinement iterations, grid points, and NLP iterations is essentially the same for all three cases. This occurs even though the objective function is significantly nonzero at the solution. Furthermore, the optimal parameter estimate is quite good, especially since the discretization error tolerance was also 10^{-7}.

Figure 3 illustrates the final mesh distribution for case 1 and case 3. The solid line plots the stepsize history as a function of time for case 3. The shading illustrates the distribution of data points over the domain—the darkest representing the highest concentration. The stepsize for case 1 is shown with dotted lines. Even though case 1 has only 11 points evenly spread over the time domain, and case 3 has 2000 data points clustered near $t = .4$, the final mesh distribution is nearly identical. Obviously for this example the location of the discrete data does not directly influence the location of the mesh points because the stepsize is constructed to control error in the differential equation. We have observed similar behavior of the mesh refinement procedure in other examples not reported here.

7.2. Reentry trajectory reconstruction. A problem of some practical interest occurs when attempting to reconstruct the trajectory of an object as it reenters the earth's atmosphere using information from radar observations. Let us consider

a nonlifting body of unknown size, shape, and mass, reentering the atmosphere over an oblate rotating earth. The translational motion is described by the differential equations:

$$\dot{\mathbf{r}} = \mathbf{v}, \tag{7.5}$$

$$\dot{\mathbf{v}} = -D\frac{\mathbf{v}_r}{\|\mathbf{v}_r\|} + \mathbf{g}(\mathbf{r}), \tag{7.6}$$

where $\mathbf{r}^\mathsf{T} = (x, y, z)$ is the earth centered inertial (ECI) position vector, $\mathbf{v}^\mathsf{T} = (\dot{x}, \dot{y}, \dot{z})$ is the ECI velocity vector, and $\mathbf{g}(\mathbf{r})$ is the gravitational acceleration. An oblate earth model including the first four zonal harmonics is used. The earth relative velocity vector is defined by

$$\mathbf{v}_r = \mathbf{v} - \boldsymbol{\omega} \times \mathbf{r}, \tag{7.7}$$

where $\boldsymbol{\omega}^\mathsf{T} = (0, 0, \omega)$ is the earth rotation rate vector. The drag on the object is given by

$$D = \frac{g_0 \rho \|\mathbf{v}_r\|^2}{2\beta}, \tag{7.8}$$

where $\rho(h)$ is the atmospheric density as a function of the altitude above the oblate spheroid, $g_0 = 32.174$, and β is the *ballistic coefficient*. For this application the atmospheric density is computed using a cubic spline approximation to the 1962 Standard Atmosphere.

The goal is to reconstruct the position and velocity time history from radar information. Thus we would like to minimize

$$F = \frac{1}{2}\sum_{k=1}^{N} \mathbf{q}_k^\mathsf{T}\mathbf{q}_k \tag{7.9}$$

with

$$\mathbf{q}_k = \begin{bmatrix} (\psi_k - \hat{\psi}_k)/\sigma_1 \\ (\eta_k - \hat{\eta}_k)/\sigma_2 \\ (s_k - \hat{s}_k)/\sigma_3 \\ (\dot{s}_k - \hat{\dot{s}}_k)/\sigma_4 \end{bmatrix}, \tag{7.10}$$

where $\psi_k = \psi(\mathbf{r}(\theta_k), \mathbf{v}(\theta_k), \theta_k)$ is the azimuth angle from the radar site to the object evaluated at time θ_k, and $\hat{\psi}_k$ is the corresponding radar measurement data, with standard deviation σ_1. Similarly, η_k is the elevation, s_k is the slant range, and $\dot{s}_k$ is the (slant) range rate. In order to restate the problem involving residuals of the form (2.12) we introduce the algebraic variables (u_1, u_2, u_3, u_4) and the corresponding algebraic path equations

$$0 = \psi(\mathbf{r}, \mathbf{v}, t) - u_1(t), \tag{7.11}$$

$$0 = \eta(\mathbf{r}, \mathbf{v}, t) - u_2(t), \tag{7.12}$$

$$0 = s(\mathbf{r}, \mathbf{v}, t) - u_3(t), \tag{7.13}$$

$$0 = \dot{s}(\mathbf{r}, \mathbf{v}, t) - u_4(t). \tag{7.14}$$

After introducing the new algebraic variables it is clear that (7.9) can be rewritten in the form (2.12).

To complete the definition of the problem it is sufficient to describe how the radar quantities in (7.11)–(7.14) are computed. The position of the radar site at time t is given by

$$\mathbf{w}(t) = r_e \begin{bmatrix} \cos\theta_s \cos\varphi \\ \cos\theta_s \sin\varphi \\ \sin\theta_s \end{bmatrix}, \tag{7.15}$$

where θ_s is the geocentric latitude of the radar site, $\varphi = \phi_s + \phi_0 + \omega t$ is the inertial longitude of the radar site, ϕ_s is the longitude of the radar site, and r_e is the radius to the site. The inertial velocity of the radar site is

$$\dot{\mathbf{w}}(t) = \begin{bmatrix} -\omega\left[\sin(\omega t) r_1 + \cos(\omega t) r_2\right] \\ \omega\left[\cos(\omega t) r_1 - \sin(\omega t) r_2\right] \\ 0 \end{bmatrix}. \tag{7.16}$$

The line-of-sight vector from the radar site to the vehicle is given by

$$\mathbf{s} = \mathbf{r} - \mathbf{w} \tag{7.17}$$

which yields the slant range

$$s(\mathbf{r}, \mathbf{v}, t) = \|\mathbf{s}\|. \tag{7.18}$$

The range rate is then given by

$$\dot{s}(\mathbf{r}, \mathbf{v}, t) = \frac{\mathbf{s}^{\mathsf{T}}(\mathbf{v} - \dot{\mathbf{w}})}{\|\mathbf{s}\|}. \tag{7.19}$$

The azimuth angle is given by

$$\psi(\mathbf{r}, \mathbf{v}, t) = \arctan\left[\frac{w_1 s_2 - w_2 s_1}{\left[(w_1^2 + w_2^2) s_3 - w_3(w_1 s_1 + w_2 s_2)\right] r_e^{-1}}\right]. \tag{7.20}$$

Now the local geodetic vertical direction at the radar site is

$$\mathbf{d} = \begin{bmatrix} \cos(\varphi)\cos(\theta_d) \\ \sin(\varphi)\cos(\theta_d) \\ \sin(\theta_d) \end{bmatrix}, \tag{7.21}$$

where θ_d is the geodetic latitude of the radar site, and the geodetic elevation is given by

$$\eta(\mathbf{r}, \mathbf{v}, t) = \frac{\pi}{2} - \arccos\left(\frac{\mathbf{d}^{\mathsf{T}}\mathbf{s}}{\|\mathbf{s}\|}\right). \tag{7.22}$$

7.2.1. Compton Gamma Ray Observatory reentry. On June 4, 2000 the NASA Compton Gamma Ray Observatory satellite reentered the atmosphere, and a portion of the trajectory was observed by the Kaena Point tracking station in Hawaii. The 17 ton spacecraft, one of the largest ever launched by NASA, was deliberately deorbited after one of the observatory's three attitude-control gyros failed in December 1999. The radar site provided azimuth, elevation, range, and range rate data for a portion of the trajectory above 70nm altitude during a time span of approximately four

TABLE 2
Mesh refinement summary.

k	M	NGC	NHC	NFE	NRHS	ϵ	Time (sec)
1	25	8	4	184	9016	1.24×10^{-2}	2.40
2	49	5	2	110	10670	8.27×10^{-4}	2.41
3	97	5	2	110	21230	4.02×10^{-5}	4.71
4	193	3	1	59	22715	2.40×10^{-6}	5.11
5	385	3	1	59	45371	1.53×10^{-7}	9.09
6	410	4	2	96	78624	9.91×10^{-8}	14.1
Total	410	28	12	618	187626		37.84

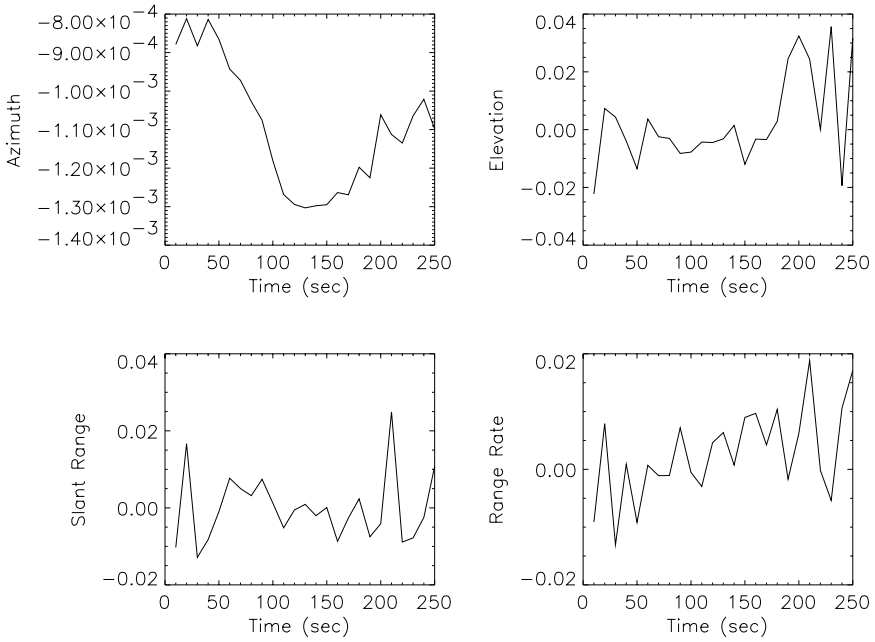


FIG. 4. *Normalized residual errors.*

minutes. The authors gratefully acknowledge the assistance provided by Dr. Wayne Hallman of The Aerospace Corporation concerning this example.

The parameter estimation method was used to reconstruct the reentry trajectory, and the results of the algorithm are summarized in Table 2. The algorithm began with 25 equally spaced grid points, and after six refinement iterations increased the number of points to 410 (cf. column 2). This refinement reduced the discretization error ϵ from 1.24×10^{-2} to 9.91×10^{-8} as shown in column 7. The number of gradient, Hessian, function, and right-hand side evaluations are given by the columns labeled NGC, NHC, NFE, and NRHS, respectively. Figure 4 displays the normalized error residuals, i.e., the components of $\mathbf{q}_k$ given by (7.10) for each set of data. The total normalized residual $\|\mathbf{q}_k^{\mathsf{T}}\mathbf{q}_k\|$ is plotted in Figure 5.

7.3. Commercial aircraft rotational dynamics analysis. When constructing a dynamic simulation of a commercial aircraft, flight test data is used to refine analytic models of the aerodynamic characteristics. A representative example of a parameter estimation problem occurs when attempting to estimate rotational accel-

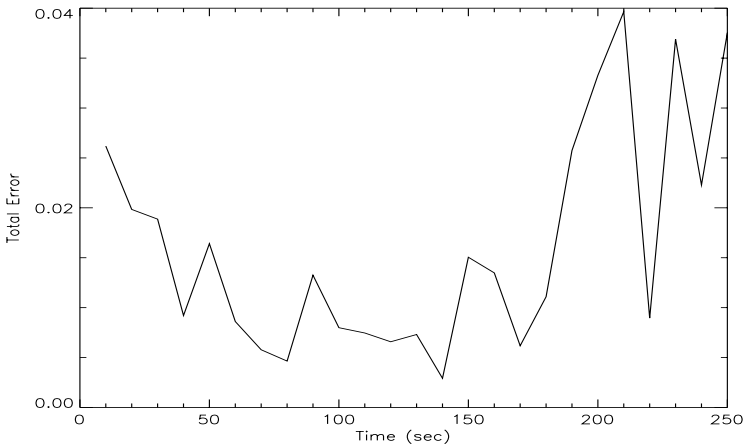


FIG. 5. *Total normalized residual error.*

erations from measured information about the aircraft orientation. We consider a particular maneuver called a "windup turn" for a 767-ER aircraft and gratefully acknowledge the contributions of Dr. Jia Luo of The Boeing Company for information related to this example. The rotational dynamics are described by

$$\dot{\phi} = p + q\frac{\sin\phi\sin\theta}{\cos\theta} + r\frac{\cos\phi\sin\theta}{\cos\theta}, \tag{7.23}$$

$$\dot{\theta} = q\cos\phi - r\sin\phi, \tag{7.24}$$

$$\dot{\psi} = q\frac{\sin\phi}{\cos\theta} + r\frac{\cos\phi}{\cos\theta}, \tag{7.25}$$

$$\dot{p} = a_p(\mathbf{b}), \tag{7.26}$$

$$\dot{q} = a_q(\mathbf{b}), \tag{7.27}$$

$$\dot{r} = a_r(\mathbf{b}), \tag{7.28}$$

where ϕ is the bank angle (rad), θ is the pitch angle (rad), ψ is the heading angle (rad), p is the roll rate (rad/sec), q is the pitch rate (rad/sec), and r is the yaw rate (rad/sec). During flight testing measurements of the bank, pitch, and heading angle are made; i.e., we have measured values $\hat{\phi}_k$, $\hat{\theta}_k$, and $\hat{\psi}_k$ at a sequence of time points—in this case 1841 values corresponding to measurements every .05 secs for 92 seconds. We would like to compute the unknown accelerations a_p, a_q, and a_r such that the objective

$$F = \frac{1}{2}\sum_{k=1}^{N}\left[\frac{\phi_k - \hat{\phi}_k}{\sigma_1}\right]^2 + \left[\frac{\theta_k - \hat{\theta}_k}{\sigma_2}\right]^2 + \left[\frac{\psi_k - \hat{\psi}_k}{\sigma_3}\right]^2 \tag{7.29}$$

is minimized, where the standard deviations on the data are given by σ_j. It should be clear that the residuals are of the form given by (2.11) with weights $w_k = 1/\sigma_\nu$. Note for this example the symbol θ is used to denote a state variable, and *not* an evaluation time as in (2.11).

There are many ways to parameterize the accelerations a_p, a_q, and a_r. Since the accelerations are smooth functions of time, a particularly effective approach is to

TABLE 3
Mesh refinement summary.

k	M	NGC	NHC	NFE	NRHS	ϵ	Time (sec)
1	200	10	1	105	39900	3.51×10^{-4}	13.0
2	380	3	1	38	28120	2.32×10^{-5}	6.19
3	740	3	1	38	55480	1.48×10^{-6}	10.9
4	1429	3	1	38	107844	9.39×10^{-8}	20.9
Total	1429	19	4	219	231344		51.07

utilize piecewise polynomial approximations. Let us introduce N_p *phases*, where the independent variable t for phase k is defined in the region $t_I^{(k)} \le t \le t_F^{(k)}$ and the phases are sequential, that is, $t_I^{(k+1)} = t_F^{(k)}$. In addition let us construct the beginning of the first phase to coincide with the beginning of the problem $t_I^{(1)} = 0$, and the end of the last phase to coincide with the end of the problem $t_F^{(N_p)} = 92$. If we treat the values of the acceleration and their slopes at the phase boundaries as parameters the accelerations within a phase are of the form

$$a(\mathbf{b}) = \mathcal{H}\left[a(t_I^{(k)}), \dot{a}(t_I^{(k)}), a(t_F^{(k)}), \dot{a}(t_F^{(k)})\right] \tag{7.30}$$

for $k = 1, \ldots, N_p$. In this expression the Hermite interpolation $\mathcal{H}$ is given by (5.1), with the appropriate definition of symbols. Finally, we require continuity and differentiability in the state variables and accelerations across the phase boundaries. It is worth noting that in general there are three distinct levels of discretization. Within a phase, there may be many grid points selected to satisfy the differential equation accuracy requirements. Furthermore, the data observation points may or may not coincide with the phase times and/or the differential equation grid. For this 20 phase example, $N_p = 20$ and the total number of parameters $\mathbf{p}$ is $n_p = 12N_p = 240$. The particular data set used for this illustration had $N = 1841$ data points or 5523 residuals in (7.29). A summary of the mesh refinement procedure is presented in Table 3. The process was initiated with 10 grid points per phase or a total of $M = 200$. The first NLP problem was solved after 10 gradient evaluations (NGC), and one Hessian evaluation (NHC), which required 39900 evaluations of the right-hand sides of the differential equations (NRHS). This problem was solved in 13 seconds of CPU time with a discretization error of $\epsilon = 3.51 \times 10^{-4}$. The mesh was refined three more times as tabulated in rows 2–4. The overall solution was obtained in 51.07 seconds, and required 1429 mesh points. Notice that only one Hessian evaluation was required for each NLP problem, even though the objective function is quite nonlinear and the optimal value $F^* = 1.715105 \times 10^{-2} \neq 0$.

From this information it is also possible to infer how the new approach compares with a more traditional shooting method. Suppose we assume that a fourth order Runge–Kutta scheme is used to integrate the trajectory (which requires four right-hand side evaluations per step), and there are 1429 steps (corresponding to the final grid size $M = 1429$). Then, the number of right-hand side evaluations (231344) required by the new approach is equivalent to $231344/(1429 \times 4) \approx 41$ integrated trajectories. In comparison at least 240 trajectories would be required just to compute a single finite difference gradient in a traditional shooting method! Furthermore, if a quasi-Newton method is used to optimize this function with 132 degrees of freedom, one would expect that at least 132 iterations (and gradient evaluations) would be required to converge. An estimate of the total number of trajectories for a traditional

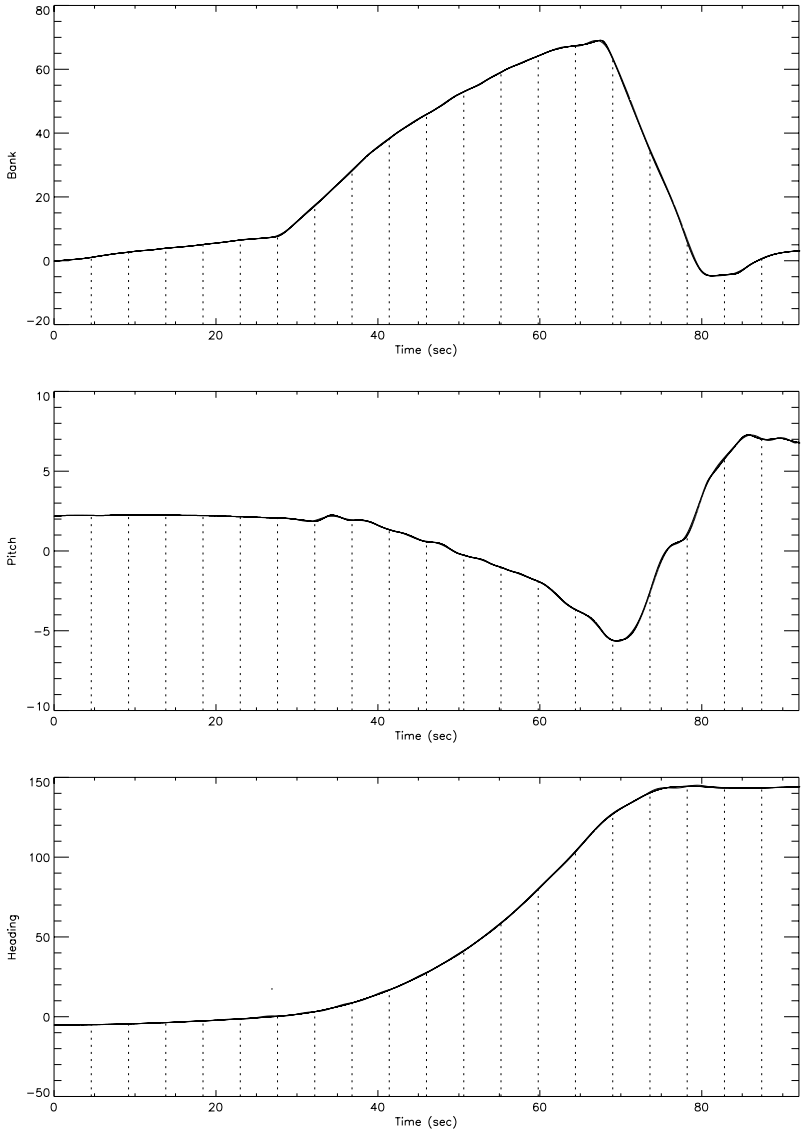


FIG. 6. *Angular variable history.*

shooting method is $(132 \times 240 = 31680)$. Thus, comparing the new versus the old algorithm suggests a ratio of $41 : 31680 \approx 1 : 773$. In short, a traditional shooting method would be extremely impractical for this application! The cost of computing first derivatives could be reduced somewhat for this problem by using a multiple shooting method; however, this approach still lacks quadratic convergence because it does not provide Hessian information.

Figure 6 presents the optimal time history for all of the angles as well as the data. Figure 7 illustrates the angular rates for the optimal solution and Figure 8 plots the corresponding accelerations. The phase boundaries are illustrated in all figures.

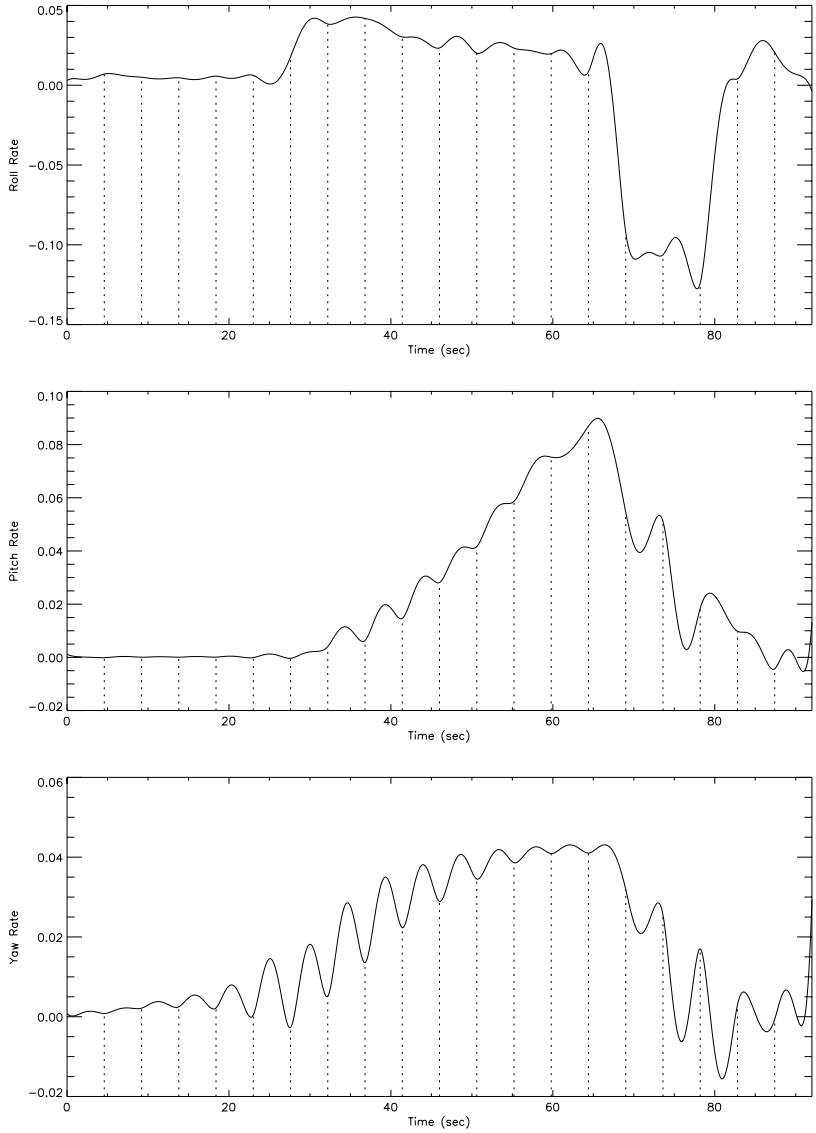


FIG. 7. *Angular rate history.*

8. Summary and conclusions. This paper describes an algorithm for parameter estimation that exploits state of the art methods for sparse NLP. The new method is unique because it exploits a full second order approximation to the Hessian matrix. As a consequence very large scale parameter estimation problems can be solved efficiently. In contrast, most traditional parameter estimation algorithms are based on a Gauss–Newton method, which is linearly convergent unless the residuals are linear and/or approach zero. In contrast to methods based on "shooting," the new algorithm constructs the relevant second order information without integrating the system dynamics. In short, the method is fast because it does not integrate the differential

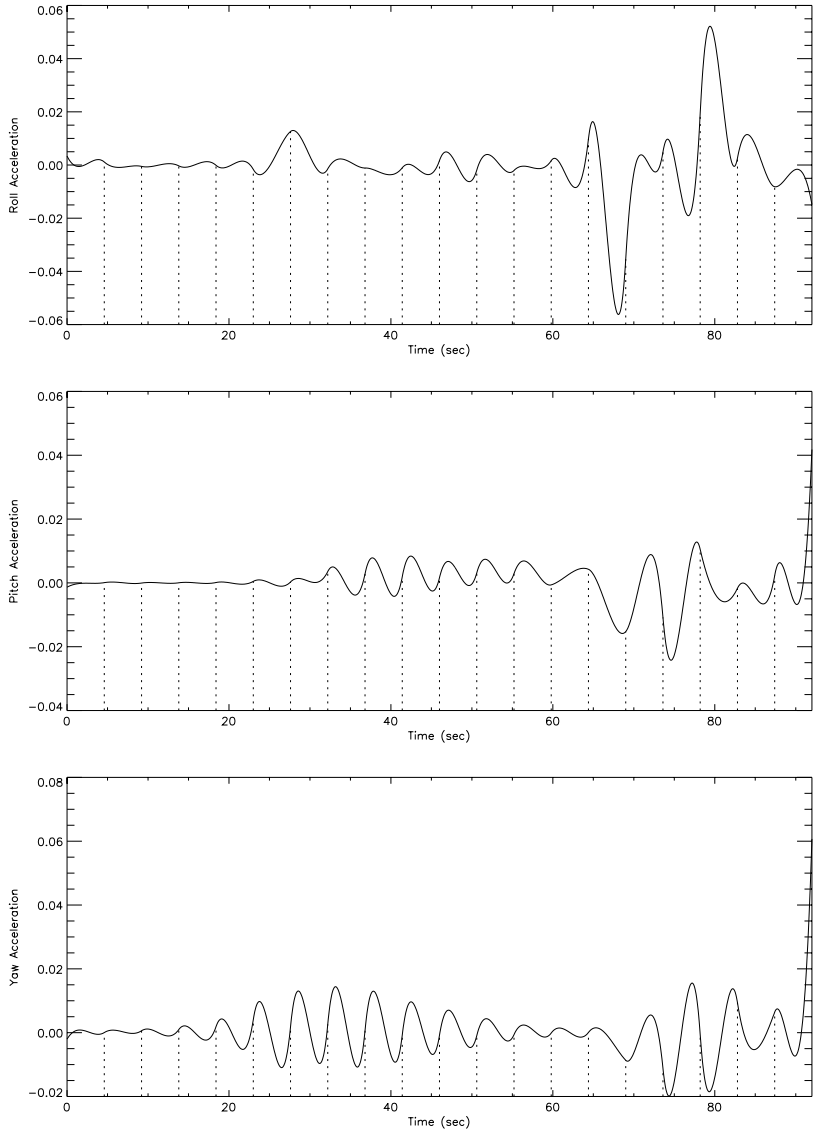


FIG. 8. *Angular acceleration history.*

equations but does use full second order information in the optimization process. The software implementation is illustrated on a set of realistic aerospace applications.

REFERENCES

[1] J. T. BETTS, *Practical Methods for Optimal Control Using Nonlinear Programming*, Adv. Des. Control 3, SIAM, Philadelphia, 2001.

[2] J. T. BETTS, M. J. CARTER, AND W. P. HUFFMAN, *Software for Nonlinear Optimization*, Mathematics and Engineering Analysis Library Report MEA-LR-83 R1, Boeing Information and Support Services, The Boeing Company, Seattle, WA, 1997.

[3] J. T. Betts and P. D. Frank, *A sparse nonlinear optimization algorithm*, J. Optim. Theory Appl., 82 (1994), pp. 519–541.

[4] J. T. Betts and W. P. Huffman, *Sparse Optimal Control Software* SOCS, Mathematics and Engineering Analysis Technical Document MEA-LR-085, Boeing Information and Support Services, The Boeing Company, Seattle, WA, 1997.

[5] R. Bulirsch, *Die Mehrzielmethode zur numerischen Lösung von nichtlinearen Randwertproblemen und Aufgaben der optimalen Steuerung*, Report of the Carl-Cranz Gesellschaft, Carl-Cranz Gesellschaft, Oberpfaffenhofen, Germany, 1971.

[6] P. J. Enright and B. A. Conway, *Discrete approximations to optimal trajectories using direct transcription and nonlinear programming*, AIAA Journal of Guidance, Control, and Dynamics, 15 (1992), pp. 994–1002.

[7] H. G. Bock, *Recent advances in parameter identification techniques for O.D.E.*, in Numerical Treatment of Inverse Problems in Differential and Integral Equations, P. Deuflhard and E. Hairer, eds., Birkhäuser Verlag, Heidelberg, 1982, pp. 95–121.

[8] E. Hairer, S. P. Norsett, and G. Wanner, *Solving Ordinary Differential Equations* I. *Nonstiff Problems*, Springer-Verlag, New York, 1993.

[9] C. R. Hargraves and S. W. Paris, *Direct trajectory optimization using nonlinear programming and collocation*, AIAA Journal of Guidance, Control, and Dynamics, 10 (1987), pp. 338–342.

[10] K. Schittkowski, *Parameter estimation in dynamic systems*, in Progress in Optimization, X. Yang, ed., Kluwer Academic Publishers, Dordrecht, The Netherlands, 2000, pp. 183–204.

[11] O. von Stryk, *Numerical solution of optimal control problems by direct collocation*, in Optimal Control, Internat. Ser. Numer. Math. 111, R. Bulirsch, A. Miele, J. Stoer, and K. H. Well, eds., Birkhäuser Verlag, Basel, 1993, pp. 129–143.

SIAM J. OPTIM.
Vol. 14, No. 1, pp. 245–267

NEW RESULTS ON QUADRATIC MINIMIZATION*

YINYU YE† AND SHUZHONG ZHANG‡

Abstract. In this paper we present several new results on minimizing an indefinite quadratic function under quadratic/linear constraints. The emphasis is placed on the case in which the constraints are two quadratic inequalities. This formulation is termed *the extended trust region subproblem* in this paper, to distinguish it from the ordinary trust region subproblem, in which the constraint is a single ellipsoid. The computational complexity of the extended trust region subproblem in general is still unknown. In this paper we consider several interesting cases related to this problem and show that for those cases the corresponding semidefinite programming relaxation admits no gap with the true optimal value, and consequently we obtain polynomial-time procedures for solving those special cases of quadratic optimization. For the extended trust region subproblem itself, we introduce a parameterized problem and prove the existence of a trajectory that will lead to an optimal solution. Combining this with a result obtained in the first part of the paper, we propose a polynomial-time solution procedure for the extended trust region subproblem arising from solving nonlinear programs with a single equality constraint.

Key words. quadratic minimization, SDP relaxation, parameterization

AMS subject classifications. 90C20, 90C22, 90C26

DOI. 10.1137/S105262340139001X

1. Introduction.

This paper is concerned with solving quadratic optimization problems by means of semidefinite programming (SDP). In particular, we focus on indefinite quadratic optimization with two or more quadratic constraints.

In the literature, quadratic optimization has received much attention. It is a fundamental problem in optimization theory and practice. Economic equilibrium, combinatorial optimization, numerical partial differential equations, and general nonlinear programming are all sources of quadratic optimization.

Recent results on quadratic optimization include the following: Bellare and Rogaway [1] established several negative results on approximating this problem; Goemans and Williamson [5], using an SDP relaxation, proved an approximation result for the Maxcut problem, which is a special quadratic optimization problem; Nesterov [11] and Ye [19] extended their SDP relaxation to approximate quadratic optimization with simple bound and diagonally homogeneous quadratic constraints; Nesterov [12] and Nemirovskii, Roos, and Terlacky [10] established a quality bound when the constraints are convex and homogeneous; and Fu, Luo, and Ye [4] constructed a quality bound for approximating quadratic optimization for general convex quadratic constraints.

More recently, Sturm and Zhang [18] proposed a quite different approach to quadratic optimization. They introduced a concept called matrix copositivity over a domain; that is a set of matrices which, in the quadratic form, is nonnegative over

*Received by the editors May 31, 2001; accepted for publication (in revised form) December 26, 2002; published electronically July 18, 2003.
http://www.siam.org/journals/siopt/14-1/39001.html

†Department of Management Sciences, University of Iowa, Iowa City, IA and Department of Systems Engineering and Engineering Management, The Chinese University of Hong Kong, Shatin, Hong Kong. Current address: Department of Management Science and Engineering, Stanford University, Stanford, CA (yinyu-ye@stanford.edu). This author's research was partially supported by NSF DMI-0231600.

‡Department of Systems Engineering and Engineering Management, The Chinese University of Hong Kong, Shatin, Hong Kong (zhang@se.cuhk.edu.hk). This author's research was supported by Hong Kong RGC earmarked grant CUHK4181/00E.

the given domain. For several specific choices of the domain, Sturm and Zhang [18] proved that such a matrix set can be characterized using linear matrix inequalities (LMI). Examples of such domains are (1) the level set of an arbitrary quadratic function, (2) the contour of a strictly concave quadratic function at zero level, and (3) the intersection of the level set of a convex quadratic function and a half-space. The key techniques used in [18] include a dual cone representation approach and a specific matrix rank-one decomposition scheme. As a consequence of the results in [18], optimizing an indefinite quadratic function under a single (nonconvex) quadratic constraint (equality or inequality), or under a convex quadratic inequality constraint and a linear inequality constraint, can be done in polynomial time, by first solving a specific form of SDP relaxation, followed by a matrix decomposition procedure.

In the current paper, we consider quadratic optimization directly. It turns out that there are more classes of quadratic optimization problems for which the SDP relaxation is exact, in the sense that its optimal value is equal to the true optimal value, and an optimal solution for the original problem can be obtained from the optimal solution of the SDP relaxation. More specifically, in section 2 we extend the matrix decomposition idea to solve the following classes of nonconvex quadratic minimization problems with two quadratic constraints:

(1) one of the two constraints in the SDP relaxation is not binding;
(2) the two constraint functions and the objective are all homogeneous quadratic functions;
(3) there are one ellipsoidal and one linear complementarity constraint.

To see why these cases are of interest, we mention that a special case of (1) is a problem studied by Stern and Wolkowicz in [16], where the analysis is lengthy and technical. The classical trust region problem (see [3]), namely, minimizing a quadratic function subject to an ellipsoid constraint, is a special case of (2). To see this, we note that the classical trust region problem can be homogenized, so that the problem becomes that of minimizing a homogeneous quadratic function, subject to two homogeneous quadratic constraints. Problem (3) is a typical problem known as *mathematical program with equilibrium constraint* (MPEC); see [9]. The MPEC problems have many practical applications and are very hard to solve in general. As far as we know, the computational complexity of the problems in (1) and (3) described above was unknown (see [14] for an alternative solution for (2)), and their solutions turn out to indeed be both interesting and nontrivial, as the current paper reveals.

The problem of minimizing an indefinite quadratic function with two (general) convex quadratic constraints arises from applying the trust region method to solving equality constrained nonlinear programs. Such a method was first proposed by Celis, Dennis, and Tapia in [3]. To distinguish it from the usual trust region subproblem, which is minimizing an indefinite quadratic function over a unit ball, we call the above problem *the extended trust region subproblem*. Although some of the cases discussed in the previous paragraph can be considered as special cases of the extended trust region subproblem, the computational complexity of the latter problem is still unknown. In section 3 of the current paper, we introduce a parameterized problem and show that by following a trajectory generated by the parameterized problem, one will arrive at the optimal solution of the original problem. Some examples are worked out in the same section to show how the method works.

Finally, we consider the extended trust region subproblem for nonlinear programming with one equality constraint. By combining results from sections 2 and 3, we present in section 4 a polynomial-time procedure for solving the subproblem. Some discussion and conclusions can be found in the same section.

Notation and convention. We let $\|\cdot\|$ denote the Euclidean norm. e_i is the unit vector, where the ith component is 1 and others are all 0. "$X \succeq 0$" stands for the fact that the symmetric matrix X is positive semidefinite. "$X \bullet Y := \text{tr}\ (X^T Y)$" is the usual matrix inner product. For a quadratic function $q(x) = x^T Q x - 2b^T x + c$ we denote the matrix representation of the function $q(\cdot)$ as $M(q(\cdot)) = \left[\begin{smallmatrix} c, & -b^T \\ -b, & Q \end{smallmatrix}\right]$. "$SOC$" stands for the second order cone, namely, $SOC = \{\left[\begin{smallmatrix} t \\ x \end{smallmatrix}\right] \,|\, t \geq \|x\|\}$. All the vectors and matrices are assumed to have appropriate dimensions, which we suppress in several places for the sake of simplicity, in such a way that the operations to follow are validated. To simplify the expression, the notion of "polynomial-time solvability" in this paper is used in the following loose sense. All basic operations such as addition, subtraction, multiplication, division, and comparison are considered as real number operations and are assumed to be executed exactly. Hence, a procedure is called polynomial if the total number of basic operations is bounded by a polynomial of the problem data. When an optimization model is under consideration, the problem data include the problem dimension and $\log 1/\epsilon$, where $\epsilon > 0$ is the precision of the solution. Porkolab and Khachiyan [15] proved the following complexity result for SDP. Consider a standard SDP problem. Let n be the order of the decision matrix, and let m be the number of inequality constraints. Then the problem can be solved in $mn^{O(\min\{m,n^2\})}$ basic operations over $Ln^{O(\min\{m,n^2\})}$-bit numbers, where L is the input length of the SDP problem. A consequence of this result is that, when m is a fixed constant, which is the case in this paper, the SDP relaxation problem can be solved in polynomial time.

2. Exact SDP relaxations. This section is concerned with quadratic optimization whose SDP relaxation admits no gap with the true optimal value, and whose optimal solution can be found in polynomial time using the SDP optimal solution.

Formally we consider the following general quadratic optimization problem:

$$\begin{array}{lll} \text{(Q)} & \text{minimize} & x^T Q_0 x - 2b_0^T x \\ & \text{subject to} & x^T Q_i x - 2b_i^T x + c_i \leq 0, \quad i = 1, \ldots, m. \end{array}$$

Let $q_i(x) = x^T Q_i x - 2b_i^T x + c_i$, $i = 1, \ldots, m$.

We assume throughout the paper that the Slater regularity condition is satisfied; i.e., there exists x_0 such that $q_i(x^0) < 0$ for all $i = 1, \ldots, m$.

For convenience, we adopt the following notation. For a quadratic function $q(x) = x^T Q x - 2b^T x + c$ we denote its matrix representation by

$$M(q(\cdot)) = \begin{bmatrix} c, & -b^T \\ -b, & Q \end{bmatrix}.$$

The homogenized version of (Q) is

$$\begin{array}{lll} \text{(HQ)} & \text{minimize} & x^T Q_0 x - 2b_0^T x t \\ & \text{subject to} & x^T Q_i x - 2b_i^T x t + c_i t^2 \leq 0, \quad i = 1, \ldots, m, \\ & & t^2 = 1. \end{array}$$

Clearly, if $\left[\begin{smallmatrix} t \\ x \end{smallmatrix}\right]$ solves (HQ), then x/t solves (Q).

The so-called SDP relaxation of (HQ) is

$$\begin{array}{lll} \text{(SP)} & \text{minimize} & M(q_0(\cdot)) \bullet X \\ & \text{subject to} & M(q_i(\cdot)) \bullet X \leq 0, \quad i = 1, \ldots, m, \\ & & X_{00} = 1,\ X \succeq 0, \end{array}$$

where $X = \left[\begin{smallmatrix} X_{00}, & x_0^T \\ x_0, & \bar{X} \end{smallmatrix}\right]$, $q_i(x) = x^T Q_i x - 2b_i^T x + c_i$ for $i = 0, 1, \ldots, m$.

The SDP problem (SP) has a dual, which is given by

$$\begin{array}{lll} \text{(SD)} & \text{maximize} & y_0 \\ & \text{subject to} & Z = M(q_0(\cdot)) - y_0 \left[\begin{smallmatrix} 1, & 0 \\ 0, & 0 \end{smallmatrix}\right] + \sum_{i=1}^m y_i M(q_i(\cdot)), \\ & & Z \succeq 0,\ y_i \geq 0, \quad i = 1, \ldots, m. \end{array}$$

Since (Q) satisfies the Slater condition, it follows that (SP) satisfies the Slater condition too.

Additionally, we assume that (SD) satisfies the Slater condition as well. This is true at least for the following two interesting cases, as shown by the next proposition.

PROPOSITION 2.1. *The problem* (SD) *satisfies the Slater regularity condition when either at least one of the m constraints is ellipsoidal or the objective function is strictly convex.*

Proof. In the first case, let us assume without loss of generality that the first constraint is ellipsoidal. In mathematical terms, this means that $Q_1 \succ 0$ and $c_1 - b_1^T Q_1^{-1} b_1 < 0$.

By fixing $y_2 = \cdots = y_m = 1$ and letting $y_1 > 0$ be sufficiently large, we will have

$$Q_0 + \sum_{i=1}^m y_i Q_i \succ 0.$$

Then we let $y_0 < 0$ be sufficiently large in absolute value to obtain

$$M(q_0(\cdot)) - y_0 \left[\begin{array}{ll} 1, & 0 \\ 0, & 0 \end{array}\right] + \sum_{i=1}^m y_i M(q_i(\cdot)) \succ 0.$$

In the second case, the objective function is strictly convex, i.e., $Q_0 \succ 0$. In that case, we let $y_i = \epsilon > 0$ be sufficiently small, $i = 1, \ldots, m$, and $y_0 < 0$ be sufficiently large in absolute value. The Slater condition follows from the fact that $M(q_0(\cdot)) - y_0 \left[\begin{smallmatrix} 1, & 0 \\ 0, & 0 \end{smallmatrix}\right] \succ 0$. □

Following a well-known result in optimization (see, e.g., [13]), if both (SP) and (SD) satisfy the Slater condition, then they have complementary optimal solutions.

In the subsequent discussion we are mainly concerned with the case in which $m = 2$, and we assume that the assumptions in Proposition 2.1 are satisfied.

Before proceeding we first quote a matrix decomposition result from [18]. For the sake of completeness, we also provide a proof here.

LEMMA 2.2. *Let G be an arbitrary symmetric matrix. Let X be a positive semidefinite matrix with rank r. Suppose that $G \bullet X \leq 0$. Then there exists a rank-one decomposition for X such that*

$$X = \sum_{i=1}^r x_i x_i^T$$

and $x_i^T G x_i \leq 0$ for all $i = 1, \ldots, r$. If, in particular, $G \bullet X = 0$, then $x_i^T G x_i = 0$ for all $i = 1, \ldots, r$.

Proof. The proof is constructive. Let

$$X = \sum_{i=1}^r u_i u_i^T$$

be an arbitrary rank-one decomposition, using, e.g., the Cholesky decomposition of X.

If $u_i^T G u_i \le 0$ for all $i = 1, \dots, r$, then we let $x_i := u_i$, $i = 1, \dots, r$, and the proof is complete.

If there is a u_i with $u_i^T G u_i > 0$ for some $1 \le i \le r$, then, due to the fact that

$$G \bullet X = \sum_{i=1}^{r} u_i^T G u_i \le 0,$$

there must exist j with $1 \le j \le r$ such that $u_j^T G u_j < 0$.

We may rename the indices if necessary so that we assume $i = 1$ and $j = 2$ for simplicity.

Consider the following quadratic equation in t:

$$0 = (tu_1 + u_2)^T G(tu_1 + u_2) = t^2(u_1^T G u_1) + 2t(u_1^T G u_2) + u_2^T G u_2.$$

This equation must have two distinct real roots with opposite signs since we have $(u_1^T G u_1)(u_2^T G u_2) < 0$. Let $\bar{t}$ be one of the roots. Let

$$\bar{u}_1 = \frac{\bar{t}}{\sqrt{\bar{t}^2 + 1}} u_1 + \frac{1}{\sqrt{\bar{t}^2 + 1}} u_2$$

and

$$\bar{u}_2 = -\frac{1}{\sqrt{\bar{t}^2 + 1}} u_1 + \frac{\bar{t}}{\sqrt{\bar{t}^2 + 1}} u_2.$$

Obviously we have

$$\bar{u}_1 \bar{u}_1^T + \bar{u}_2 \bar{u}_2^T = u_1 u_1^T + u_2 u_2^T$$

and $\bar{u}_1^T G \bar{u}_1 = 0$.

Now, recall $u_1 := \bar{u}_1$ and $u_2 := \bar{u}_2$. The decomposition

$$X = \sum_{i=1}^{r} u_i u_i^T$$

still holds. Moreover, the total number of nonzeros in the set $\{u_i^T G u_i \mid i = 1, \dots, r\}$ is strictly decreased by 1. Therefore this procedure must terminate in at most $r - 1$ steps, with $u_i^T G u_i \le 0$ for all $i = 1, \dots, r$. Then we let $x_i := u_i$, $i = 1, \dots, r$, and the lemma is proven by this construction.

If $G \bullet X = 0$, then the procedure terminates with $x_i^T G x_i = 0$, $i = 1, \dots, r$. □

2.1. Nonbinding SDP relaxation. In this subsection we consider (Q) with $m = 2$, and at least one of the two constraints $M(q_i(\cdot)) \bullet X \le 0$, $i = 1, 2$, is not binding at the optimality. Without loss of generality, suppose that $M(q_2(\cdot)) \bullet X < 0$. This implies, by complementarity, that $y_2 = 0$ at optimality. Let X^* be an optimal solution of (SP). By applying Lemma 2.2 we get a rank-one decomposition of X^* such that

$$X^* = \sum_{j=1}^{r} x_j^* (x_j^*)^T \quad \text{and} \quad (x_j^*)^T M(q_1(\cdot)) x_j^* = 0 \qquad \text{for all } j = 1, \dots, r, \tag{2.1}$$

where

$$0 \neq x_j^* = \begin{bmatrix} t_j^* \\ \bar{x}_j^* \end{bmatrix}, \quad j = 1, \ldots, r.$$

Since $M(q_2(\cdot)) \bullet X^* = \sum_{j=1}^r (x_j^*)^T M(q_2(\cdot)) x_j^* < 0$, there must exist k with $1 \leq k \leq r$ such that

$$(x_k^*)^T M(q_2(\cdot)) x_k^* \leq 0. \tag{2.2}$$

Since (SD) satisfies the Slater condition, we have $t_k^* \neq 0$, because otherwise the primal optimal set will be unbounded, which is impossible due to the dual Slater condition. (For a detailed account of the duality relations for conic optimization, one is referred to the Ph.D. thesis of Sturm; see [17].)

It follows from (2.1) and (2.2) that

$$\begin{cases} M(q_1(\cdot)) \bullet \left(\begin{bmatrix} 1 \\ \bar{x}_k^*/t_k^* \end{bmatrix} \cdot \left[1, (\bar{x}_k^*/t_k^*)^T\right] \right) = 0, \\ \\ M(q_2(\cdot)) \bullet \left(\begin{bmatrix} 1 \\ \bar{x}_k^*/t_k^* \end{bmatrix} \cdot \left[1, (\bar{x}_k^*/t_k^*)^T\right] \right) \leq 0. \end{cases} \tag{2.3}$$

Let (y^*, Z^*) be an optimal solution for (SD). By complementarity we have $X^* Z^* = 0$. It follows therefore that

$$\sum_{j=1}^r (x_j^*)^T Z^* x_j^* = 0,$$

and consequently,

$$(x_j^*)^T Z^* x_j^* = 0$$

for all $j = 1, \ldots, r$. In particular,

$$(x_k^*)^T Z^* x_k^* = 0,$$

and so

$$Z^* \bullet \left(\begin{bmatrix} 1 \\ \bar{x}_k^*/t_k^* \end{bmatrix} \cdot \left[1, (\bar{x}_k^*/t_k^*)^T\right] \right) = 0. \tag{2.4}$$

Combining (2.3) and (2.4) and noting that

$$M(q_1(\cdot)) \bullet \left(\begin{bmatrix} 1 \\ \bar{x}_k^*/t_k^* \end{bmatrix} \cdot \left[1, (\bar{x}_k^*/t_k^*)^T\right] \right) = 0 \quad \text{and} \quad y_2^* = 0,$$

we conclude that $\left[\begin{smallmatrix} 1 \\ \bar{x}_k^*/t_k^* \end{smallmatrix}\right] \cdot [1, (\bar{x}_k^*/t_k^*)^T]$ is an optimal solution for (SP) as well. Note that (SP) is a relaxation of (Q). Therefore, $\left[\begin{smallmatrix} 1 \\ \bar{x}_k^*/t_k^* \end{smallmatrix}\right]$ must be an optimal solution for (Q). All the procedures described above, including solving the SDP relaxation (SP) and the rank-one decomposition procedure in Sturm and Zhang [18] (see Lemma 2.2), are polynomial. This leads to the following result.

THEOREM 2.3. *Suppose that* (SP) *and* (SD) *both satisfy the Slater condition and* $m = 2$. *Furthermore, suppose that the primal problem* (SP) *has at least one nonbinding constraint at optimality. Then* (Q) *can be solved in polynomial time.*

One consequence of Theorem 2.3 is the following.

COROLLARY 2.4. *Suppose that* (SP) *and* (SD) *both satisfy the Slater condition and* $m = 2$. *Furthermore, suppose that* $q_1(x) \le 0$ *for all* x *with* $q_2(x) \ge 0$. *Moreover, suppose that* $q_1(x)$ *and* $q_2(x)$ *do not share any common root. Then,* (Q) *can be solved in polynomial time.*

Proof. If $q_2(x) \le 0$ for all x, then the second constraint in (Q), namely, $q_2(x) \le 0$ itself, is redundant, in which case the polynomial-time solvability of (Q) is well known, as (Q) has only one quadratic inequality constraint.

Let us assume that there is an $\hat{x}$ such that $q_2(\hat{x}) > 0$. Then, by the S-Lemma (see [2]), since $-q_1(x) \ge 0$ for all $q_2(x) \ge 0$, there must exist $t \ge 0$ such that

$$-M(q_1(\cdot)) - tM(q_2(\cdot)) \succeq 0. \tag{2.5}$$

If $-M(q_1(\cdot)) \succeq 0$, then the constraint $q_1(x) \le 0$ is redundant, and the problem (Q) again has only one quadratic inequality constraint and hence is solvable in polynomial time. Thus, for the interesting case we may assume $t > 0$.

Now we wish to show that the SDP relaxation (SP) cannot be binding at any feasible solution. Suppose by contradiction that there is a feasible $X \succeq 0$ for (SP) such that $M(q_1(\cdot)) \bullet X = 0$ and $M(q_2(\cdot)) \bullet X = 0$. Then

$$(-M(q_1(\cdot)) - tM(q_2(\cdot))) \bullet X = 0. \tag{2.6}$$

By Lemma 2.2, we can get a rank-one decomposition of X, $X = \sum_{i=1}^{r} x_i x_i^T$, such that

$$x_i^T M(q_1(\cdot)) x_i = 0$$

for all $i = 1, \ldots, r$. By (2.5) and (2.6) we also have

$$x_i^T(-M(q_1(\cdot)) - tM(q_2(\cdot)))x_i = 0$$

for all $i = 1, \ldots, r$. Because $t > 0$, it follows that

$$x_i^T M(q_2(\cdot)) x_i = 0$$

for all $i = 1, \ldots, r$.

Since $X_{00} = 1$, there must exist $x_j = \left[\begin{smallmatrix} t_j \\ \bar{x}_j \end{smallmatrix}\right]$ such that its first component t_j is nonzero. Then we have

$$q_1(x_j/t_j) = x_j^T M(q_1(\cdot)) x_j / t_j^2 = x_j^T M(q_2(\cdot)) x_j / t_j^2 = q_2(x_j/t_j) = 0.$$

This contradicts the assumption that there is no common root for $q_1(x)$ and $q_2(x)$. ∎

As an application, we consider the following quadratic program:

$$\begin{array}{ll} \text{minimize} & q_0(x) \\ \text{subject to} & l \le q_1(x) \le u, \end{array}$$

where $l < u$. This problem was analyzed thoroughly by Stern and Wolkowicz in [16].

This problem clearly satisfies the conditions in Corollary 2.4, because the two constraints are

$$q_1(x) - u \leq 0 \quad \text{and} \quad l - q_1(x) \leq 0.$$

Therefore,

$$q_1(x) - u \geq 0 \quad \text{implies} \quad l - q_1(x) = l - u - (q_1(x) - u) \leq 0.$$

Moreover,

$$q_1(x) - u = 0 \quad \text{and} \quad l - q_1(x) = 0$$

cannot hold at the same time. Therefore, we can apply Corollary 2.4 to conclude that in this case (Q) is solvable in polynomial time.

We remark that the model investigated in Stern and Wolkowicz [16] assumes that $q_1(x)$ is a pure quadratic form. In this sense, our result is also a little bit more general.

Interestingly, this lends itself to solving the following problem:

$$\begin{array}{ll} \text{minimize} & |q_0(x)| \\ \text{subject to} & q_1(x) \leq 0. \end{array}$$

The key is to rewrite the problem as

$$\begin{array}{ll} \text{minimize} & t \\ \text{subject to} & q_1(x) \leq 0, \\ & -t \leq q_0(x) \leq t, \end{array}$$

and observe that for any fixed $t \geq 0$ the feasibility check of the above problem reduces to

$$\begin{array}{ll} \text{minimize} & q_1(x) \\ \text{subject to} & -t \leq q_0(x) \leq t. \end{array}$$

If the optimal value of this problem is positive, then the original problem is infeasible for that given t; otherwise, it is feasible.

For $t = 0$, this problem can be solved by the SDP relaxation method, as it reduces to one quadratic equality constraint. Otherwise, $t > 0$, and we may resort to Theorem 2.3 for its solution.

Since the feasibility check can be done in polynomial time for any given objective value t, we may solve the optimization problem using bisection on the objective value.

There are other nontrivial domains that are claimed by Corollary 2.4, such as the whole space with two nonintersecting ellipsoids taken away. Minimizing an indefinite quadratic function over such a domain, claims Corollary 2.4, is easy.

2.2. Homogeneous quadratic functions. Another polynomially solvable special case of (Q) is $m = 2$, and all the functions involved, $q_0(x)$, $q_1(x)$, and $q_2(x)$, are homogeneous, i.e., there are no linear terms. Hence, the problem can be simply written as

$$\begin{array}{ll} \text{minimize} & x^T Q_0 x \\ \text{subject to} & x^T Q_1 x \leq 1, \\ & x^T Q_2 x \leq 1. \end{array}$$

Due to its homogeneous form, the corresponding SDP relaxation is

$$\begin{array}{ll} \text{minimize} & Q_0 \bullet X \\ \text{subject to} & Q_1 \bullet X \leq 1, \\ & Q_2 \bullet X \leq 1, \\ & X \succeq 0. \end{array}$$

Its dual problem is

$$\begin{array}{ll} \text{maximize} & y_1 + y_2 \\ \text{subject to} & Z = Q_0 - y_1 Q_1 - y_2 Q_2, \\ & Z \succeq 0,\ y_1 \leq 0,\ y_2 \leq 0. \end{array}$$

Suppose that the primal-dual problems have a pair of complementary optimal solutions. Again, a sufficient condition to ensure this is that one of the Q_1, Q_2 matrix is positive definite.

If one of the two constraints in the primal SDP relaxation is not binding at the optimality, then the results in subsection 2.1 apply and the problem is solved.

Consider the case in which they are both binding at the optimality. Let the primal optimal solution be X^*, and the dual optimal solution be (y_1^*, y_2^*, Z^*).

We now apply Lemma 2.2 to generate

$$X^* = \sum_{i=1}^{r} x_i^* (x_i^*)^T$$

such that $(x_i^*)^T (Q_1 - Q_2) x_i^* = 0$ for all $i = 1, \ldots, r$.

Since $\sum_{i=1}^{r} (x_i^*)^T Q_1 x_i^* = 1$, we may select x_j^*, $1 \leq j \leq r$, such that $(x_j^*)^T Q_1 x_j^* =: \tau > 0$. By our construction, $(x_j^*)^T Q_2 x_j^* = \tau$.

Let

$$x^* = \frac{x_j^*}{\sqrt{\tau}}.$$

We see that $x^*(x^*)^T$ is a primal feasible solution for the SDP relaxation. Moreover, it is optimal, because

$$0 \leq (x^*)^T Z^* x^* \leq \frac{1}{\tau} X^* \bullet Z^* = 0$$

and $[1 - (x^*)^T Q_i x^*] y_i^* = 0 \times y_i^* = 0$ for $i = 1, 2$, and hence the primal-dual complementarity conditions are satisfied.

This shows that the SDP relaxation admits no gap with the true optimal value, and an optimal solution for the original quadratic optimization problem can be constructed in polynomial time.

The above result was first proved by Polyak in [14].[1] However, the method used in Polyak [14] is based on an extension of Hausdorff's result [6] regarding the convexity of the image of homogeneous quadratic mapping: $(x^T A_0 x, x^T A_1 x, x^T A_2 x)^T$. Hence, the underlying methodologies are quite different.

One advantage of our approach is that it allows some room for approximation when $m > 2$, as we see below.

[1] We are indebted to a referee for pointing the reference [14] out to us.

Consider the SDP problem

$$\begin{array}{lll} (\text{SDP}) \quad z^{SDP} := & \text{minimize} & Q_0 \bullet X \\ & \text{subject to} & Q_i \bullet X \leq (=)\ 1, \quad i = 1, \ldots, m, \\ & & X \succeq 0, \end{array}$$

where $2 \leq m \leq n$.

We assume that the problem satisfies the Slater condition. Then we have the following result.

THEOREM 2.5. *Let X^* be a minimizer of* (SDP). *Then we can compute another minimizer of* (SDP) *whose rank is no more than $m-1$ in polynomial time.*[2]

Proof. Suppose that the rank of an initial minimizer X^* is r and $r > m-1$. Without losing generality, let $Q_1 \bullet X^* = 1$ so that

$$(Q_2 - Q_1) \bullet X^* \leq (=)\ 0.$$

From Lemma 2.2, there exist column vectors x_j, $j = 1, \ldots, r$, such that

$$X^* = \sum_{j=1}^{r} x_j x_j^T,$$

and for every j

$$(Q_2 - Q_1) \bullet x_j x_j^T \leq (=)\ 0.$$

Let

$$a_{ij} = Q_i \bullet x_j x_j^T = x_j^T Q_i x_j,$$

and consider a linear program

$$\begin{array}{ll} \text{minimize} & \sum_{j=1}^{r} a_{0j} v_j \\ \text{subject to} & \sum_{j=1}^{r} a_{ij} v_j \leq (=)\ 1, \quad i = 1, \ldots, m, \\ & v_j \geq 0, \quad j = 1, \ldots, r. \end{array}$$

Since, for every j,

$$a_{2j} = Q_2 \bullet x_j x_j^T \leq (=)\ Q_1 \bullet x_j x_j^T = a_{1j},$$

constraint $\sum_{j=1}^{r} a_{1j} v_j \leq (=)\ 1$ implies that $\sum_{j=1}^{r} a_{2j} v_j \leq (=)\ 1$. That is, the second constraint in the linear programming (LP) problem is redundant. Therefore, the LP problem is equivalent to

$$\begin{array}{ll} \text{minimize} & \sum_{j=1}^{r} a_{0j} v_j \\ \text{subject to} & \sum_{j=1}^{r} a_{ij} v_j \leq (=)\ 1, \quad i = 1, 3, \ldots, m, \\ & v_j \geq 0, \quad j = 1, \ldots, r. \end{array}$$

Note that $v_1 = \cdots = v_r = 1$ is an optimal solution for the LP problem, since for any $v_j \geq 0$, $j = 1, \ldots, r$,

$$X = \sum_{j=1}^{r} v_j \cdot x_j x_j^T$$

[2] This theorem first appeared in an unpublished discussion note by Kim, Kojima, and Ye [8]. It was also given as a quiz to the students of an optimization course in Tokyo Institute of Technology, 2001. One of the students, Hayato Waki, successfully proved the theorem.

is a feasible solution for the SDP problem. Thus, the LP minimal value is also $z^{SDP} = Q_0 \bullet x^*$, which corresponds to $v_1 = \cdots = v_r = 1$.

Since the LP problem is bounded, it must have a basic optimal feasible solution. At a basic optimal solution, we should have at least r inequalities active or binding. Thus, we should have at most $m-1$ inequalities inactive, since the total inequalities of the LP problem is $m-1+r$. Thus, among r of v_j variables, at most $m-1$ of them are positive at the optimal basic solution. Let it be v^*. Then,

$$X^{**} = \sum_{j=1}^{r} v_j^* \cdot x_j x_j^T$$

is also a minimizer for the SDP, and its rank is no more than $m-1$. □

Consider the homogeneous quadratic minimization with $m \geq 2$ homogeneous quadratic constraints:

$$\begin{array}{lll} \text{(QP)} \quad z^* := & \text{minimize} & x^T Q_0 x \\ & \text{subject to} & x^T Q_i x \leq 1, \quad i = 1, \ldots, m. \end{array}$$

Its SDP relaxation is the one presented above, and

$$z^{SDP} \leq z^* \leq 0.$$

COROLLARY 2.6. *Let X^* be the low-rank minimizer of* (SDP) *with rank $r \leq \min\{m-1, n\}$. Then we can quickly (in polynomial time) compute a feasible solution of* (QP) *such that*

$$x^T Q_0 x \leq \frac{1}{r} z^{SDP} \leq \frac{1}{r} z^*$$

if $Q_i \succeq 0$ for $i = 1, \ldots, m$.

Proof. Let

$$X^* = \sum_{j=1}^{r} x_j x_j^T.$$

Then, for every i and j,

$$Q_i \bullet x_j x_j^T \leq Q_i \bullet X^* \leq 1,$$

so that x_j is a feasible solution to (QP) for every j. Without losing generality, assume that

$$Q_0 \bullet x_1 x_1^T = \min_j \{Q_0 \bullet x_j x_j^T\}.$$

Then

$$Q_0 \bullet x_1 x_1^T \leq \frac{1}{r} \sum_{j=1}^{r} Q_0 \bullet x_j x_j^T = \frac{1}{r} Q_0 \bullet \left(\sum_{j=1}^{r} x_j x_j^T \right) = \frac{1}{r} Q_0 \bullet X^* = \frac{1}{r} z^{SDP}. \qquad \square$$

Note that if $m = 2$, the approximation ratio is 1.

2.3. Complementary linear constraints. In this subsection, we consider the following special case of (Q):

$$\begin{array}{lll} \text{(CL)} & \text{minimize} & q_0(x) \\ & \text{subject to} & \|x\|^2 \le 1, \\ & & \bar{a}^T x \le a_0, \\ & & \bar{b}^T x \le b_0, \\ & & (a_0 - \bar{a}^T x)(b_0 - \bar{b}^T x) = 0. \end{array}$$

The last constraint is a complementarity condition. The problem of this type is known as the MPEC; see [9]. The MPEC problems have many practical applications and are very hard to solve in general.

In [18], the above problem with $b_0 = 0$ and $\bar{b} = 0$ is solved via a special type of SDP relaxation. We now extend the method to solve (CL). Let

$$J = \begin{bmatrix} 1, & 0 \\ 0, & -I \end{bmatrix}, \quad a = \begin{bmatrix} a_0 \\ -\bar{a} \end{bmatrix}, \quad b = \begin{bmatrix} b_0 \\ -\bar{b} \end{bmatrix}.$$

Recall that we denote the standard second order cone in $\Re^{n+1}$ by

$$SOC = \left\{ \begin{bmatrix} t \\ x \end{bmatrix} \;\middle|\; t \ge \|x\| \right\}.$$

Consider the following SDP relaxation for the homogenized version of (CL):

$$\begin{array}{lll} \text{(CLSP)} & \text{minimize} & M(q_0(\cdot)) \bullet X \\ & \text{subject to} & J \bullet X \ge 0, \\ & & Xa \in SOC, \\ & & Xb \in SOC, \\ & & a^T X b = 0, \\ & & X_{00} = 1, \\ & & X = \begin{bmatrix} X_{00}, & x_0^T \\ x_0, & \bar{X} \end{bmatrix} \succeq 0. \end{array}$$

Clearly, (CLSP) is a relaxation of (CL), since, if X is rank one, then its eigenvector is simply a solution of (CL). This problem has a dual, given as follows:

$$\begin{array}{lll} \text{(CLSD)} & \text{maximize} & y_1 \\ & \text{subject to} & Z = M(q_0(\cdot)) - y_0 J - y_1 e_1 e_1^T \\ & & \quad -(a y_a^T + y_a a^T) - (b y_b^T + y_b b^T) - y_2(ab^T + ba^T), \\ & & y_a \in SOC, \\ & & y_b \in SOC, \\ & & y_0 \ge 0, \\ & & Z \succeq 0. \end{array}$$

Let us now assume that the regularity condition is satisfied so that (CLSP) and (CLSD) have complementary optimal solutions, denoted by X^* and $(y_0^*, y_1^*, y_2^*, y_a^*, y_b^*, Z^*)$, respectively. That is,

$$X^* Z^* = 0, \quad y_0^*(J \bullet X^*) = 0, \quad (y_a^*)^T X^* a = 0, \quad (y_b^*)^T X^* b = 0. \tag{2.7}$$

The main result in this subsection is the following assertion.

THEOREM 2.7. *Suppose that the SDP relaxation* (CLSP) *and it dual problem* (CLSD) *have complementary optimal solutions. Then the optimal value of* (CLSP)*, which equals that of* (CLSD) *by strong duality, is identical to the optimal value of* (CL)*. In other words, the relaxation admits no gap. Moreover, an optimal solution for* (CL) *can be obtained in polynomial time, provided that the solution for its SDP relaxation* (CLSP) *is available.*

Proof. The proof below uses the matrix rank-one decomposition procedure proposed by Sturm and Zhang in [18]. The idea is to construct a rank-one feasible solution for (CLSP), based on X^*, such that the complementarity conditions are still satisfied, thus ensuring the optimality. We proceed by considering several possible cases regarding the status of X^*.

Case 1. $X^*a = 0$ and $X^*b = 0$. Applying Lemma 2.2, we obtain a decomposition for X^*:

$$X^* = \sum_{i=1}^{r} x_i^*(x_i^*)^T,$$

where r is the rank of X^* such that $J \bullet [x_i^*(x_i^*)^T] \geq 0$ for all $i = 1, \ldots, r$. Moreover, $J \bullet [x_i^*(x_i^*)^T] = 0$ for all $i = 1, \ldots, r$ if $J \bullet X^* = 0$. We may choose the sign of the first component in x_i^* to ensure that $x_i^* \in SOC$, $i = 1, \ldots, r$.

By linear independence of x_i^*'s, we get $a^T x_i^* = 0$ and $b^T x_i^* = 0$, $i = 1, \ldots, r$. Let $x_i^* = \left[\begin{smallmatrix} t_i^* \\ \bar{x}_i^* \end{smallmatrix}\right]$, $i = 1, \ldots, r$. Since $x_i^* \in SOC$ and $x_i^* \neq 0$, we have $t_i^* > 0$, $i = 1, \ldots, r$. Take any $1 \leq j \leq r$; it follows that $\left[\begin{smallmatrix} 1 \\ \bar{x}_j^*/t_j^* \end{smallmatrix}\right][1, (\bar{x}_j^*/t_j^*)^T]$ is optimal for (CLSP).

Case 2. $J \bullet X^* > 0$ and $X^*a \neq 0$. (The case $J \bullet X^* > 0$ and $X^*b \neq 0$ is similar). In this case, $y_0^* = 0$ and $X^*a \neq 0$. Let $x_a^* := X^*a = \left[\begin{smallmatrix} t_a^* \\ \bar{x}_a^* \end{smallmatrix}\right] \neq 0$. Since $x_a^* \in SOC$, by feasibility, we know that $t_a^* > 0$. Moreover, $J \bullet [x_a^*(x_a^*)^T] = (t_a^*)^2 - \|\bar{x}_a^*\|^2 \geq 0$, $x_a^*(x_a^*)^T b = 0$, and $x_a^*(x_a^*)^T a = (a^T X^* a) X^* a \in SOC$. Therefore, $x_a^*(x_a^*)^T/(t_a^*)^2$ is optimal for (CLSP) as well, as it is feasible and satisfies the complementarity conditions stipulated in (2.7), after replacing X^* by $x_a^*(x_a^*)^T/(t_a^*)^2$.

Case 3. $J \bullet X^* = 0$, $X^*a \neq 0$, and $X^*b = 0$. Denote $x_a^* = X^*a \neq 0$. Let $\tilde{X} := X^* - \frac{X^* a a^T X^*}{a^T X^* a} \succeq 0$. It is easy to see that $\tilde{X}a = 0$ and $\tilde{X}b = 0$. There are two possibilities.

Case 3.1. $J \bullet [x_a^*(x_a^*)^T] = 0$. In this subcase, we have that $x_a^*(x_a^*)^T/(t_a^*)^2$ is optimal for (CLSP).

Case 3.2. $J \bullet [x_a^*(x_a^*)^T] > 0$. In this subcase,

$$J \bullet \tilde{X} = J \bullet X^* - J \bullet [x_a^*(x_a^*)^T]/a^T X^* a < 0. \tag{2.8}$$

Now let us decompose $\tilde{X}$ as

$$\tilde{X} = \sum_{i=1}^{s} \tilde{x}_i \tilde{x}_i^T,$$

where $s = \text{rank}\,(\tilde{X}) > 0$. Since $\tilde{X}a = 0$ and $\tilde{X}b = 0$, we have

$$\tilde{x}_i^T a = 0 \quad \text{and} \quad \tilde{x}_i^T b = 0$$

for all $i = 1, \ldots, s$. Choose j such that

$$J \bullet [\tilde{x}_j(\tilde{x}_j)^T] < 0.$$

Such j must exist due to (2.8).

Consider the following quadratic equation:

$$J \bullet [(x_a^* + \alpha \tilde{x}_j)(x_a^* + \alpha \tilde{x}_j)^T] = 0.$$

This equation has two distinct real roots with opposite signs. Choose α with an appropriate sign so that the first component in $x_a^* + \alpha \tilde{x}_j$ is positive. Denote

$$x_a^* + \alpha \tilde{x}_j =: \begin{bmatrix} t^* \\ \bar{x}^* \end{bmatrix}.$$

In this case, since $J \bullet [x_a^*(x_a^*)^T] > 0$, it follows that x_a^* is in the strict interior of the cone SOC. Due to the complementarity, we must have $y_a^* = 0$. Let us consider the solution $\begin{bmatrix} 1 \\ \bar{x}^*/t^* \end{bmatrix}[1, (\bar{x}^*/t^*)^T]$ for (CLSP). It is readily verified that this solution is both feasible and complementary to the dual optimal solution $(y_0^*, y_1^*, y_2^*, y_a^*, y_b^*, Z^*)$. Hence it is optimal for (CLSP).

Case 4. $J \bullet X^* = 0$, $X^*a \neq 0$, and $X^*b \neq 0$. This case is treated in the same way as for Case 3.

Again, denote $x_a^* = X^*a$ and $x_b^* = X^*b$. Certainly, in this particular case, $x_a^* \neq 0$ and $x_b^* \neq 0$.

Observe that

$$\tilde{X} := X^* - \frac{X^*aa^TX^*}{a^TX^*a} - \frac{X^*bb^TX^*}{b^TX^*b} \succeq 0.$$

We have $\tilde{X}a = 0$ and $\tilde{X}b = 0$.

As before, consider two more possibilities.

Case 4.1. Either $J \bullet [x_a^*(x_a^*)^T] = 0$ or $J \bullet [x_b^*(x_b^*)^T] = 0$. Let us assume $J \bullet [x_a^*(x_a^*)^T] = 0$. In this particular case, $x_a^* \neq 0$. Hence, $x_a^*(x_a^*)^T/(t_a^*)^2$ is optimal for (CLSP).

Case 4.2. $J \bullet [x_a^*(x_a^*)^T] > 0$ and $J \bullet [x_b^*(x_b^*)^T] > 0$. In this case,

$$J \bullet \tilde{X} = J \bullet X^* - J \bullet \frac{[x_a^*(x_a^*)^T]}{a^TX^*a} - J \bullet \frac{[x_b^*(x_b^*)^T]}{b^TX^*b} < 0. \tag{2.9}$$

Now let us decompose $\tilde{X}$ as

$$\tilde{X} = \sum_{i=1}^{s} \tilde{x}_i \tilde{x}_i^T,$$

where $s = \text{rank } (\tilde{X}) > 0$. Since $\tilde{X}a = 0$ and $\tilde{X}b = 0$, we have

$$\tilde{x}_i^T a = 0 \quad \text{and} \quad \tilde{x}_i^T b = 0$$

for all $i = 1, \ldots, s$. Choose j such that

$$J \bullet [\tilde{x}_j(\tilde{x}_j)^T] < 0.$$

Such j must exist due to (2.9).

Consider the following quadratic equation:

$$J \bullet [(x_a^* + \alpha \tilde{x}_j)(x_a^* + \alpha \tilde{x}_j)^T] = 0.$$

This equation has two distinct real roots with opposite signs. Choose α with an appropriate sign so that the first component in $x_a^* + \alpha\tilde{x}_j$ is positive. Define

$$x_a^* + \alpha\tilde{x}_j =: \begin{bmatrix} t^* \\ \bar{x}^* \end{bmatrix}.$$

In this case, since $J \bullet [x_a^*(x_a^*)^T] > 0$ and $J \bullet [x_b^*(x_b^*)^T] > 0$, it follows that x_a^* and x_b^* are in the strict interior of the cone SOC. Due to the complementarity we must have $y_a^* = 0$ and $y_b^* = 0$. Let us consider the solution $[\begin{smallmatrix} 1 \\ \bar{x}^*/t^* \end{smallmatrix}][1, (\bar{x}^*/t^*)^T]$ for (CLSP). As for the case before, we can easily check that this solution is both feasible and complementary to the dual optimal solution $(y_0^*, y_1^*, y_2^*, y_a^*, y_b^*, Z^*)$. Hence it is optimal to (CLSP). □

We remark here that the solution methodology readily extends to the following more general setting:

$$\begin{array}{ll} \text{minimize} & q_0(x) \\ \text{subject to} & \|x\|^2 \le 1, \\ & \bar{a}_i^T x \le a_{i0}, \quad i = 1, \ldots, m, \\ & (a_{i0} - \bar{a}_i^T x)(a_{j0} - \bar{a}_j^T x) = 0 \quad \text{for all } i \ne j. \end{array}$$

3. Two convex quadratic constraints. Now we move on to consider the problem of minimizing a nonconvex quadratic function with two convex quadratic constraints. We assume that one of the constraints is simply ellipsoidal. More specifically, without losing generality, we assume it to be a unit spherical constraint.

Let

$$\begin{aligned} q_0(x) &= \tfrac{1}{2}x^T Q_0 x - b_0^T x, \\ q_1(x) &= \tfrac{1}{2}x^T Q_1 x - b_1^T x + \tfrac{c_1}{2}, \end{aligned}$$

where Q_0 is indefinite and $Q_1 \succeq 0$. Hence, $q_1(x)$ is convex.

The problem that we consider in this section is

$$\begin{array}{lll} \text{(P)} & \text{minimize} & q_0(x) \\ & \text{subject to} & \|x\|^2 \le 1, \\ & & q_1(x) \le 0. \end{array}$$

As we discussed in section 1, this problem arises from the application of the trust region method for equality constrained nonlinear programming. More discussion on this subject can be found in section 4.

Throughout our discussion we assume that the above problem satisfies the Slater condition; i.e., there is x such that $q_1(x) < 0$ and $\|x\|^2 < 1$. Let us denote the feasible region of (P) as Ω. Obviously, Ω is a compact convex set with a nonempty interior. Since at least one of the two constraints will be binding at optimum for the interesting cases, let us assume for simplicity that $\|x\|^2 \le 1$ is a binding constraint.

Let the optimal value of (P) be v^*.

Consider the following parameterized problem:

$$\begin{array}{lll} \text{(H}_\lambda\text{)} & \text{minimize} & q_0(x) + \lambda q_1(x) \\ & \text{subject to} & \|x\|^2 \le 1, \\ & & q_1(x) \le 0, \end{array}$$

with $\lambda \ge 0$. Let the optimal value of (H$_\lambda$) be $h(\lambda)$.

LEMMA 3.1. *The value function $h(\lambda)$ is nonincreasing and is concave. Moreover, $h(\lambda) \leq h(0) = v^*$ for all $\lambda \geq 0$.*

Proof. The concavity of $h(\lambda)$ follows from the fact that for any fixed x, $q_0(x) + \lambda q_1(x)$ is linear, and hence concave, in λ. Moreover, it is nonincreasing since $q_1(x) \leq 0$ for all $x \in \Omega$. The second assertion is obvious. □

We may introduce a perturbation if necessary, $[\epsilon_1, \epsilon_2, \ldots, \epsilon_n] > 0$, on the diagonal elements of Q_0, so that the matrix $Q_0 + \lambda Q_1$ will always have at most two identical eigenvalues for any $\lambda \geq 0$. In the rest of the paper, we assume that such is the case.

Consider another relaxed problem,

$$\begin{array}{lll} (\mathrm{F}_\lambda) & \text{minimize} & q_0(x) + \lambda q_1(x) \\ & \text{subject to} & \|x\|^2 \leq 1, \end{array}$$

with $\lambda \geq 0$. Let the optimal value of (F_λ) be $f(\lambda)$.

Using a similar argument as that used for Lemma 3.1, the following relation is readily seen.

LEMMA 3.2. *The function $f(\lambda)$ is concave, and furthermore, it holds that*

$$f(\lambda) \leq h(\lambda) \leq v^*$$

for all $\lambda \geq 0$.

For any fixed λ, (F_λ) can be easily solved, e.g., by solving its SDP relaxation followed by a matrix decomposition procedure; see [18]. Among other things, this implies that $f(\lambda)$ can be evaluated in polynomial time. In particular, for fixed λ, the optimality condition for (F_λ) is

$$\begin{cases} (Q_0 + \lambda Q_1 + \mu I)x = b_0 + \lambda b_1, \\ \mu(\|x\|^2 - 1) = 0, \quad \mu \geq 0, \quad \|x\|^2 - 1 \leq 0, \\ Q_0 + \lambda Q_1 + \mu I \succeq 0, \end{cases}$$

where the first two conditions are simply KKT and the last one follows from the SDP duality.

Let X_λ be the set of optimal solutions for (F_λ). In our case, $|X_\lambda| \leq 2$. Then

$$\partial f(\lambda) = \text{conv } \{q_1(x) \mid x \in X_\lambda\},$$

where $\partial f(\lambda)$ stands for the set of supergradients (see, e.g., Theorem 4.4.2 in [7]).

Let

$$\hat{\lambda} = \text{argmax } \{f(\lambda) \mid \lambda \geq 0\}.$$

We remark here that, due to the Slater condition, we have $f(\lambda) \to -\infty$ as $\lambda \to \infty$. Hence $\hat{\lambda}$ exists and is finite. Moreover, since $f(\lambda)$ is concave, using bisection, one can find $\hat{\lambda}$ in polynomial time with any given precision.

By the concavity of f, we conclude that X_λ contains only infeasible solutions $(q_1(x) > 0)$ of (F_λ) for $\lambda < \hat{\lambda}$, and contains only feasible solutions of (F_λ) for $\lambda > \hat{\lambda}$.

Now that $\hat{\lambda}$ is a maximum point for $f(\lambda)$, we have

$$0 \in \partial f(\hat{\lambda}).$$

If $X_{\hat{\lambda}}$ is a singleton, then its element is also optimal for (P), and we are done. If $X_{\hat{\lambda}}$ contains two elements, say $\{x^+, x^-\}$, then we must have

$$q_1(x^-) \leq 0 \leq q_1(x^+).$$

If any of the above two inequalities is actually an equality, then again the corresponding solution is optimal to (P), and we are done. Next we are concerned with the remaining case, i.e.,

$$q_1(x^-) < 0 < q_1(x^+).$$

In that case, $x^- \in \operatorname{int} \Omega$ and $x^+ \notin \Omega$.

According to Lemma 3.2 we know that $h(\hat{\lambda}) \geq f(\hat{\lambda})$. Due to the fact that $(\mathrm{F}_{\hat{\lambda}})$ is a relaxation of $(\mathrm{H}_{\hat{\lambda}})$, we conclude that x^- is optimal to $(\mathrm{H}_{\hat{\lambda}})$, and consequently $0 > q_1(x^-) \in \partial h(\hat{\lambda})$. It can be shown that $\|x^-\| = 1$.

Now consider a set of KKT solutions, denoted by S_λ^0, such that the Hessian matrix of the Lagrangian function is positive semidefinite. In particular, S_λ^0 contains all x such that there exists $\mu \geq 0$ satisfying the following conditions:

$$\begin{cases} (Q_0 + \lambda Q_1 + \mu I)x = b_0 + \lambda b_1, \\ \|x\|^2 - 1 = 0, \quad \mu \geq 0, \\ Q_0 + \lambda Q_1 + \mu I \text{ has no negative eigenvalue.} \end{cases}$$

One can easily verify that, if $Q_0 + \lambda Q_1$ has distinct eigenvalues, then $|S_\lambda^0| \leq 3$. Furthermore, let

$$(\mathrm{S}_\lambda^0)^* := \arg \min_{x \in S_\lambda^0} \; q_0(x) + \lambda q_1(x).$$

In the same vein, let us define S_λ^1 to be the set of such KKT solutions x that have one negative eigenvalue in the Hessian matrix of the Lagrangian; i.e., there is $\mu \geq 0$ satisfying

$$\begin{cases} (Q_0 + \lambda Q_1 + \mu I)x = b_0 + \lambda b_1, \\ \|x\|^2 - 1 = 0, \quad \mu \geq 0, \\ Q_0 + \lambda Q_1 + \mu I \text{ has exactly one negative eigenvalue.} \end{cases}$$

Similarly, define

$$(\mathrm{S}_\lambda^1)^* := \arg \min_{x \in S_\lambda^1} \; q_0(x) + \lambda q_1(x).$$

Our first result is the following.

THEOREM 3.3. *For any λ with $h(\lambda) < h(0) = v^*$, the optimal solution for (H_λ) is always contained in $(\mathrm{S}_\lambda^0)^* \cup (\mathrm{S}_\lambda^1)^*$.*

Proof. Let y_λ be an optimal solution of (H_λ). Since $h(\lambda) < h(0)$, i.e., λ is not a maximum point for h, it follows that $q_1(y_\lambda) < 0$. By the local optimality of y_λ we have

$$q_0(y_\lambda + d) + \lambda q_1(y_\lambda + d) \geq q_0(y_\lambda) + \lambda q_1(y_\lambda) \tag{3.1}$$

for all $\|y_\lambda + d\| \leq 1$ and $\|d\|$ sufficiently small. Moreover, y_λ must be a KKT point, i.e.,

$$(Q_0 + \lambda Q_1 + \mu I)y_\lambda = b_0 + \lambda b_1 \tag{3.2}$$

for some $\mu \geq 0$. Equations (3.1) and (3.2) imply that

$$\frac{1}{2} d^T (Q_0 + \lambda Q_1) d \geq \mu d^T y_\lambda$$

for all $\|y_\lambda + d\| \le 1$ and $\|d\|$ sufficiently small. We may rewrite this relation as

$$\begin{aligned}\frac{1}{2}d^T(Q_0 + \lambda Q_1 + \mu I)d &\ge \mu d^T y_\lambda + \frac{\mu}{2}d^T d \\ &= \frac{\mu}{2}(y_\lambda + d)^T(y_\lambda + d) - \frac{\mu}{2}y_\lambda^T y_\lambda\end{aligned}$$

for all $\|y_\lambda + d\| \le 1$ and $\|d\|$ sufficiently small. In particular,

$$\frac{1}{2}d^T(Q_0 + \lambda Q_1 + \mu I)d \ge 0 \tag{3.3}$$

for $\|y_\lambda + d\| = 1$ and $\|d\|$ sufficiently small.

Consider a fixed $\bar{d}$ satisfying $\bar{d}^T y_\lambda = 0$. Let $\epsilon > 0$ be a small number. Let $y_\lambda + d_\epsilon$ be the projection of $y_\lambda + \epsilon\bar{d}$ onto the unit sphere $\{y \mid \|y\| \le 1\}$. Since $\|y_\lambda + \epsilon\bar{d}\| > 1$, we conclude that $\|y_\lambda + d_\epsilon\| = 1$. Let $\Delta d_\epsilon = \epsilon\bar{d} - d_\epsilon$. It follows that

$$\|\Delta d_\epsilon\| = o(\epsilon). \tag{3.4}$$

Using (3.3), we have

$$\begin{aligned}0 &\le d_\epsilon^T(Q_0 + \lambda Q_1 + \mu I)d_\epsilon \\ &= (\epsilon\bar{d})^T(Q_0 + \lambda Q_1 + \mu I)(\epsilon\bar{d}) - 2(\Delta d_\epsilon)^T(Q_0 + \lambda Q_1 + \mu I)(\epsilon\bar{d}) \\ &\quad + (\Delta d_\epsilon)^T(Q_0 + \lambda Q_1 + \mu I)(\Delta d_\epsilon).\end{aligned}$$

Dividing ϵ^2 on the both sides of the above inequality and letting $\epsilon \to 0$, we get

$$\bar{d}^T(Q_0 + \lambda Q_1 + \mu I)\bar{d} \ge 0. \tag{3.5}$$

Note that the inequality (3.5) holds for any $\bar{d}$ with $\bar{d}^T y_\lambda = 0$. Hence

$$Q_0 + \lambda Q_1 + \mu I$$

can have at most one negative eigenvalue. Taking this together with (3.2), we conclude $y_\lambda \in (S_\lambda^0)^* \cup (S_\lambda^1)^*$. □

Theorem 3.3 suggests the following scheme to solve (P) by means of tracking the paths S_λ^0 and S_λ^1.

A KKT Solution Path(s) Tracking Procedure.

Step 1. Find $\hat{\lambda} = \text{argmax}\{f(\lambda) \mid \lambda \ge 0\}$. If there is $x^* \in X_{\hat{\lambda}}$ such that $q_1(x^*) = 0$, then stop with x^* being optimal to (P). Otherwise, go to Step 2.

Step 2. Track all the paths in S_λ^0 and S_λ^1 by reducing λ, starting from $\lambda = \hat{\lambda}$. If $x_\lambda \in S_\lambda^1$ and $q_1(x_\lambda) = 0$, then store such x_λ as a candidate for optimal solution and stop searching along this path. If this does not happen, then the search along each path stops either when the path ceases to exist or when $\lambda = 0$.

Step 3. Track all the paths in S_λ^0 and S_λ^1 by increasing λ, starting from $\lambda = \hat{\lambda}$. If $x_\lambda \in S_\lambda^1$ and $q_1(x_\lambda) = 0$, then store such x_λ as a candidate for optimal solution and stop searching along this path. Otherwise, the search along each path stops when the path ceases to exist.

Step 4. Among all the candidate solutions (namely the stored x_λ's) and the feasible solutions in S_0^1, pick up the one with the minimum $q_0(x)$ value, which is then optimal to (P).

This tracking scheme can be accomplished by using Newton's method to solve the parameterized equation

$$(\mathrm{E}_\lambda) \qquad \begin{cases} (Q_0 + \lambda Q_1 + \mu I)x = b_0 + \lambda b_1, \\ \|x\|^2 = 1. \end{cases}$$

The equation (E_λ) may have multiple solutions, and we need to follow all the solutions of (E_λ) belonging to S^i_λ, $i = 0, 1$.

To see that this procedure indeed solves (P), we further quote a result proved by Yuan in [20].

THEOREM 3.4. *There is an optimal solution of* (P) *such that the Hessian matrix of the Lagrangian function when evaluated with optimal Lagrangian multipliers at the optimal solution can have at most one negative eigenvalue.*

Therefore the optimal solution must be contained in S^i_λ, $i = 0, 1$, for $\lambda \geq 0$, and as a consequence the theorem below follows.

THEOREM 3.5. *The above-described KKT solution path(s) tracking procedure solves* (P) *correctly.*

To illustrate how the procedure works, let us consider three examples. The first is

$$(\mathrm{EX}_1) \quad \begin{array}{ll} \text{minimize} & -x_1^2 + x_1 + 4x_2^2 \\ \text{subject to} & x_1^2 + x_2^2 \leq 4, \\ & x_1^2 - 4x_1 + \frac{1}{4}x_2^2 \leq 0. \end{array}$$

Its SDP relaxation is

$$\text{minimize} \quad \begin{bmatrix} 0 & 0.5 & 0 \\ 0.5 & -1 & 0 \\ 0 & 0 & 4 \end{bmatrix} \bullet X$$

$$\text{subject to} \quad \begin{bmatrix} 0 & -2 & 0 \\ -2 & 1 & 0 \\ 0 & 0 & 0.25 \end{bmatrix} \bullet X \leq 0, \qquad \begin{bmatrix} -4 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \bullet X \leq 0,$$

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \bullet X = 1, \quad X \succeq 0.$$

The optimal solution is

$$X^* = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 4 & 0 \\ 0 & 0 & 0 \end{bmatrix},$$

with the optimal value $v^* = -3$. The functions are

$$f(\lambda) = \begin{cases} 12\lambda - 6 & \text{if } 0 \leq \lambda \leq 0.25, \\ -4\lambda - 2 & \text{if } \lambda \geq 0.25, \end{cases}$$

and $h(\lambda) = -4\lambda - 2$ for all $\lambda \geq 0$. We see that $\hat{\lambda} = 0.25$ and $x^- = \left[\begin{smallmatrix} 2 \\ 0 \end{smallmatrix}\right]$. We then follow the trajectory while reducing λ. In this case, $x_\lambda \equiv x^-$, and this leads us to $x^* = x^-$

at $\lambda = 0$. The true optimal value of the original problem is -2. At the optimality, $\mu = 0.25$, and the Hessian matrix of the Lagrangian function is

$$\begin{bmatrix} -1 & 0 \\ 0 & 4 \end{bmatrix} + \mu \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix},$$

which has a negative eigenvalue.

The second example we study is

$$\begin{array}{lll} (\mathrm{EX}_2) & \text{minimize} & -x_1^2 + x_1 + x_2^2 \\ & \text{subject to} & x_1^2 + x_2^2 \le 4, \\ & & (x_1 + x_2)^2 + x_2^2 - 2x_1 \le 0. \end{array}$$

The corresponding value function $f(\lambda)$ attains its maximum at $\hat{\lambda} = 0.5$, and $f(0.5) = -2.3851$. (The numerical values for the computation of this example are rounded.) Moreover, $x^- = \begin{bmatrix} 1.9639 \\ -0.3782 \end{bmatrix}$ and $x^+ = \begin{bmatrix} -1.9639 \\ 0.3782 \end{bmatrix}$. At x^-, $\lambda = \hat{\lambda} = 0.5$ and $\mu = 1.1926$. In our case, Newton's equation amounts to

$$\begin{bmatrix} Q_0 + \lambda Q_1 + \mu I, & x \\ x^T, & 0 \end{bmatrix} \begin{bmatrix} \Delta x \\ \Delta \mu \end{bmatrix} = \begin{bmatrix} (Q_0 + \lambda Q_1 + \mu I)x - b_0 - \lambda b_1 \\ \frac{1}{2}\|x\|^2 - 2 \end{bmatrix},$$

with $Q_0 = \begin{bmatrix} -2 & 0 \\ 0 & 2 \end{bmatrix}$ and $Q_1 = \begin{bmatrix} 2 & 4 \\ 2 & 4 \end{bmatrix}$. Applying Newton's method, the trajectory can be computed as follows:

$$x_{0.5} = \begin{bmatrix} 1.9639 \\ -0.3782 \end{bmatrix}, \quad x_{0.4} = \begin{bmatrix} 1.9731 \\ -0.3267 \end{bmatrix}, \quad x_{0.3} = \begin{bmatrix} 1.9823 \\ -0.2656 \end{bmatrix},$$

$$x_{0.2} = \begin{bmatrix} 1.9907 \\ -0.1925 \end{bmatrix}, \quad x_{0.1} = \begin{bmatrix} 1.9973 \\ -0.1048 \end{bmatrix}, \quad x_0 = \begin{bmatrix} 2 \\ 0 \end{bmatrix}.$$

The optimal solution is found by tracing this trajectory until $\lambda = 0$, i.e., $x^* = x_0$ with $v^* = -2$. The q_1 values at these points are

$$q_1(x_{0.5}) = -1.2703, \quad q_1(x_{0.4}) = -1.1289, \quad q_1(x_{0.3}) = -0.9469,$$

$$q_1(x_{0.2}) = -0.7107, \quad q_1(x_{0.1}) = -0.4023, \quad q_1(x_0) = 0.$$

The last example[3] is

$$\begin{array}{lll} (\mathrm{EX}_3) & \text{minimize} & -x_1^2 + x_1 \\ & \text{subject to} & (x_1 - 2)^2 + (x_2 - 1)^2 - 16 \le 0, \\ & & x_1^2 + x_2^2 - 4 \le 0. \end{array}$$

The function $f(\lambda)$ attains its maximum at $\hat{\lambda} = 0.25$, and $f(0.25) = -5.8125$. The corresponding solutions are $x^- = \begin{bmatrix} 1.9843 \\ 0.2500 \end{bmatrix}$ and $x^+ = \begin{bmatrix} -1.9843 \\ 0.2500 \end{bmatrix}$. The global minimum solution $x^* = \begin{bmatrix} -1.9568 \\ 0.4134 \end{bmatrix}$ is found by increasing λ from $\hat{\lambda}$ and tracking the path starting at x^+ until $\lambda = 0.3649$.

[3]We thank Dr. Ai Wenbao for suggesting this example to us.

4. An extended trust region subproblem and conclusions. In [3], a trust region method was proposed for solving the nonlinear program

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & c(x) = 0, \end{array}$$

where $c(x) : \Re^n \mapsto \Re^m$, i.e., there are m equality constraints.

The subproblem to be solved at each iterative point x^k amounts to

$$\begin{array}{ll} \text{minimize} & d^T \nabla f(x^k) + \frac{1}{2} d^T B_k d \\ \text{subject to} & \|c(x^k) + \nabla c(x^k)^T d\| \leq \xi_k, \\ & \|d\| \leq \Delta_k, \end{array}$$

where $\nabla c(x^k)$ stands for the Jacobian matrix of c evaluated at x^k.

In case $m = 1$, the above problem can be formally written as

$$\begin{array}{lll} \text{(TR)} & \text{minimize} & q_0(x) \\ & \text{subject to} & \|x\|^2 \leq 1, \\ & & -1 \leq \bar{a}^T x - a_0 \leq 1. \end{array}$$

The last constraint is equivalent to $q_1(x) = (\bar{a}^T x - a_0)^2 - 1 \leq 0$.

The optimal solution for (TR) may be binding at the constraint $q_1(x) \leq 0$, or it may not be. However, these two possibilities can be treated separately using the techniques developed in section 3 and subsection 2.3.

The binding case can be immediately dealt with by solving the following quadratic optimization with complementary linear constraints:

$$\begin{array}{ll} \text{minimize} & q_0(x) \\ \text{subject to} & \|x\|^2 \leq 1, \\ & -1 \leq \bar{a}^T x - a_0 \leq 1, \\ & (\bar{a}^T x - a_0 - 1)(\bar{a}^T x - a_0 + 1) = 0. \end{array}$$

As we discussed in subsection 2.3, this can be solved by an SOC-based SDP relaxation in polynomial time.

The remaining task now is to consider the possibility that the constraint $q_1(x) \leq 0$ may not be binding at optimality. If that happens, then the value function $h(\lambda)$ as defined in section 3 attains its maximum value only at $\lambda = 0$. Using Theorem 3.5, we need only to consider solutions generated by the following equation:

$$(\mathrm{E}_0) \begin{cases} (Q_0 + \mu I)x = b_0, \\ \|x\|^2 = 1 \end{cases}$$

for given μ so that $Q_0 + \mu I$ has at most one negative eigenvalue, where we assume $q_0(x) = \frac{1}{2} x^T Q_0 x - b_0^T x$.

The key to note here is that any solution of (E_0) yields the same objective value under q_0, as shown below.

LEMMA 4.1. *Suppose that x and x' both satisfy* (E_0). *Then $q_0(x) = q_0(x')$.*

Proof. Multiplying x^T on both sides of the first equation and rearranging yields

$$q_0(x) = -\frac{\mu}{2} - \frac{1}{2} b_0^T x.$$

On the other hand, we have

$$b_0^T x = b_0^T x' = x^T (Q_0 + \mu I) x'.$$

Hence $q_0(x) = q_0(x')$, as desired. □

The procedure for finding the solution works as follows. We first compute the μ values such that (E_0) has a solution and $Q_0 + \mu I$ has at most one negative eigenvalue. This will result in at most three different μ values. Then, for each of these μ's, solve the following quadratic optimization problem:

$$\begin{array}{ll} \text{minimize} & q_1(x) \\ \text{subject to} & (Q_0 + \mu I)x = b_0, \\ & \|x\|^2 = 1. \end{array}$$

This problem, after variable reduction if necessary, can be solved easily using the SDP relaxation plus decomposition approach; see [18]. If the optimal value of q_1 is positive for every computed μ, then we simply take the solution generated under the binding assumption. Otherwise, we take the solution with the lowest q_0 value among the selected μ's. Summarizing, we have shown the following result.

THEOREM 4.2. *The trust region subproblem arising from a single equality constraint nonlinear program can be solved in polynomial time.*

The computational complexity for minimizing an indefinite quadratic function subject to two convex quadratic constraints remains unsettled. However, as shown in section 3, there exist effective solution procedures for solving the problem. As we saw in section 2.1, there are interesting cases of quadratic optimization with indefinite objective function that can be solved in polynomial time using the SDP relaxation approach. There are several other related unsolved problems. For instance, how to minimize a nonhomogeneous indefinite quadratic function with two homogeneous quadratic constraints? Is there an exact SDP relaxation (cf. section 2.2)? Another problem one may attempt to solve is: Can one formulate an exact SDP relaxation for (TR) (cf. section 4)?

Acknowledgment. We would like to thank Dr. Ai Wenbao for reading this paper carefully and pointing out several errors in an earlier version.

REFERENCES

[1] M. BELLARE AND P. ROGAWAY, *The complexity of approximating a nonlinear program*, Math. Programming, 69 (1995), pp. 429–442.

[2] S. BOYD, L. EL GHAOUI, E. FERON, AND V. BALAKRISHNAN, *Linear Matrix Inequalities in Systems and Control Theory*, SIAM Stud. Appl. Math. 15, SIAM, Philadelphia, 1994.

[3] M.R. CELIS, J.E. DENNIS, AND R.A. TAPIA, *A trust region strategy for nonlinear equality constrained optimization*, in Numerical Optimization 1994 (Proceedings of the SIAM conference on Numerical Optimization, Boulder, CO), P.T. Boggs, R.H. Byrd, and R.B. Schnabel, eds., SIAM, Philadelphia, 1985, pp. 71–82.

[4] M. FU, Z.-Q. LUO, AND Y. YE, *Approximation algorithms for quadratic programming*, J. Comb. Optim., 2 (1998), pp. 29–50.

[5] M.X. GOEMANS AND D.P. WILLIAMSON, *Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming*, J. ACM, 42 (1995), pp. 1115–1145.

[6] F. HAUSDORFF, *Der Wervorrat einer Bilinearform*, Math. Z., 2 (1919), pp. 314–316.

[7] J.B. HIRIART-URRUTY AND C. LEMARÉCHAL, *Convex Analysis and Minimization Algorithms* I, Springer-Verlag, Berlin, 1993.

[8] S. KIM, M. KOJIMA, AND Y. YE, *On Quadratic Minimization with Quadratic Constraints*, discussion note, Department of Management Sciences, The University of Iowa, Iowa City, IA, 2001.

[9] Z.-Q. LUO, J.S. PANG, AND D. RALPH, *Mathematical Programs with Equilibrium Constraints*, Cambridge University Press, Cambridge, UK, 1996.

[10] A. NEMIROVSKII, C. ROOS, AND T. TERLAKY, *On maximization of quadratic form over intersection of ellipsoids with common centers*, Math. Program., 86 (1999), pp. 463–473.

[11] Yu.E. Nesterov, *Semidefinite relaxation and nonconvex quadratic optimization*, Optim. Methods Softw., 9 (1998), pp. 141–160. Special Issue Celebrating the 60th Birthday of Professor Naum Shor.

[12] Yu.E. Nesterov, *Global quadratic optimization via conic relaxation*, in Handbook of Semidefinite Programming, Theory, Algorithms, and Applications, H. Wolkowicz, R. Saigal, and L. Vandenberghe, eds., Kluwer Academic Publishers, Norwell, MA, 2000, pp. 363–387.

[13] Y. Nesterov and A. Nemirovskii, *Interior-Point Polynomial Algorithms in Convex Programming*, SIAM Stud. Appl. Math. 13, Philadelphia, 1994.

[14] B.T. Polyak, *Convexity of quadratic transformations and its use in control and optimization*, J. Optim. Theory Appl., 99 (1998), pp. 553–583.

[15] L. Porkolab and L. Khachiyan, *On the complexity of semidefinite programs*, J. Global Optim., 10 (1997), pp. 351–365.

[16] R.J. Stern and H. Wolkowicz, *Indefinite trust region subproblems and nonsymmetric eigenvalue perturbations*, SIAM J. Optim., 5 (1995), pp. 286–313.

[17] J.F. Sturm, *Theory and algorithms of semidefinite programming*, in High Performance Optimization, H. Frenk, C. Roos, T. Terlaky, and S. Zhang, eds., Kluwer Academic Publishers, Dordrecht, The Netherlands, 2000, pp. 3–194.

[18] J.F. Sturm and S. Zhang, *On cones of nonnegative quadratic functions*, Math. Oper. Res., to appear.

[19] Y. Ye, *Approximating quadratic programming with bound and quadratic constraints*, Math. Program., 84 (1999), pp. 219–226.

[20] Y. Yuan, *On a subproblem of trust region algorithms for constrained optimization*, Math. Programming, 47 (1990), pp. 53–63.

SIAM J. OPTIM.
Vol. 14, No. 1, pp. 268–283

FURTHER RESULTS ON APPROXIMATING NONCONVEX QUADRATIC OPTIMIZATION BY SEMIDEFINITE PROGRAMMING RELAXATION*

PAUL TSENG†

Abstract. We study approximation bounds for the semidefinite programming (SDP) relaxation of quadratically constrained quadratic optimization: $\min f^0(x)$ subject to $f^k(x) \le 0$, $k = 1, \ldots, m$, where $f^k(x) = x^T A^k x + (b^k)^T x + c^k$. In the special case of ellipsoid constraints with interior feasible solution at 0, we show that the SDP relaxation, coupled with a rank-1 decomposition result of Sturm and Zhang [*Math. Oper. Res.*, to appear], yields a feasible solution of the original problem with objective value at most $(1-\gamma)^2/(\sqrt{m}+\gamma)^2$ times the optimal objective value, where $\gamma = \sqrt{\max_k f^k(0) + 1}$. For the single trust-region problem corresponding to $m = 1$, this yields an exact optimal solution. In the general case, we extend some bounds derived by Nesterov [*Optim. Methods Softw.*, 9 (1998), pp. 141–160; working paper, CORE, Université Catholique de Louvain, Louvain-la-Neuve, Belgium, 1998], Ye [*Math. Program.*, 84 (1999), pp. 219–226], and Nesterov, Wolkowicz, and Ye [in *Handbook of Semidefinite Programming*, H. Wolkowicz, R. Saigal, and L. Vandenberghe, eds., Kluwer Academic Publishers, Dordrecht, The Netherlands, 2000, pp. 360–419] for the special case where A^k is diagonal and $b^k = 0$ for $k = 1, \ldots, m$. We also discuss the generation of approximate solutions with high probability.

Key words. quadratically constrained quadratic optimization, semidefinite programming relaxation, approximation algorithm

AMS subject classifications. 90C20, 90C22, 90C26, 90C59

DOI. 10.1137/S1052623401395899

1. Introduction.

Consider the quadratically constrained quadratic program (QP):

$$\text{(1)} \qquad \begin{array}{rcll} v_{\text{QP}} & := & \min & f^0(x) \\ & & \text{s.t.} & f^k(x) \le 0, \quad k = 1, \ldots, m, \end{array}$$

where $f^k(x) = x^T A^k x + (b^k)^T x + c^k$, with $A^k \in \Re^{n \times n}$ symmetric, $b^k \in \Re^n$, $c^k \in \Re$ for $k = 0, 1, \ldots, m$. We assume $c^0 = 0$. If $c^0 \neq 0$, our results still hold by suitably replacing $f^0(x)$ with $f^0(x) - f^0(0)$. This problem is NP-hard.

It was known through the work of Lovász and Schrijver [9], Shor [20], and others that certain NP-hard combinatorial optimization problems can be approximated by semidefinite programming (SDP) problems, for which efficient solution methods exist [1, 14, 15]. This motivated an important work of Goemans and Williamson [8] showing that, for special cases of (1) corresponding to certain NP-hard problems like maximum cut, SDP relaxation yields very good (randomized) approximation algorithms. This work was subsequently extended by Nesterov and colleagues [11, 15], Ye [23, 24], Nemirovski, Roos, and Terlaky [10], and Zhang [26] to other cases of (1), as well as to other combinatorial optimization problems—see [2, 22] and references therein.

*Received by the editors September 29, 2001; accepted for publication (in revised form) January 31, 2003; published electronically July 18, 2003. This research is supported by National Science Foundation grant CCR-9731273.

http://www.siam.org/journals/siopt/14-1/39589.html

†Department of Mathematics, University of Washington, Seattle, WA 98195 (tseng@math.washington.edu).

Motivated by the aforementioned work, in this paper we make further study of the SDP relaxation of (1). In particular, by defining

$$B^k := \begin{bmatrix} A^k & b^k/2 \\ (b^k)^T/2 & c^k \end{bmatrix}, \quad k = 0, 1, \ldots, m,$$

and introducing $x_{n+1} = 1$, we rewrite (1) equivalently as

$$\begin{array}{ll} \min & \sum_{i,j=1}^{n+1} B_{ij}^0 x_i x_j \\ \text{s.t.} & \sum_{i,j=1}^{n+1} B_{ij}^k x_i x_j \le 0, \quad k = 1, \ldots, m, \\ & x_{n+1} = 1. \end{array}$$

By further making the transformation $X = xx^T = [x_i x_j]_{i,j=1}^{n+1}$ for $x \in \Re^{n+1}$ with $x_{n+1} = 1$, we write the above problem equivalently as

$$\begin{array}{ll} \min & \langle B^0, X \rangle \\ \text{s.t.} & \langle B^k, X \rangle \le 0, \quad k = 1, \ldots, m, \\ & X_{n+1n+1} = 1, \quad X \succeq 0, \quad \text{Rank} X = 1. \end{array}$$

Relaxing the rank-1 constraint yields the SDP relaxation of (1):

$$\begin{array}{rcll} \upsilon_{\text{SDP}} & := & \min & \langle B^0, X \rangle \\ & & \text{s.t.} & \langle B^k, X \rangle \le 0, \quad k = 1, \ldots, m, \\ & & & X_{n+1n+1} = 1, \quad X \succeq 0. \end{array} \tag{2}$$

Clearly $\upsilon_{\text{SDP}} \le \upsilon_{\text{QP}}$. Let ρ_{SDP} denote the optimal objective value of (2) but with minimization replaced by maximization. Below we discuss known upper bounds on υ_{QP} in terms of υ_{SDP} and ρ_{SDP}.

Goemans and Williamson [8] showed that if $m = n$, $A^k = e^k(e^k)^T$, $b^k = 0$, $c^k = -1$ for $k = 1, \ldots, m$, $b^0 = 0$, and $-A^0$ is positive semidefinite with nonpositive off-diagonals and zero row sums, then

$$\upsilon_{\text{QP}} \le (0.87856\ldots)\, \upsilon_{\text{SDP}}. \tag{3}$$

Nesterov [11] showed that if $-A^0$ is allowed to be any positive semidefinite matrix, then

$$\upsilon_{\text{QP}} \le \frac{2}{\pi}\, \upsilon_{\text{SDP}}. \tag{4}$$

Ye [23] and Nesterov [12] showed that this still holds if $A^1, \ldots, A^m$ are further allowed to be diagonal (or mutually commute). Zhang [26] showed that the Goemans–Williamson bound (3) still holds if $A^1, \ldots, A^m$ are similarly allowed to be diagonal (or mutually commute) and $-A^0$ is allowed to have nonzero row sums. Zhang also showed that if instead $-A^0$ has nonnegative off-diagonals, then $\upsilon_{\text{QP}} = \upsilon_{\text{SDP}}$ and an optimal solution of (1) can be easily found from an optimal solution of (2).

Of special interest is the case of ellipsoid constraints:

$$A^k = (F^k)^T F^k, \quad b^k = 2(F^k)^T g^k, \quad c^k = \|g^k\|^2 - h^k, \quad k = 1, \ldots, m, \tag{5}$$

where $F^k \in \Re^{n \times n}$, $g^k \in \Re^n$, $h^k \in \{0, 1\}$, and $\|\cdot\|$ denotes the Euclidean norm. Then $f^k(x) = \|F^k x + g^k\|^2 - h^k$, $k = 1, \ldots, m$. Nemirovski, Roos, and Terlaky [10] showed

that if, in addition, the ellipsoids have a common center and nonempty interior (i.e., $g^k = 0$, $h^k = 1$ for all k) and $\sum_{k=1}^m A^k \succ 0$, then a feasible solution x satisfying

$$f^0(x) \leq \frac{1}{2\ \ln(2(m+1)\mu)}\ v_{\text{SDP}},$$

with $\mu := \min\{m+1, \max_{k=1,\ldots,m} \text{Rank}A^k\}$, can be found from the SDP relaxation using a randomization scheme and then derandomizing. Moreover, they constructed for each $m \geq 3$ a problem instance for which $v_{\text{QP}} \geq \frac{1}{O(\ln m)} v_{\text{SDP}}$. Notice that $\sum_{k=1}^m A^k \succ 0$ implies $\max_k \text{Rank}A^k \geq n/m$. Also, if (1) has a ball constraint, then $\mu = \min\{m+1, n\}$. This result was extended by Ye [24] to allow the ellipsoids not to have a common center but assuming $A^0 \preceq 0$, $b^0 = 0$, and that the origin is an interior feasible solution. Ye showed that a feasible solution $\tilde{x}$ can be randomly generated such that

$$\text{E}\left[f^0(\tilde{x})\right] \leq \frac{(1 - \max_k \|g^k\|)^2}{4\ \ln\left(4mn \cdot \max_{k=1,\ldots,m} \text{Rank}A^k\right)}\ v_{\text{SDP}}.$$

Ye remarked that, for general A^0 and b^0, an additional term depending on v_{SDP} and ρ_{SDP} appears on both sides. We will show that if, in addition to (5), the origin is a relatively interior feasible solution and (2) has an optimal solution, then a feasible solution x satisfying

$$f^0(x) \leq \frac{(1-\gamma)^2}{(\sqrt{\kappa}+\gamma)^2}\ v_{\text{SDP}}, \tag{6}$$

where $\kappa := \text{Card}\{k \in \{1,\ldots,m\} : h^k = 1\}$ and $\gamma := \max_{k:h_k=1} \|g^k\|$, can be found using a rank-1 decomposition procedure of Sturm and Zhang [21, Procedure 1]. Thus, in contrast to the bound of Ye, no assumption is made on A^0 or b^0, and (6) does not involve expectation or n. Also, unlike previous work on SDP relaxation, rank reduction does not involve randomization, and the feasible set need not be bounded. In the case of ellipsoids with a common center (i.e., $g^k = 0$ and $h^k = 1$ for all k), (6) reduces to

$$f^0(x) \leq \frac{1}{m} v_{\text{SDP}}.$$

For $m \leq 11$, this improves on the above bound of Nemirovski, Roos, and Terlaky. For $m = 1$, (1) and (5) correspond to the single trust-region problem (see [2, 6]), and (6) implies that an exact optimal solution can be found by solving the SDP relaxation (2). A similar result was obtained in [21] in a more general context. For $m = 2$, (1) and (5) correspond to the two-ellipsoid trust-region problem, which is also of importance—see [2, 16, 25], [6, section 15.4.3], and references therein. Our result provides a practical way to compute an approximate solution to this problem by solving a single SDP. Such an approximate solution is sufficient for achieving convergence of the associated trust-region method. In the special case of *homogeneous* objective and two ellipsoids with a common center (i.e., $b^0 = 0$, $m = 2$, $g^k = 0$, and $h^k = 1$ for $k = 1, 2$), it was shown by Polyak [17, section 6.1] and rediscovered by Ye and Zhang [25, Theorem 2] that an exact optimal solution can be found by solving an SDP.

The work of Nesterov, Ye, and colleagues [11, 12, 15, 23] showed a more general result than (4); namely, if

$$\mathcal{I} := \{k \in \{1,\ldots,m\} : A^k \text{ is diagonal and } b^k = 0\} \tag{7}$$

equals $\{1, \ldots, m\}$, then

$$v_{\text{QP}} \leq \frac{2}{\pi}\, v_{\text{SDP}} + \left(1 - \frac{2}{\pi}\right) \rho_{\text{SDP}}.$$

Our second result is an extension of the above bound to the general case where $\mathcal{I} \neq \{1, \ldots, m\}$. In particular, we show that an $\tilde{x}$ can be randomly generated to satisfy $f^k(\tilde{x}) \leq 0$, $k \in \mathcal{I}$, with probability 1 and

$$\text{E}\left[f^0(\tilde{x})\right] \leq \frac{2}{\pi}\, v_{\text{SDP}} + \left(1 - \frac{2}{\pi}\right) \rho^0_{\text{SDP}}, \tag{8}$$

$$\text{E}\left[f^\ell(\tilde{x})\right] \leq \left(1 - \frac{2}{\pi}\right) \rho^\ell_{\text{SDP}}, \quad \ell \in \{1, \ldots, m\} \setminus \mathcal{I}, \tag{9}$$

where ρ^ℓ_{SDP}, $\ell \notin \mathcal{I}$, are defined similarly as ρ_{SDP}, but with B^0 replaced by B^ℓ and with the inequality constraints not indexed by $\mathcal{I}$ dropped—see (19). An alternative bound that seems generally sharper is also considered. By using a large deviation result, the above bounds holding in expectation can be replaced by bounds holding with high probability—see section 4. In the case where the constraints not indexed by $\mathcal{I}$ are ellipsoid constraints, we discuss ways to randomly generate feasible solutions that, with high probability, satisfy related bounds on the objective value—see Theorem 4 and the subsequent discussions.

Other approximation results for special cases of (1), *not* based on SDP, are discussed in [7, 12, 13, 15, 23]. In particular, for ellipsoid constraints with feasible set having nonempty bounded interior, Fu, Luo, and Ye [7] showed that, for a fixed $\epsilon > 0$ near 0, a feasible solution x with

$$f^0(x) \leq \frac{1 - \epsilon}{m^2(1 + \epsilon)^2}\, v_{\text{QP}}$$

can be found by using a column generation method to find an inexact analytic center of the feasible set and then minimizing f^0 over a Dikin ellipsoid centered there. The computational effort depends on $\ln(1/\epsilon)$ and $\ln(1/\delta)$, where δ is the radius of a Euclidean ball contained in the feasible set. The bound (6) improves on the above bound by a factor of $O(m)$, provided that γ is uniformly bounded away from zero. Some results of Nesterov [13], [15, pp. 376, 377] suggest that, for simplex-type constraints, approximation techniques not based on SDP relaxation might be preferable. However, very recently Bomze and de Klerk [5] presented an SDP-based polynomial-time approximation scheme (PTAS) for QP with a simplex constraint. It would be very interesting if their scheme could be extended to more general constraints. Also, Barvinok [3] showed that, for a bounded number of homogeneous quadratic equations, existence of a nontrivial solution is decidable in a polynomial number of arithmetic operations, using results from real algebraic geometry. It was pointed out by a referee that Barvinok's result can be used to decide in a polynomial number of arithmetic operations whether the system of quadratic inequalities

$$x^T A^0 x \leq v, \qquad x^T A^k x \leq 1, \quad k = 1, \ldots, m,$$

has a solution ($v \in \Re$), assuming $m = O(1)$ and a mild constraint qualification, namely, that the cone generated by $A^0, \ldots, A^m$ contains a positive definite matrix. This in turn can be used to develop, under the same constraint qualification, a PTAS

for QP with a bounded number of common-center-ellipsoid constraints. While SDP relaxation does not provide as sharp an approximation bound in this case, it can be efficiently solved (approximately) and yields a practical algorithm as opposed to (exact) algorithms based on real algebraic geometry.

Throughout, $\Re^n$ denotes the space of n-dimensional column vectors, $\mathcal{S}^n$ denotes the space of $n \times n$ real symmetric matrices, and ^{T} denotes transpose. For $x \in \Re^n$, x_j denotes jth component of x and $\|x\| = \sqrt{x^T x}$. Also, e^k denotes the kth coordinate vector. For $A \in \Re^{m\times n}$, A_{ij} denotes the (i,j)th entry of A. For $A \in \mathcal{S}^n$ with $|A_{ij}| \le 1$ for all i,j, $\arcsin(A)$ denotes the matrix in $\mathcal{S}^n$ with (i,j)th entry $\arcsin(A_{ij})$. For $A, B \in \mathcal{S}^n$, we denote $\langle A, B\rangle = \sum_{i,j} A_{ij}B_{ij}$, and $A \succeq B$ (respectively, $A \succ B$) means that $A - B$ is positive semidefinite (respectively, positive definite). Also, ":=" means "define."

2. SDP relaxation bounds: The ellipsoid constraints case. In this section, we make in addition to (5) the following assumption.

ASSUMPTION 1. *The origin $0 \in \Re^n$ is a feasible solution of* (1) *and $f^k(0) < 0$ whenever $h^k = 1$.*

Assumption 1 is equivalent to $g^k = 0$ whenever $h^k = 0$, and $\|g^k\| = \sqrt{f^k(0) + 1} < 1$ whenever $h^k = 1$. It implies that 0 is in the relative interior of the feasible set of (1), but the converse does not hold. To satisfy Assumption 1, it suffices to find a feasible solution of (1) satisfying strictly those constraints with $h^k = 1$ and then to translate the origin there. Such a feasible solution can be found efficiently by solving

$$\begin{array}{ll} \min & \max\limits_{k:h^k=1} f^k(x) \\ \text{s.t.} & f^k(x) \le 0 \quad \forall k \text{ with } h_k = 0 \end{array}$$

as a second-order cone programming problem [14, p. 221]. Notice that those constraints with $h^k = 0$ are in effect linear constraints. We also make the following assumption.

ASSUMPTION 2. (2) *has an optimal solution X^*.*

It can be seen by using (5) that if the feasible set of (1) is bounded, then so is the feasible set of (2), so that Assumption 2 holds. In the footnote below, we show that if $\{u \in \Re^n : u^T A^0 u \le 0, F^1 u = 0, \ldots, F^m u = 0\} = \{0\}$, then (feasible set of (2)) $\cap$ $\{X : \langle B^0, X\rangle \le 0\}$ is nonempty and bounded, so that Assumption 2 again holds.

We show below that a feasible solution x satisfying (6) can be found efficiently from X^*. Our analysis is based on the following rank-1 decomposition result of Sturm and Zhang [21, Proposition 3].

LEMMA 1. *Let $X \in \mathcal{S}^{n+1}$ be a positive semidefinite matrix of rank r. Let $B \in \mathcal{S}^{n+1}$. Then, $\langle B, X\rangle \le 0$ if and only if there exist $w_j \in \Re^{n+1}$, $j = 1, \ldots, r$, such that*

$$X = \sum_{j=1}^{r} w_j w_j^T \quad \text{and} \quad w_j^T B w_j \le 0, \quad j = 1, \ldots, r.$$

The proof of Lemma 1 is constructive (see [21, Procedure 1]): Given X and B with $\langle B, X\rangle \le 0$, choose any $w_1, \ldots, w_r$ satisfying $X = \sum_{j=1}^r w_j w_j^T$. If $w_j^T B w_j > 0$ for some j, then there is some ℓ with $w_\ell^T B w_\ell < 0$, and we swap w_j and w_ℓ with the linear combinations $(w_j + \alpha w_\ell)/\sqrt{1+\alpha^2}$ and $(w_\ell - \alpha w_j)/\sqrt{1+\alpha^2}$, where α solves $(w_j + \alpha w_\ell)^T B (w_j + \alpha w_\ell) = 0$. Each swap increases the number of w_j with $w_j^T B w_j = 0$ by at least 1, so the desired $w_1, \ldots, w_r$ are found after at most $r - 1$ replacements.

For

$$B := \begin{bmatrix} A^0 & b^0/2 \\ (b^0)^T/2 & -\upsilon_{\rm SDP} \end{bmatrix},$$

we have from $X^*_{n+1n+1} = 1$ and $c^0 = 0$ that $\langle B, X^*\rangle = \langle B^0, X^*\rangle - \upsilon_{\rm SDP} = 0$. Applying Lemma 1 to X^* and B, we can find $w_j = (u_j, t_j) \in \Re^n \times \Re$, $j = 1, \ldots, n+1$, such that

$$X^* = \sum_{j=1}^{n+1} w_j w_j^T \quad \text{and} \quad w_j^T B w_j \le 0, \quad j = 1, \ldots, n+1.$$

Since X^* is a feasible solution of (2), this and (5) yield

$$u_j^T A^0 u_j + t_j (b^0)^T u_j \le \upsilon_{\rm SDP} t_j^2, \quad j = 1, \ldots, n+1, \tag{10}$$

$$\sum_{j=1}^{n+1} \left(u_j^T A^k u_j + t_j (b^k)^T u_j + t_j^2 c^k \right) = \langle B^k, X^*\rangle \le 0, \quad k = 1, \ldots, m, \tag{11}$$

$$\sum_{j=1}^{n+1} t_j^2 = X^*_{n+1n+1} = 1. \tag{12}$$

Using (5), we obtain from (11) and (12) that

$$\sum_{j=1}^{n+1} \|F^k u_j + t_j g^k\|^2 \le h^k, \quad k = 1, \ldots, m. \tag{13}$$

Notice that the above results can be generalized to any feasible solution of (2).[1] If $h^k = 0$, then $g^k = 0$, and thus (13) yields $\|F^k u_j\|^2 = 0$ for all j. Also, summing (13) over all k with $h^k = 1$ yields

$$\sum_{j=1}^{n+1} \sum_{k:h^k=1} \|F^k u_j + t_j g^k\|^2 \le \kappa, \tag{14}$$

where $\kappa := \text{Card}\{k : h^k = 1\}$. We need the following fact.

LEMMA 2. *For any scalars $\kappa \ge 0$, $\alpha_j \ge 0$ and $\beta_j \ge 0$, $j = 1, \ldots, \ell$ ($\ell \ge 1$), such that $\sum_{j=1}^{\ell} \alpha_j \le \kappa$ and $\sum_{j=1}^{\ell} \beta_j = 1$, there exists $\bar{j} \in \{1, \ldots, \ell\}$ such that $\beta_{\bar{j}} > 0$ and $\alpha_{\bar{j}}/\beta_{\bar{j}} \le \kappa$.*

Proof. If the assertion is false, then for every $j \in \{1, \ldots, \ell\}$ such that $\beta_j > 0$ we would have $\alpha_j/\beta_j > \kappa$ or, equivalently, $\alpha_j > \kappa\beta_j$. Then we would have

$$\kappa \ge \sum_{j=1}^{\ell} \alpha_j \ge \sum_{j:\beta_j>0} \alpha_j > \sum_{j:\beta_j>0} \kappa\beta_j = \kappa,$$

[1] In particular, Assumption 1 implies that $X = e^{n+1}(e^{n+1})^T$ is a feasible solution of (2) with $\langle B^0, X\rangle = 0$. Thus, $\mathcal{X}^0 :=$ (feasible set of (2)) $\cap \{X : \langle B^0, X\rangle \le 0\}$ is nonempty. Then for any $X \in \mathcal{X}^0$, repeating the above argument with X and 0 in place of X^* and $\upsilon_{\rm SDP}$ yields $X = \sum_{j=1}^{n+1} w_j w_j^T$ for some $w_j = (u_j, t_j) \in \Re^n \times \Re$ satisfying $u_j^T A^0 u_j + t_j (b^0)^T u_j \le 0$, (12), and (13). For each j, (12) implies that t_j is bounded, while (13) implies $\|F^k u_j + t_j g^k\|^2 \le h^k$. If u_j is unbounded for some j, then dividing by $\|u_j\|^2$ and taking the limit yields a cluster point u of $u_j/\|u_j\|$ satisfying $\|u\| = 1$, $u^T A^0 u \le 0$ and $\|F^k u\|^2 \le 0$, $k = 1, \ldots, m$. Thus, if $\{u \in \Re^n : u^T A^0 u \le 0, F^1 u = 0, \ldots, F^m u = 0\} = \{0\}$, then $\mathcal{X}^0$ is bounded.

a clear contradiction. ∎

By (12) and (14), we can apply Lemma 2 to $\alpha_j = \sum_{k:h^k=1} \|F^k u_j + t_j g^k\|^2$ and $\beta_j = t_j^2$ to conclude the existence of $\bar{j} \in \{1, \ldots, n+1\}$ such that

$$t_{\bar{j}}^2 > 0 \quad \text{and} \quad \sum_{k:h^k=1} \|F^k u_{\bar{j}} + t_{\bar{j}} g^k\|^2 / t_{\bar{j}}^2 \le \kappa.$$

In particular, we can choose $\bar{j}$ to minimize the ratio α_j/β_j over all j with $\beta_j > 0$. Thus

$$\|F^k u_{\bar{j}}/t_{\bar{j}} + g^k\| \le \sqrt{\kappa} \quad \text{whenever } h^k = 1. \tag{15}$$

Let

$$\bar{x} := \begin{cases} u_{\bar{j}}/t_{\bar{j}} & \text{if } (b^0)^T u_{\bar{j}}/t_{\bar{j}} \le 0, \\ -u_{\bar{j}}/t_{\bar{j}} & \text{otherwise,} \end{cases}$$
$$\bar{\tau} := \max\{\tau \in [0,1] : f^k(\tau \bar{x}) \le 0,\ k = 1, \ldots, m\}.$$

Using (10) and (15), we prove below the following result.

THEOREM 1. *Under Assumptions 1, 2 and (5), the above construction yields a feasible solution $x = \bar{\tau}\bar{x}$ of (1) satisfying (6), where $\kappa := \text{Card}\{k \in \{1, \ldots, m\} : h^k = 1\}$ and $\gamma := \max_{k:h_k=1} \|g^k\|$.*

Proof. We estimate $\bar{\tau}$. Fix any $k \in \{1, \ldots, m\}$. Suppose $h^k = 0$. Then we have from $\|F^k u_{\bar{j}}\|^2 = 0$ that $f^k(\tau \bar{x}) = 0$ for all $\tau \in [0,1]$. Suppose $h^k = 1$. Then we see from (15) that if $(b^0)^T u_{\bar{j}}/t_{\bar{j}} \le 0$, then $\|F^k \bar{x} + g^k\| \le \sqrt{\kappa}$, and otherwise

$$\|F^k \bar{x} + g^k\| = \left\| -\left(\frac{F^k u_{\bar{j}}}{t_{\bar{j}}} + g^k \right) + 2g^k \right\| \le \left\| \frac{F^k u_{\bar{j}}}{t_{\bar{j}}} + g^k \right\| + 2\|g^k\| \le \sqrt{\kappa} + 2\|g^k\|.$$

Thus for any $\tau \in [0,1]$ we have

$$\|F^k(\tau \bar{x}) + g^k\| = \|\tau(F^k \bar{x} + g^k) + (1-\tau) g^k\| \le \tau(\sqrt{\kappa} + 2\|g^k\|) + (1-\tau)\|g^k\|.$$

Using $\|g^k\| < 1$, the right-hand side is below 1 (i.e., $f^k(\tau \bar{x}) \le 0$) whenever $\tau \le (1 - \|g^k\|)/(\sqrt{\kappa} + \|g^k\|)$. Thus,

$$\bar{\tau} \ge \min_{k:h_k=1} \frac{1 - \|g^k\|}{\sqrt{\kappa} + \|g^k\|} = \frac{1 - \max_{k:h_k=1} \|g^k\|}{\sqrt{\kappa} + \max_{k:h_k=1} \|g^k\|}, \tag{16}$$

where the equality follows from $(1-\gamma)/(\sqrt{\kappa} + \gamma)$ being a decreasing function of $\gamma \in [0, 1)$. Notice that $\bar{\tau}$ can be easily computed by solving the quadratic equation $\|\tau F^k \bar{x} + g^k\|^2 = 1$ in τ for each k such that $\|F^k \bar{x} + g^k\|^2 > 1$ and then taking the minimum of all the positive roots found.

Finally, our choice of $\bar{x}$ implies $(b^0)^T \bar{x} \le 0$ and $(b^0)^T \bar{x} \le (b^0)^T u_{\bar{j}}/t_{\bar{j}}$. Then for any $\tau \in [0,1]$ we have $\tau \ge \tau^2$ and hence

$$\begin{aligned} f^0(\tau \bar{x}) &= \tau^2 \bar{x}^T A^0 \bar{x} + \tau (b^0)^T \bar{x} \\ &\le \tau^2 \bar{x}^T A^0 \bar{x} + \tau^2 (b^0)^T \bar{x} \\ &\le \tau^2 \bar{x}^T A^0 \bar{x} + \tau^2 (b^0)^T u_{\bar{j}}/t_{\bar{j}} \\ &= \tau^2 (u_{\bar{j}}^T A^0 u_{\bar{j}} + t_{\bar{j}} (b^0)^T u_{\bar{j}})/t_{\bar{j}}^2 \\ &\le \tau^2 v_{\text{SDP}}, \end{aligned}$$

where the last inequality uses (10). Since 0 is a feasible solution of (1) so that $\upsilon_{\text{SDP}} \leq \upsilon_{\text{QP}} \leq f^0(0) = 0$, setting $\tau = \bar{\tau}$ in the above inequality and using (16) completes the proof. □

In the above construction, the main effort lies in solving the SDP relaxation (2), for which many efficient methods exist. Given that an exact optimal solution of (1) is constructed when $m = 1$, we may speculate that when $m = 2$, which is also of considerable interest for trust-region methods (see [2, 16, 25], [6, section 15.4.3], and references therein), a good approximate solution will generally be found. It is worthwhile to test this numerically. In particular, if the rank-1 decomposition given by Lemma 1 is not unique, can we choose one so that the corresponding $\bar{x}$ minimizes $\min_{\tau \in [0,\bar{\tau}]} f^0(\tau \bar{x})$?

3. SDP relaxation bounds: The general case. In this section, following Goemans and Williamson, Nesterov, and Ye, we derive approximation bounds for (1) based on the SDP relaxation (2) under Assumption 2. Notice that Assumption 2 does not guarantee the feasibility of (1), which is NP-hard to check in general. Since $X^* \succeq 0$, we can express

$$X^* = V^T V = [v_i^T v_j]_{i,j=1}^{n+1}$$

for some $V \in \Re^{n+1 \times n+1}$. Here v_i denotes the ith column of V. Choose randomly (according to uniform distribution) v on the unit sphere in $\Re^{n+1}$. Since $\|v_{n+1}\|^2 = X^*_{n+1n+1} = 1$, v_{n+1} also lies on this unit sphere. If $v^T v_{n+1} \leq 0$, then set for $i = 1, \ldots, n+1$

$$\tilde{x}_i = \begin{cases} \sqrt{X^*_{ii}} & \text{if } v^T v_i \leq 0, \\ -\sqrt{X^*_{ii}} & \text{otherwise.} \end{cases}$$

If $v^T v_{n+1} > 0$, then set for $i = 1, \ldots, n+1$

$$\tilde{x}_i = \begin{cases} -\sqrt{X^*_{ii}} & \text{if } v^T v_i \leq 0, \\ \sqrt{X^*_{ii}} & \text{otherwise.} \end{cases}$$

The above choice and $X^*_{n+1n+1} = 1$ ensure that $\tilde{x}_{n+1} = 1$ always.

For each i, j we have that $|\tilde{x}_i \tilde{x}_j| = \sqrt{X^*_{ii} X^*_{jj}}$. If $X^*_{ii} X^*_{jj} \neq 0$, then $\tilde{x}_i \tilde{x}_j = \sqrt{X^*_{ii} X^*_{jj}}$ if and only if $v^T v_i \leq 0$, $v^T v_j \leq 0$ or $v^T v_i > 0$, $v^T v_j > 0$. As was shown by Goemans and Williamson [8], the probability that this event occurs is

$$p = 1 - \frac{1}{\pi} \arccos\left(\frac{v_i^T v_j}{\|v_i\| \|v_j\|}\right) = 1 - \frac{1}{\pi} \arccos\left(\frac{X^*_{ij}}{\sqrt{X^*_{ii} X^*_{jj}}}\right).$$

Thus, the expectation of $\tilde{x}_i \tilde{x}_j$ is

$$\begin{aligned} \mathrm{E}[\tilde{x}_i \tilde{x}_j] &= \sqrt{X^*_{ii} X^*_{jj}}\, p + \left(-\sqrt{X^*_{ii} X^*_{jj}}\right)(1-p) \\ &= \frac{2}{\pi} \sqrt{X^*_{ii} X^*_{jj}} \left(\frac{\pi}{2} - \arccos\left(\frac{X^*_{ij}}{\sqrt{X^*_{ii} X^*_{jj}}}\right)\right) \\ &= \frac{2}{\pi} \sqrt{X^*_{ii} X^*_{jj}} \arcsin\left(\frac{X^*_{ij}}{\sqrt{X^*_{ii} X^*_{jj}}}\right). \end{aligned} \tag{17}$$

If $X^*_{ii}X^*_{jj} = 0$, then $X^*_{ij} = 0$ since $X^* \succeq 0$, and so (17) still holds with the convention that $0/0 = 0$.

Thus, for $k = 0, 1, \ldots, m$, since $\tilde{x}_{n+1} = 1$ always, (17) yields

$$\begin{aligned} \mathrm{E}\left[f^k(\tilde{x})\right] &= \sum_{i,j=1}^{n+1} B^k_{ij}\mathrm{E}[\tilde{x}_i\tilde{x}_j] \\ &= \sum_{i,j=1}^{n+1} B^k_{ij}\frac{2}{\pi}\sqrt{X^*_{ii}X^*_{jj}}\arcsin\left(\frac{X^*_{ij}}{\sqrt{X^*_{ii}X^*_{jj}}}\right) \\ &= \frac{2}{\pi}\langle B^k, D\arcsin(D^{-1}X^*D^{-1})D\rangle, \end{aligned} \tag{18}$$

where $D = \mathrm{diag}[\sqrt{X^*_{ii}}]_{i=1}^{n+1}$. Notice that $|(D^{-1}X^*D^{-1})_{ij}| \leq 1$ for all i, j, and thus $\arcsin(D^{-1}X^*D^{-1})$ is defined. Since $\tilde{x}_i^2 = X^*_{ii}$ always for all i, we have $f^k(\tilde{x}) = \langle B^k, X^*\rangle \leq 0$ always for $k \in \mathcal{I}$ (see (7)).

We now derive bounds on $\mathrm{E}[f^k(\tilde{x})]$, $k \notin \mathcal{I}$, by using (18) and extending an analysis of Nesterov and Ye. We will make, in addition to Assumption 2, the following assumption.

ASSUMPTION 3. *$\{x \in \Re^n : f^k(x) \leq 0,\ k \in \mathcal{I}\}$ is bounded.*

Consider for each $\ell \notin \mathcal{I}$ the following SDP problem:

$$\begin{aligned} \rho^\ell_{\mathrm{SDP}} \ := \ \max \quad & \langle B^\ell, X\rangle \\ \text{s.t.} \quad & \langle B^k, X\rangle = c^k_*, \quad k \in \mathcal{I}, \\ & \langle B^{m+1}, X\rangle = 1, \quad X \succeq 0, \end{aligned} \tag{19}$$

where $B^{m+1} := e^{n+1}(e^{n+1})^T$ and $c^k_* := \langle B^k, X^*\rangle \leq 0$. For $\ell \neq 0$, ρ^ℓ_{SDP} measures how much the ℓth inequality in (2) is violated by the feasible solutions of the inequalities indexed by $\mathcal{I}$. Here we use the tighter constraints $\langle B^k, X\rangle = c^k_*$ instead of $\langle B^k, X\rangle \leq 0$ used by Nesterov, Wolkowicz, and Ye [15]. This yields a tighter upper bound.

Let $\mathcal{X}$ denote the feasible set of (19). Since $X^* \in \mathcal{X}$, $\mathcal{X}$ is nonempty. By Assumption 3, the diagonal entries of $X \in \mathcal{X}$ are bounded, which, together with $X \succeq 0$, implies that $\mathcal{X}$ is bounded. By a result of Rockafellar [19, Theorem 30.4(i)], strong duality holds between (19) and its dual:

$$\begin{aligned} \rho^\ell_{\mathrm{SDP}} \ = \ \inf \quad & \textstyle\sum_{k\in\mathcal{I}} c^k_* y^k + y^{m+1} \\ \text{s.t.} \quad & -B^\ell + \textstyle\sum_{k\in\mathcal{I}\cup\{m+1\}} B^k y^k \succeq 0. \end{aligned} \tag{20}$$

In general, the infimum in (20) need not be attained. As in [11, 15, 23], we make use of the following result of Nesterov [11].

LEMMA 3. *For any $Y \succeq 0$ with $Y_{ii} \leq 1$ for all i, we have* $\arcsin(Y) \succeq Y$.

Fix any $\ell \notin \mathcal{I}$. For each $\epsilon > 0$, let $(y^k)_{k\in\mathcal{I}\cup\{m+1\}}$ be any feasible solution of the dual problem (20) such that $\sum_{k\in\mathcal{I}} c^k_* y^k + y^{m+1} \leq \rho^\ell_{\mathrm{SDP}} + \epsilon$. Let $D := \mathrm{diag}[\sqrt{X^*_{ii}}]_{i=1}^{n+1}$ and $Y := D^{-1}X^*D^{-1}$, with the convention that $Y_{ii} = 1$ if $X^*_{ii} = 0$, and $Y_{ij} = 0$ if $X^*_{ii}X^*_{jj} = 0$ and $i \neq j$. Since $X^* \succeq 0$, then $Y \succeq 0$ and $Y_{ii} = 1$, $i = 1, \ldots, n+1$. Thus

$$
\begin{aligned}
&\langle B^\ell, D\arcsin(Y)D\rangle \\
&\quad= \left\langle B^\ell - \sum_{k\in\mathcal{I}\cup\{m+1\}} B^k y^k, D\arcsin(Y)D \right\rangle + \sum_{k\in\mathcal{I}\cup\{m+1\}} y^k \langle B^k, D\arcsin(Y)D\rangle \\
&\quad\le \left\langle B^\ell - \sum_{k\in\mathcal{I}\cup\{m+1\}} B^k y^k, DYD \right\rangle + \sum_{k\in\mathcal{I}\cup\{m+1\}} y^k \langle B^k, D\arcsin(Y)D\rangle \\
&\quad= \langle B^\ell, X^*\rangle + \left(\frac{\pi}{2}-1\right) \sum_{k\in\mathcal{I}\cup\{m+1\}} y^k \langle B^k, X^*\rangle \\
&\quad= \langle B^\ell, X^*\rangle + \left(\frac{\pi}{2}-1\right) \left(\sum_{k\in\mathcal{I}} y^k c^k_* + y^{m+1}\right) \\
&\quad\le \langle B^\ell, X^*\rangle + \left(\frac{\pi}{2}-1\right)(\rho^\ell_{\text{SDP}} + \epsilon), \qquad (21)
\end{aligned}
$$

where the first inequality uses dual feasibility, Lemma 3, and the fact that $\langle W, Z\rangle \ge 0$ whenever $W \succeq 0, Z \succeq 0$; the second equality uses $DYD = X^*$ and the observations that B^k is diagonal for $k \in \mathcal{I}\cup\{m+1\}$ and that $D\arcsin(Y)D$ has diagonal entries $\frac{\pi}{2}X^*_{ii}$ for all i. Since (21) holds for every $\epsilon > 0$, taking the limit as $\epsilon \to 0$ and using (18) yields

$$
\begin{aligned}
\mathrm{E}\left[f^\ell(\tilde{x})\right] &= \frac{2}{\pi}\langle B^\ell, D\arcsin(D^{-1}X^*D^{-1})D\rangle \\
&= \frac{2}{\pi}\langle B^\ell, D\arcsin(Y)D\rangle \\
&\le \frac{2}{\pi}\langle B^\ell, X^*\rangle + \left(1-\frac{2}{\pi}\right)\rho^\ell_{\text{SDP}}.
\end{aligned}
$$

Since X^* is an optimal solution of (2), this establishes the following result.

THEOREM 2. *Under Assumptions* 2 *and* 3*, the bounds* (8) *and* (9) *hold.*

In the case where $\mathcal{I} = \{1,\dots,m\}$, the above bounds slightly refine analogous bounds obtained by Nesterov and colleagues [11, Theorem 3.3], [15, Theorem 13.2.1] and Ye [23, Theorem 2], [15, Theorem 13.3.2, part 2]. As was considered by Nesterov, Wolkowicz, and Ye [15] (also see [26]), the quadratic inequalities $f^k(x) \le 0$, $k \in \mathcal{I}$, can be generalized to constraints of the form $[x_i^2]_{i=1}^n \in \mathcal{F}$, where $\mathcal{F}$ is a compact convex set intersecting the positive orthant. In this general case, however, the corresponding relaxation may no longer be an SDP problem.

We can also obtain lower bounds analogous to those obtained in the above references. Consider for each $\ell \notin \mathcal{I}$ the following QP:

$$
\begin{aligned}
\upsilon^\ell_{\text{QP}} \;:=\; &\min \quad f^\ell(x) \\
&\text{s.t.} \quad f^k(x) \le 0, \quad k \in \mathcal{I}.
\end{aligned} \qquad (22)
$$

Since $\mathcal{X}$ is nonempty and bounded, then so is the feasible set of this QP. Thus $\upsilon^\ell_{\text{QP}}$ is finite. By the definition of $\mathcal{I}$, we can apply [15, Theorem 13.3.1] or [26, Theorem 1] to reformulate this QP as an equivalent nonlinear program for which X^* and Y, D defined above form a feasible solution with objective function value $\frac{2}{\pi}\langle B^\ell, D\arcsin(Y)D\rangle$. Thus,

$$
\upsilon^\ell_{\text{QP}} \le \frac{2}{\pi}\langle B^\ell, D\arcsin(Y)D\rangle = \mathrm{E}\left[f^\ell(\tilde{x})\right]. \qquad (23)
$$

Notice that $v^0_{\rm QP} \le v_{\rm QP}$, with equality holding when $\mathcal{I} = \{1,\dots,m\}$. If constraints not indexed by $\mathcal{I}$ are ellipsoid constraints, an upper bound on $v_{\rm QP}$ in terms of $v^0_{\rm QP}$ will be derived in section 4.

We can more generally replace $\mathcal{I}$ in (20) by any $\mathcal{I}' \subset \{1,\dots,m\}$ containing $\mathcal{I}$. This would yield a lower $\rho^\ell_{\rm SDP}$, but then the right-hand side of (21) would have an additional term of the form $\sum_{k\in\mathcal{I}'\setminus\mathcal{I}} y^k\langle B^k, D\arcsin(D^{-1}X^*D^{-1})D - \frac{\pi}{2}X^*\rangle$. By Lemma 4 below, this term is at most

$$\Big(\frac{\pi}{2}-1\Big)\sum_{k\in\mathcal{I}'\setminus\mathcal{I}} y^k \sum_{i\ne j}|B_{ij}X^*_{ij}|.$$

Thus, the resulting bound would involve a dual solution y^k, $k\in\mathcal{I}\setminus\mathcal{I}$, as well as X^*.

Below we consider alternative bounds that complement the bounds from Theorem 2. Since $\arcsin(t)$ is convex for $t\in(0,1]$ and has derivative 1 at $t=0$, we have that $1 \le \arcsin(t)/t \le \arcsin(1)/1 = \frac{\pi}{2}$. By symmetry, this holds for $t\in[-1,0)$ as well, so that

$$1 \le \frac{\arcsin(t)}{t} \le \frac{\pi}{2} \qquad \forall t\in[-1,0)\cup(0,1]. \tag{24}$$

Using (24), the following lemma readily follows.

LEMMA 4. *For any $X\succeq 0$ and $B\in\mathcal{S}^{n+1}$, we have*

$$\left|\frac{2}{\pi}\langle B, D\arcsin(D^{-1}XD^{-1})D\rangle - \langle B,X\rangle\right| \le \left(1-\frac{2}{\pi}\right)\sum_{i\ne j}|B_{ij}X_{ij}|,$$

where $D = \mathrm{diag}[\sqrt{X_{ii}}]_{i=1}^{n+1}$.

By using (18) and Lemma 4, we obtain

$$\mathrm{E}\left[f^\ell(\tilde{x})\right] \le \langle B^\ell, X^*\rangle + \left(1-\frac{2}{\pi}\right)\sum_{i\ne j}|B^\ell_{ij}X^*_{ij}|, \quad \ell = 0,1,\dots,m.$$

Since X^* is an optimal solution of (2), the above inequalities yield the following bounds.

THEOREM 3. *Under Assumption 2,*

$$\mathrm{E}\left[f^0(\tilde{x})\right] \le v_{\rm SDP} + \left(1-\frac{2}{\pi}\right)\delta^0,$$
$$\mathrm{E}\left[f^\ell(\tilde{x})\right] \le \left(1-\frac{2}{\pi}\right)\delta^\ell, \quad \ell\in\{1,\dots,m\}\setminus\mathcal{I},$$

where $\delta^\ell := \sum_{i\ne j}|B^\ell_{ij}X^*_{ij}|$.

The bounds in Theorem 3 depend on X^* as well as the off-diagonal quadratic coefficients A^k_{ij}, $i\ne j$, and the linear coefficients b^k_i. While these bounds might look less attractive than the bounds in Theorem 2, they were found to be sharper in all the examples this author tried. For example, if

$$m=n=2, \quad f^0(x) = x_1x_2 + x_1 + x_2, \quad f^1(x) = x_1^2 - 1, \quad f^2(x) = x_2^2 - 1,$$

then $\mathcal{I}=\{1,2\}$ and it is straightforward to verify that

$$v_{\rm QP} = -1, \quad X^* = \begin{bmatrix} 1 & -\frac{1}{2} & -\frac{1}{2} \\ -\frac{1}{2} & 1 & -\frac{1}{2} \\ -\frac{1}{2} & -\frac{1}{2} & 1 \end{bmatrix}, \quad v_{\rm SDP} = -\frac{3}{2}, \quad \rho^0_{\rm SDP} = 3, \quad \delta^0 = \frac{3}{2}.$$

Here δ^0 is smaller than $\rho^0_{\rm SDP} - v_{\rm SDP}$ by a factor of 3!

4. Generating approximate solutions: The general case. The results of section 3 show that $\tilde{x}$ is an approximate solution of (1) in expectation only. In this section we refine this result to generate approximate solutions with high probability.

The following lemma, attributed to Bernstein, refines the Chebychev inequality for bounded random variables. Its proof can be inferred from the argument in [18, pp. 385–386]; a similar result was used in [10]. We note that the probabilistic analysis of Nesterov [11, p. 159] is not applicable here since υ_{SDP} need not be below υ^0_{QP} in Theorem 2, except in the case of $\mathcal{I} = \{1, \ldots, m\}$.

LEMMA 5. *Let ξ be a random variable with standard deviation σ. Suppose $\sigma \le C$ and $|\xi - \mathrm{E}[\xi]| \le K$ always for some constants C and K. Then, for any $t \in (0, C/K]$,*

$$\text{Prob}\left[\xi - \mathrm{E}[\xi] \ge \frac{3}{2}tC\right] \le e^{-t^2/2}.$$

For each $k \notin \mathcal{I}$, let σ^k denote the standard deviation of $f^k(\tilde{x})$. Since $|\tilde{x}_i\tilde{x}_j| = \sqrt{X^*_{ii}X^*_{jj}}$ for all i and j, then $|f^0(\tilde{x}) - \mathrm{E}[f^0(\tilde{x})]| = |\sum_{i\neq j} B^0_{ij}\,(\tilde{x}_i\tilde{x}_j - \mathrm{E}[\tilde{x}_i\tilde{x}_j])\,| \le K$ for some $K > 0$ depending on B^0 and X^*. Applying Lemma 5 with $\xi = f^0(\tilde{x})$ and $t = \frac{2}{3}\epsilon^0$, we obtain that

$$\text{Prob}\left[f^0(\tilde{x}) - \mathrm{E}[f^0(\tilde{x})] \ge \epsilon^0\sigma^0\right] \le e^{-\frac{2}{9}(\epsilon^0)^2},$$

provided that $0 < \epsilon^0 \le \frac{3}{2}\sigma^0/K$. For each $k \in \{1, \ldots, m\} \setminus \mathcal{I}$, we have from the Chebychev inequality that

$$\text{Prob}\left[|f^k(\tilde{x}) - \mathrm{E}[f^k(\tilde{x})]| \ge \epsilon^k\sigma^k\right] \le (\epsilon^k)^{-2},$$

provided that $\epsilon^k > 1$. Then

$$\text{Prob}\left[\max_{k\notin\mathcal{I}}\{f^k(\tilde{x}) - \mathrm{E}[f^k(\tilde{x})] - \epsilon^k\sigma^k\} > 0\right] \;\le\; \pi,$$

where

$$\pi \;:=\; e^{-\frac{2}{9}(\epsilon^0)^2} + \sum_{k\in\{1,\ldots,m\}\setminus\mathcal{I}} (\epsilon^k)^{-2}. \tag{25}$$

For each $k \in \mathcal{I}$, we have $f^k(\tilde{x}) \le 0$ with probability 1. Thus, if we generate $\tilde{x}$ randomly and independently L times, the probability that one of these L samples satisfies

$$f^k(\tilde{x}) \le \mathrm{E}[f^k(\tilde{x})] + \epsilon^k\sigma^k, \quad k = 0, 1, \ldots, m, \tag{26}$$

is at least $1 - \pi^L$. To maintain $\pi < 1$, we must trade off between optimality (small ϵ^0) and feasibility (small ϵ^k for $k \in \{1, \ldots, m\} \setminus \mathcal{I}$). As an example, suppose $\{1, \ldots, m\} \setminus \mathcal{I} = \{1\}$ and $\frac{3}{2}\sigma^0/K^0 \ge 0.5$. If we choose $\epsilon^0 = 0.5$, $\epsilon^1 = 5$, $L = 200$, then this probability is at least 0.940. Notice that 200 is an overestimate. In practice, fewer samples would be needed. Also, $\tilde{x}$ need not be a feasible solution of (1).

To construct feasible solutions with probability 1, we consider the special case where the constraints not indexed by $\mathcal{I}$ are ellipsoid constraints, i.e.,

$$f^k(x) = \|F^k x + g^k\|^2 - 1 \quad \forall k \in \{1, \ldots, m\} \setminus \mathcal{I}, \tag{27}$$

where $F^k \in \Re^{n\times n}$, $g^k \in \Re^n$. We also assume that the origin is a feasible solution of (1) satisfying strictly those constraints not indexed by $\mathcal{I}$. This is equivalent to

$$c^k \le 0 \quad \forall k \in \mathcal{I} \qquad \text{and} \qquad \|g^k\| < 1 \quad \forall k \notin \mathcal{I}. \tag{28}$$

Then, by moving $\tilde{x}$ sufficiently close toward the origin, as was done in section 2, we will construct feasible solutions with certainty. We give more details below.

For each randomly generated $\tilde{x}$, let

$$\bar{x} := \begin{cases} \tilde{x} & \text{if } (b^0)^T\tilde{x} \le 0, \\ -\tilde{x} & \text{otherwise,} \end{cases}$$
$$\bar{\tau} := \max\{\tau \in [0,1] : f^k(\tau\bar{x}) \le 0,\ k = 1,\dots,m\},$$
$$\check{\tau} := \arg\min\{f^0(\tau\bar{x}) : \tau \in [0,\bar{\tau}]\}.$$

Notice that $\bar{\tau}$ and $\check{\tau}$ are well defined and can be easily computed. By using (26)–(28), we obtain the following main result.

THEOREM 4. *Under Assumption* 2 *and* (27), (28), *for any* $0 < \epsilon^0 \le \frac{3}{2}\sigma^0/K^0$ *and integer* $L \ge 1$ *and any* $\epsilon^k > 1$ *and* $\eta^k \ge \mathrm{E}[f^k(\tilde{x})]$ *for* $k \in \{1,\dots,m\} \setminus \mathcal{I}$, *if we generate* $\tilde{x}$ *randomly and independently* L *times as described in section* 3 *and construct* $x = \check{\tau}\bar{x}$ *as above, then each* x *is a feasible solution of* (1) *with probability* 1 *and, with probability of at least* $1 - \pi^L$, *one of these* L *samples satisfies*

$$f^0(x) \le \min_{k\notin\mathcal{I}} \left(\frac{1 - \|g^k\|}{\sqrt{1 + \eta^k + \epsilon^k\sigma^k} + \|g^k\|} \right)^2 (\mathrm{E}[f^0(\tilde{x})] + \epsilon^0\sigma^0), \tag{29}$$

where π *is given by* (25), σ^k *denotes the standard deviation of* $f^k(\tilde{x})$, *and* K *is any constant for which* $|f^0(\tilde{x}) - \mathrm{E}[f^0(\tilde{x})]| \le K$ *always.*

Proof. The probability that one of the samples satisfies (26) is at least $1 - \pi^L$. Consider one such sample $\tilde{x}$ and the corresponding $\bar{x}, \bar{\tau}, \check{\tau}$. For each $k \in \mathcal{I}$, since A^k is diagonal and $b^k = 0$, we see that

$$f^k(\tau\bar{x}) = f^k(\tau\tilde{x}) = \tau^2 f^k(\tilde{x}) + (1-\tau^2)c^k \le 0$$

for all $\tau \in [0,1]$. For each $k \in \{1,\dots,m\} \setminus \mathcal{I}$, we see from (26) and (27) that if $(b^0)^T\tilde{x} \le 0$, then $\|F^k\bar{x} + g^k\| \le \sqrt{\kappa^k}$; otherwise

$$\|F^k\bar{x} + g^k\| = \| - (F^k\tilde{x} + g^k) + 2g^k\| \le \|F^k\tilde{x} + g^k\| + 2\|g^k\| \le \sqrt{\kappa^k} + 2\|g^k\|,$$

where $\kappa^k := 1 + \mathrm{E}[f^k(\tilde{x})] + \epsilon^k\sigma^k$. Thus, arguing identically as in the proof of Theorem 1, we obtain that

$$\bar{\tau} \ge \min_{k\notin\mathcal{I}} \frac{1 - \|g^k\|}{\sqrt{\kappa^k} + \|g^k\|}. \tag{30}$$

Moreover, for all $\tau \in [0,\bar{\tau}]$, $\tau\bar{x}$ is a feasible solution of (1) with probability 1. Since $\check{\tau} \in [0,\bar{\tau}]$, then $x = \check{\tau}\bar{x}$ is a feasible solution of (1) with probability 1.

For each $k \in \{1,\dots,m\} \setminus \mathcal{I}$, since $\mathrm{E}[f^k(\tilde{x})] \le \eta^k$, then $\kappa^k \le 1 + \eta^k + \epsilon^k\sigma^k$, and it follows from (30) that

$$\bar{\tau} \ \ge \ \hat{\tau} \ := \min_{k\notin\mathcal{I}} \frac{1 - \|g^k\|}{\sqrt{1 + \eta^k + \epsilon^k\sigma^k} + \|g^k\|} > 0,$$

implying $\check{\tau} \in (0, \bar{\tau}]$. Finally, our choice of $\bar{x}$ implies $(b^0)^T \bar{x} \leq 0$ and $(b^0)^T \bar{x} \leq (b^0)^T \tilde{x}$. Then, arguing similarly as in the proof of Theorem 1, we obtain for any $\tau \in [0, \bar{\tau}]$ that

$$\begin{aligned} f^0(\check{\tau}\bar{x}) &\leq f^0(\tau \bar{x}) \\ &\leq \tau^2 (\tilde{x}^T A^0 \tilde{x} + (b^0)^T \tilde{x}) \\ &\leq \tau^2 (\mathrm{E}[f^0(\tilde{x})] + \epsilon^0 \sigma^0), \end{aligned}$$

where the last inequality uses (26). Setting $\tau = \hat{\tau}$ completes the proof. □

By Theorem 3, we can choose $\eta^k = \left(1 - \frac{2}{\pi}\right) \delta^k$ in Theorem 4. Then (29) becomes

$$f^0(x) \leq \min_{k \notin \mathcal{I}} \left(\frac{1 - \|g^k\|}{\sqrt{1 + \left(1 - \frac{2}{\pi}\right) \delta^k + \epsilon^k \sigma^k} + \|g^k\|} \right)^2 \left(\upsilon_{\text{SDP}} + \left(1 - \frac{2}{\pi}\right) \delta^0 + \epsilon^0 \sigma^0 \right).$$

If Assumption 3 also holds, then, by Theorem 2, we can choose $\eta^k = \left(1 - \frac{2}{\pi}\right) \rho^k_{\text{SDP}}$ in Theorem 4. Then (29) becomes

$$f^0(x) \leq \min_{k \notin \mathcal{K}} \left(\frac{1 - \|g^k\|}{\sqrt{1 + \left(1 - \frac{2}{\pi}\right) \rho^k_{\text{SDP}} + \epsilon} + \|g^k\|} \right)^2 \left(\frac{2}{\pi}\, \upsilon_{\text{SDP}} + \left(1 - \frac{2}{\pi}\right) \rho^0_{\text{SDP}} + \epsilon \right).$$

If Assumption 3 also holds, then (22) with $\ell = 0$ has an optimal solution, say x^0. Since $-x^0$ is also a feasible solution of (22), then $(b^0)^T x^0 \leq 0$. By an argument similar to the proof of Theorem 4, it can be shown that tx^0 is a feasible solution of (1) whenever

$$0 \leq t \leq \min_{k : \|F^k x^0 + g^k\| > 1} \frac{1 - \|g^k\|}{\|F^k x^0 + g^k\| - \|g^k\|}.$$

Moreover, $f^0(tx^0) \leq t^2 f^0(x^0) = t^2 \upsilon^0_{\text{QP}}$. Since tx^0 is a feasible solution of (1), this implies $\upsilon_{\text{QP}} \leq f^0(tx^0) \leq t^2 \upsilon^0_{\text{QP}}$. Thus, we obtain the following upper bound on υ_{QP} in terms of υ^0_{QP}:

$$\upsilon_{\text{QP}} \leq \min_{k : \|F^k x^0 + g^k\| > 1} \left(\frac{1 - \|g^k\|}{\|F^k x^0 + g^k\| - \|g^k\|} \right)^2 \upsilon^0_{\text{QP}}.$$

This can be combined with the lower bound (23) and the upper bound (8) to yield bounds involving mainly $\upsilon_{\text{QP}}, \upsilon_{\text{SDP}}, \rho^0_{\text{SDP}}$ and quantities depending on x^0.

The bound (29) works best when σ^k is small for all $k \notin \mathcal{I}$ and when σ^0/K is not too small. Then we can set $\epsilon^0 = \frac{3}{2}\sigma^0/K$ and choose moderately large ϵ^k, $k \in \{1, \ldots, m\} \setminus \mathcal{I}$, to satisfy $\pi < 1$. Both σ^k and K can be estimated as follows. Since $\tilde{x}_i^2 = \mathrm{E}[\tilde{x}_i^2]$ and $|\tilde{x}_i \tilde{x}_j| = \sqrt{X^*_{ii} X^*_{jj}}$ for all i and j, we have

$$\begin{aligned} (\sigma^k)^2 &= \mathrm{E}\left[\left(\sum_{i \neq j} B^k_{ij} \tilde{x}_i \tilde{x}_j - B^k_{ij} \mathrm{E}[\tilde{x}_i \tilde{x}_j] \right)^2 \right] \\ &= \mathrm{E}\left[\left(\sum_{i \neq j} B^k_{ij} \tilde{x}_i \tilde{x}_j \right)^2 \right] - \left(\sum_{i \neq j} B^k_{ij} \mathrm{E}[\tilde{x}_i \tilde{x}_j] \right)^2 \\ &\leq \mathrm{E}\left[\left(\sum_{i \neq j} |B^k_{ij}| |\tilde{x}_i \tilde{x}_j| \right)^2 \right] \quad = \quad (\Delta^k)^2, \end{aligned}$$

where $\Delta^k := \sum_{i\neq j} |B^k_{ij}|\sqrt{X^*_{ii}X^*_{jj}}$. In fact, it can be shown by using (17), [8, Lemma 7.3.1], and [4, Corollary 18.6.10] that

$$\begin{aligned}(\sigma^k)^2 &= \sum_{i\neq j}\sum_{p\neq q} B^k_{ij}B^k_{pq}\left(\mathrm{E}[\tilde{x}_i\tilde{x}_j\tilde{x}_p\tilde{x}_q] - \mathrm{E}[\tilde{x}_i\tilde{x}_j]\mathrm{E}[\tilde{x}_p\tilde{x}_q]\right)\\ &\leq 4\sum_{i\neq j}\sum_{p\neq q} |B^k_{ij}||B^k_{pq}|\sqrt{X^*_{ii}X^*_{jj}X^*_{pp}X^*_{qq}}\left(1-\max\{|\nu_{ij}|,|\nu_{pq}|\}\right),\end{aligned}$$

where $\nu_{ij} := \frac{2}{\pi}\arcsin(X^*_{ij}/\sqrt{X^*_{ii}X^*_{jj}})$. Thus, σ^k is small if either B^k has small off-diagonal entries or if $|\nu_{ij}|$ is near 1 (i.e., $|X^*_{ij}|$ is near $\sqrt{X^*_{ii}X^*_{jj}}$) for many index pairs $i\neq j$. Also, since $|\tilde{x}_i\tilde{x}_j| = \sqrt{X^*_{ii}X^*_{jj}}$ and, by (17) and $|t| \leq |\arcsin(t)| \leq \frac{\pi}{2}|t|$ (see (24)), we have $|\mathrm{E}[\tilde{x}_i\tilde{x}_j]| \leq \sqrt{X^*_{ii}X^*_{jj}}$ for all i,j, implying

$$|f^k(\tilde{x}) - \mathrm{E}[f^k(\tilde{x})]| = \left|\sum_{i\neq j} B^k_{ij}(\tilde{x}_i\tilde{x}_j - \mathrm{E}[\tilde{x}_i\tilde{x}_j])\right| \leq 2\Delta^k$$

and hence $K \leq 2\Delta^0$. Notice that since $X^* \succeq 0$ so that $|X^*_{ij}| \leq \sqrt{X^*_{ii}X^*_{jj}} = |\tilde{x}_i\tilde{x}_j|$ and $X^*_{n+1n+1} = \tilde{x}_{n+1} = 1$, we have $\Delta^k = \sum_{i\neq j}|A^\ell_{ij}\tilde{x}_i\tilde{x}_j| + \sum_i |b^\ell_i\tilde{x}_i| \geq \delta^k$.

If the ellipsoid constraint functions (27) have some special structure, then the bound (29) can be sharpened. For example, suppose these ellipsoid constraints come in pairs of the form

$$\frac{1}{\alpha}(x_i - x_j)^2 \leq 1, \qquad \frac{1}{\beta}(x_i + x_j)^2 \leq 1, \tag{31}$$

for some $i\neq j$ and some $\alpha > 0, \beta > 0$. Then $X^*_{ii} + X^*_{jj} - 2X^*_{ij} \leq \alpha$ and $X^*_{ii} + X^*_{jj} + 2X^*_{ij} \leq \beta$, implying $2X^*_{ii} + 2X^*_{jj} \leq \alpha + \beta$. Since $\tilde{x}^2_i = X^*_{ii}$ and $\tilde{x}^2_j = X^*_{jj}$, this yields

$$(\tilde{x}_i - \tilde{x}_j)^2 = \tilde{x}^2_i + \tilde{x}^2_j - 2\tilde{x}_i\tilde{x}_j \leq X^*_{ii} + X^*_{jj} + 2\sqrt{X^*_{ii}X^*_{jj}} \leq 2X^*_{ii} + 2X^*_{jj} \leq \alpha + \beta$$

always. Similarly, $(\tilde{x}_i + \tilde{x}_j)^2 \leq \alpha + \beta$ always. It follows that $x = \tau\bar{x}$ satisfies (31) for all $\tau \in [0, \sqrt{\min\{\alpha,\beta\}/(\alpha+\beta)}]$. Then it is readily seen from the proof of Theorem 4 that

$$\mathrm{E}[f^0(\check{\tau}\bar{x})] \leq \min\left\{\frac{\min\{\alpha,\beta\}}{\alpha+\beta}\right\}\mathrm{E}[f^0(\tilde{x})],$$

where the first minimum is taken over all pairs. If only a subset of the ellipsoid constraints come in pairs of the form (31), the bound (29) can be sharpened analogously.

Acknowledgments. The author thanks the two referees for their helpful comments on the original version of this paper. One referee, in particular, pointed out the references [3, 5, 17].

REFERENCES

[1] F. Alizadeh, *Interior point methods in semidefinite programming with applications to combinatorial optimization*, SIAM J. Optim., 5 (1995), pp. 13–51.

[2] K. Anstreicher and H. Wolkowicz, *On Lagrangian relaxation of quadratic matrix constraints*, SIAM J. Matrix Anal. Appl., 22 (2000), pp. 41–55.

[3] A. I. Barvinok, *Feasibility testing for systems of real quadratic functions*, Discrete Comput. Geom., 10 (1993), pp. 1–13.
[4] M. Berger, *Geometry* II, Springer-Verlag, Berlin, 1987.
[5] I. M. Bomze and E. de Klerk, *Solving standard quadratic optimization problems via linear, semidefinite and copositive programming*, J. Global Optim., 24 (2002), pp. 163–185.
[6] A. R. Conn, N. I. M. Gould, and Ph. L. Toint, *Trust-Region Methods*, MPS-SIAM Ser. Optim. MP01, SIAM, Philadelphia, PA, 2000.
[7] M. Fu, Z.-Q. Luo, and Y. Ye, *Approximation algorithms for quadratic programming*, J. Comb. Optim., 2 (1998), pp. 29–50.
[8] M. X. Goemans and D. P. Williamson, *Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming*, J. ACM, 42 (1995), pp. 1115–1145.
[9] L. Lovász and A. Schrijver, *Cones of matrices and set-functions and* 0–1 *optimization*, SIAM J. Optim., 1 (1991), pp. 166–190.
[10] A. Nemirovski, C. Roos, and T. Terlaky, *On maximization of quadratic form over intersection of ellipsoids with common center*, Math. Program., 86 (1999), pp. 463–473.
[11] Y. Nesterov, *Semidefinite relaxation and nonconvex quadratic optimization*, Optim. Methods Softw., 9 (1998), pp. 141–160.
[12] Y. Nesterov, *Global Quadratic Optimization via Conic Relaxation*, working paper, CORE, Université Catholique de Louvain, Louvain-la-Neuve, Belgium, 1998.
[13] Y. Nesterov, *Global Quadratic Optimization on the Sets with Simplex Structure*, working paper, CORE, Université Catholique de Louvain, Louvain-la-Neuve, Belgium, 1999.
[14] Y. Nesterov and A. Nemirovskii, *Interior-Point Polynomial Algorithms in Convex Programming*, SIAM Stud. Appl. Math. 13, SIAM, Philadelphia, PA, 1994.
[15] Y. Nesterov, H. Wolkowicz, and Y. Ye, *Semidefinite programming relaxations of nonconvex quadratic optimization*, in Handbook of Semidefinite Programming, H. Wolkowicz, R. Saigal, and L. Vandenberghe, eds., Kluwer Academic Publishers, Dordrecht, The Netherlands, 2000, pp. 360–419.
[16] J.-M. Peng and Y.-X. Yuan, *Optimality conditions for the minimization of a quadratic with two quadratic constraints*, SIAM J. Optim., 7 (1997), pp. 579–594.
[17] B. T. Polyak, *Convexity of quadratic transformations and its use in control and optimization*, J. Optim. Theory Appl., 99 (1998), pp. 553–583.
[18] A. Rényi, *Probability Theory*, North-Holland, Amsterdam, 1970.
[19] R. T. Rockafellar, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.
[20] N. Z. Shor, *Quadratic optimization problems*, Soviet J. Comput. Systems Sci., 25 (1987), pp. 1–11.
[21] J. F. Sturm and S. Zhang, *On cones of nonnegative quadratic functions*, Math. Oper. Res., to appear.
[22] H. Wolkowicz, R. Saigal, and L. Vandenberghe, eds., *Handbook of Semidefinite Programming*, Kluwer Academic Publishers, Dordrecht, The Netherlands, 2000.
[23] Y. Ye, *Approximating quadratic programming with bound and quadratic constraints*, Math. Program., 84 (1999), pp. 219–226.
[24] Y. Ye, *Approximating global quadratic optimization with convex quadratic constraints*, J. Global Optim., 15 (1999), pp. 1–17.
[25] Y. Ye and S. Zhang, *New results on quadratic minimization*, SIAM J. Optim., 14 (2003), pp. 245–267.
[26] S. Zhang, *Quadratic maximization and semidefinite relaxation*, Math. Program., 87 (2000), pp. 453–465.

SIAM J. OPTIM.
Vol. 14, No. 1, pp. 284–305

RADIAL EPIDERIVATIVES AND ASYMPTOTIC FUNCTIONS IN NONCONVEX VECTOR OPTIMIZATION*

FABIÁN FLORES-BAZÁN†

Abstract. The notions of lower and upper radial epiderivatives (a radial notion different from that generally discussed) for nonconvex vector-valued functions are introduced, many of their properties are established, and some of the optimality conditions for a point to be an ideal, Pareto, or weak-Pareto minimum involving these epiderivatives are also presented. In particular, a characterization for the ideal minima in terms of the lower radial epiderivative is proved. Such a result appears to be new in the literature. Under some mild assumptions on the given function, it is proved that the asymptotic cone of its epigraph is the epigraph of its upper radial epiderivative. Moreover, given a vector minimization problem, we describe the asymptotic behavior of its solution set by introducing some cones of asymptotic directions of the function involved. Finally, we define the lower and upper radial subdifferentials and express the optimality conditions by means of these subdifferentials. Certainly these optimality conditions subsume various necessary or sufficient conditions found in the literature for convex or nonconvex functions.

Key words. nonconvex vector-valued optimization, strong minima, Pareto minima, weak-Pareto minima, closed radial cone, epiderivatives, optimality conditions, subdifferential, nonsmooth analysis

AMS subject classifications. 46G05, 49J52, 90C26, 90C29, 90C46, 90C48

DOI. 10.1137/S1052623401392111

1. Introduction. Although the study of vector-valued functions, motivated mainly by optimization problems, started around 1970, there is not a unified theory on this subject. However, a common approach to carrying out the study of such problems, as well as set-valued problems, seems to be the use of a notion of derivatives for vector (set)-valued functions. In this regard, the contingent cone (historically the tangent cone of Bouligand, or the set of adherent displacements) has been widely applied in developing a nonsmooth analysis or in deriving optimality conditions; see [A, AF, C, CJ, GJ, JR, HU1, HU2, L1, L2, P1, P2, P3] among others. After the introduction of Clarke's tangent cone [Cl], many people used this generalized tangent cone (see [Th1, Th2, Th3] and references therein) to develop a similar theory suitable for (nonconvex) locally Lipschitz continuous vector-valued functions.

In [P1] several notions of (what are now called) epiderivatives (according to the cone involved: contingent, Clarke's, ...) for real-valued functions were considered. Later, a more complete treatment for vector-valued functions was presented in [P2], where some necessary optimality conditions were derived in terms of these kinds of derivatives. Independently, the notion of a contingent epiderivative for functions taking values in the real line appears in [A], under the name of "upper contingent derivative"; see also [AF, section 6.1] for further developments.

Very recently, perhaps motivated by the real case, the notion of the (single-valued) contingent epiderivative for set-valued maps was introduced in [JR]. More precisely,

*Received by the editors July 11, 2001; accepted for publication (in revised form) January 10, 2003; published electronically July 18, 2003. Part of this work was carried out while the author was visiting ICTP (Trieste, Italy) as a regular associate. This research was also partly supported by FONDECYT 101-0116 and FONDAP-Matemáticas Aplicadas II (Chile).

http://www.siam.org/journals/siopt/14-1/39211.html

†Universidad de Concepción, Facultad de Ciencias Físicas y Matemáticas, Grupo de Investigación Interdisciplinario en Matemática Aplicada, Departamento de Ingeniería Matemática, Casilla 160-C, Concepción, Chile (fflores@ing-mat.udec.cl).

if $F : X \to 2^Y$, then its contingent epiderivative at $(\bar{x}, \bar{y})$, $\bar{y} \in F(\bar{x})$, is defined by Jahn and Rauh in [JR] as the single-valued function $D_eF(\bar{x}, \bar{y}) : X \to Y$ such that

$$\operatorname{epi}(D_eF(\bar{x}, \bar{y})) = T(\operatorname{epi} F; (\bar{x}, \bar{y})). \tag{1.1}$$

Here, $T(A; a)$ stands for the contingent cone of A at $a \in A$. By using this notion, one can formulate optimality conditions that are necessary and sufficient under the convexity assumption, as shown in [JR, CJ], and a Lagrange multiplier rule, as established in [GJ]. The convexity is imposed to guarantee the global character of the contingent cone. However, a drawback of considering such a notion of epiderivative is the nonexistence of a formula for it, except for the special cases when $D_eF(\bar{x}, \bar{y})(u) \in Y$ for all $u \in X$ [JR, CJ] and when $Y = \mathbb{R}$ [JR], [AF, section 6.1]. In both cases, one obtains

$$D_eF(\bar{x}, \bar{y}) = \inf\left\{v \in Y : (u, v) \in T(\operatorname{epi} F; (\bar{x}, \bar{y}))\right\}, \tag{1.2}$$

where the infimum is taken with respect to the ordering cone in Y.

In order to deal with nonconvex set-valued optimization problems, the author proposes in [F3] an alternative definition of epiderivative based on the closed radial cone, which differs from the radial notion discussed in [P1, P2]: the closed radial cone of A at $a \in A$, denoted by $R(A; a)$, is the smallest closed cone containing $A - a$, which, in the case in which A is convex, coincides with the contingent cone. Thus, a necessary and sufficient optimality condition in nonconvex set-valued optimization is established in [F3], without assuming the existence of the alternative epiderivative as a function taking values in Y, but under a restrictive assumption on the ordering cone (but including the scalar case). In both cases, we also have a formula for the radial epiderivative like (1.2). We point out that the closed radial cone was also used in [T] in a different framework.

In the present paper, we restrict ourselves to vector-valued functions, since one cannot expect, even in the single-valued case ($Y \neq \mathbb{R}$), that a condition like (1.1) will be satisfied for any kind of epiderivative, as Propositions 3.1 and 3.8 show; see also Proposition 3.2 in [P2]. Thus, the purpose of this paper is to contribute to a better understanding of the phenomena arising in the study of nonconvex vector optimization problems. Our approach is based on some derivative notions defined in terms of the closed radial cone.

In the next section we recall some basic definitions and facts related to vector-valued functions. In section 3, after defining the lower and upper radial epiderivatives for vector-valued functions, we study some of their main properties to be used in what follows. In particular, we characterize the ideal (strong) minimizers in terms of the lower radial epiderivative, as well as some sufficient or necessary conditions for the Pareto or weak-Pareto minima in terms of these radial epiderivatives (see Corollary 3.5). We establish, under some mild assumptions, that the epigraph of the upper radial epiderivative is the asymptotic cone of the epigraph of the function involved (Theorem 3.10). Thus, it gives rise to a notion of asymptotic function for vector-valued functions. We close section 3 by establishing a formula for the upper radial epiderivative in the finite dimensional case and under convexity assumptions.

Section 4 is devoted to describing the asymptotic behavior of the solution set for a vector minimization problem by applying the previous results. Indeed, two cones of asymptotic directions for the function involved are introduced. Such cones are estimates for the asymptotic cone of the solution set to the minimization problem (Theorem 4.2). In particular, we find an expression for the asymptotic cone of the set

of ideal (strong) minimizers in terms of the upper radial epiderivative (Theorems 4.3 and 4.4). Finally, some kinds of subdifferentials involving the lower or upper radial epiderivatives are introduced, and some optimality conditions are written by means of these subdifferentials. Certainly such optimality conditions give more information than those appearing in the literature.

2. Some preliminary facts. As we said in the introduction, our main concern is dealing with nonconvex vector-valued functions. Thus, we need an object taking into account the global behavior of the function involved. Although the so-called contingent cone has proved to be very useful in the study of nonsmooth problems in scalar and set-valued optimization [A, AF, C, CJ, GJ, JR, L1, L2], it captures only the local nature of the function. Therefore, we shall use the closed radial cone to define some kinds of epiderivatives for vector-valued functions. This will be carried out in the next section.

In all this paper X stands for any real normed vector space, and, given any set C in X, $\overline{C}$ will denote its closure. We first recall some basic notions.

DEFINITION 2.1. *Given any nonempty set $C \subset X$, $\bar{x} \in \overline{C}$, we define the following cones:*

(i) *the contingent cone of C (or tangent cone of Bouligand) at $\bar{x}$, denoted by $T(C;\bar{x})$, is the set of all $v \in X$ such that there exist sequences $t_n \downarrow 0$ and $v_n \to v$ with $\bar{x} + t_n v_n \in C$ for all $n \in \mathbb{N}$;*

(ii) *the closed radial cone of C at $\bar{x}$, denoted by $R(C;\bar{x})$, is the set of all $v \in X$ such that there exist sequences $t_n > 0$, $v_n \to v$, and $\bar{x} + t_n v_n \in C$ for all $n \in \mathbb{N}$;*

(iii) *the interiorly radial cone of C at $\bar{x}$, denoted by $R^i(C;\bar{x})$, is the set of all $v \in X$ such that there exists $\varepsilon > 0$ satisfying $\bar{x} + tv' \in C$ for all $t > 0$, $||v' - v|| < \varepsilon$.*

By a cone we mean a set K satisfying $\lambda K \subset K$ for all $\lambda \geq 0$, so $0 \in K$.

Remark 2.2. (a) Some of the equivalent definitions for $T(C;\bar{x})$ are the following:

- $v \in T(C;\bar{x})$ if and only if there exist sequences $t_n > 0$ and $v_n \to v$ such that $t_n v_n \to 0$ and $\bar{x} + t_n v_n \in C$ for all $n \in \mathbb{N}$;
- $v \in T(C;\bar{x})$ if and only if there exist sequences $t_n > 0$ and $x_n \in C$ such that $x_n \to \bar{x}$ and $t_n(x_n - \bar{x}) \to v$.

(b) $T(C;\bar{x})$ and $R(C;\bar{x})$ are nonempty closed cones.

(c) $T(C;\bar{x}) = R(C;\bar{x})$ for all $\bar{x} \in C$ whenever C is a convex set.

(d) $R^i(C;\bar{x})$ is an open set whenever $\bar{x}$ is a boundary point of C. Furthermore, $\lambda R^i(C;\bar{x}) \subset R^i(C;\bar{x})$ for all $\lambda > 0$. Indeed $R^i(C;\bar{x}) = X \setminus R(X \setminus C;\bar{x})$. Moreover,

$$\begin{aligned} v \in R^i(C;\bar{x}) &\iff \exists\, \varepsilon > 0 \text{ such that } v' \in \bigcap_{t>0} t(C - \bar{x}), \qquad ||v' - v|| < \varepsilon, \\ &\iff v \in \operatorname{int}\left(\bigcap_{t>0} t(C - \bar{x})\right). \end{aligned} \tag{2.1}$$

Given any closed set $K \subset X$, we define the asymptotic cone of K as the closed cone

$$K^\infty = \Big\{ v \in X : \exists\, t_n \downarrow 0, \exists\, x_n \in K,\ t_n x_n \to v \Big\}.$$

We set $\emptyset^\infty = \emptyset$. The term "recession cone" is used when the set is also convex.

A closed set K it said to be radiant at $\bar{x} \in K$ if there exists some $\delta \in\ ]0,1]$ such that $\bar{x} + t(x - \bar{x}) \in K$ for all $x \in K$, for all $t \in\ ]0,\delta]$.

We recall that a subset K is star-shaped with respect to some point $\bar{x} \in K$ if $\bar{x} + t(x - \bar{x}) \in K$ for all $t \in [0,1]$ and all $x \in K$. Thus, one immediately sees that every set that is star-shaped with respect to a point in the set is radiant at the same point. In particular, closed convex sets are radiant at any point belonging to the set. Let K be radiant at $\bar{x} \in K$; it is proved in [D1, D2] that

$$K^\infty = \bigcap_{t>0} t(K - \bar{x}). \tag{2.2}$$

Consequently, in the case in which K is convex, given any $\bar{x} \in K$,

$$K^\infty = \Big\{ v \in X : \ \bar{x} + tv \in K \ \ \forall\, t > 0 \Big\} = \bigcap_{t>0} t(K - \bar{x}).$$

Here K^∞ is independent of $\bar{x}$. For general closed sets, we have

$$\bigcap_{t>0} t(K - \bar{x}) \subset K^\infty.$$

Hence $R^i(K;\bar{x}) \subset \operatorname{int}(K^\infty)$ for all $\bar{x} \in K$. The following proposition summarizes the previous results.

PROPOSITION 2.3. *Let $C \subset X$ be a closed set, $\bar{x} \in C$. If C is radiant at $\bar{x}$, then*

$$R^i(C;\bar{x}) = \operatorname{int}\left(\bigcap_{t>0} t(C - \bar{x}) \right) = \operatorname{int}(C^\infty). \tag{2.3}$$

When C is convex, $R^i(C;\bar{x})$ is independent of $\bar{x}$.

By $\overline{\text{cone}}\ A$ we denote the smallest closed cone containing A, which is the closure of the smallest cone containing A. More precisely,

$$\overline{\text{cone}}\ A = \overline{\text{cone}\ (A)} \ \text{ and } \ \text{cone}\ A = \bigcup_{t \geq 0} tA.$$

The next proposition justifies the term "closed radial" for the set $R(C;\bar{x})$.

PROPOSITION 2.4. *Given any nonempty set C and $\bar{x} \in C$, we have*

(a) $R(C;\bar{x}) = \overline{\text{cone}}(C - \bar{x})$;

(b) $R(C;\bar{x}) = T(C;\bar{x})$, *provided that C is star-shaped with respect to $\bar{x}$.*

Proof. Part (a) follows directly from the very definition of $R(C;\bar{x})$ and by noticing that $\overline{\text{cone}(C - \bar{x})} = \overline{\text{cone}}(C - \bar{x})$. Part (b) is Corollary 4.11 in [J]. □

In addition to the normed space X, let Y be another real normed vector space. We shall require that Y be an ordered space, with ordering cone P being closed convex and pointed ($P \cap (-P) = \{0\}$); eventually we will require $\operatorname{int} P \neq \emptyset$. The cone P will determine the "preference relation." We recall that P introduces on Y a partial ordering by defining $y_1 \geq y_2$ (equivalently, $y_2 \leq y_1$) if and only if $y_1 - y_2 \in P$. This ordering is reflexive, transitive, and antisymmetric; i.e., ($y \geq 0$ and $y \leq 0$) implies $y = 0$.

For $B \subset Y$ nonempty, $y_0 \in Y$ is a lower bound of B if and only if $y_0 \leq b$ for all $b \in B$. An *infimum* of B is a greatest lower bound, i.e., a lower bound y_0 of B such that $y_0 \geq y_1$ for every other lower bound y_1 of B. From antisymmetry of the ordering, the infimum is unique; we denote it by $\inf B$, provided that it exists ($\inf B \in Y$). An element $y_0 \in Y$ is a *minimum* of B, denoted $y_0 = \min B$, if and only if $y_0 = \inf B$ and

$y_0 \in B$. This is equivalent to $y_0 \in B$ and $y_0 \leq b$ for all $b \in B$. Analogous definitions hold for the upper bound and $\sup B$, $\max B$.

The crucial order theoretic assumption we have to make is that (Y, P) is *order-complete*. This means that every nonempty subset of Y, which has an upper bound, also has a supremum. If $\operatorname{int} P \neq \emptyset$, any two-element subset $\{a, b\} \subset Y$ has an upper bound. In fact, choose $t \in \operatorname{int} P$ and $\lambda \in]0, \infty[$ so small that $t + \lambda(a - b) \in P$. Then $a + t/\lambda$ is an upper bound for both a and b. Hence $\sup\{a, b\}$ exists for any $a, b \in Y$, due to order-completeness. Similarly one can show that $\inf\{a, b\}$ also exists. Finally, we adjoin to Y two artificial elements, $-\infty$ and $+\infty$, say, and denote the extended space by $\overline{Y} = Y \cup \{\pm\infty\}$. We suppose that $-\infty \leq y \leq +\infty$ for every $y \in Y \cup \{\pm\infty\}$. Furthermore, the following conventions are assumed:

$$(\pm\infty) + y = y + (\pm\infty) = \pm\infty \quad \forall\, y \in Y, \qquad (\pm\infty) + (\pm\infty) = \pm\infty,$$

$$\lambda(\pm\infty) = \pm\infty \ \ \forall\, \lambda > 0, \quad \text{and} \quad \lambda(\pm\infty) = \mp\infty \ \ \forall\, \lambda < 0.$$

Thus, every nonempty subset of Y has a infimum in $Y \cup \{-\infty\}$ (resp., a supremum in $Y \cup \{+\infty\}$), and the infima (resp., suprema) occurring in the next section are always to be understood in this sense.

We shall also need the notion of the epigraph of a single-valued map $f : X \to \overline{Y}$. It is, as usual, the set

$$\operatorname{epi} f = \Big\{(x, y) \in X \times Y : \ f(x) \in Y, y \in f(x) + P\Big\} \cup \Big\{(x, y) \in X \times Y : \ f(x) = -\infty\Big\},$$

and the hypograph of f is the set

$$\operatorname{hyp} f = \Big\{(x, y) \in X \times Y : \ f(x) \in Y, y \in f(x) - P\Big\} \cup \Big\{(x, y) \in X \times Y : \ f(x) = +\infty\Big\},$$

while the *effective domain* of f is, as usual,

$$\operatorname{dom} f = \Big\{x \in X : \ f(x) \neq +\infty\Big\}.$$

Finally, we set $\inf A = -\infty$ (resp., $\sup A = +\infty$) if A has no lower (resp., upper) bound, and $\inf \emptyset = +\infty$ (resp., $\sup \emptyset = -\infty$).

DEFINITION 2.5. *The vector-valued function $f : X \to \overline{Y}$ is said to be*

(i) *P-convex if for all $x, y \in \operatorname{dom} f$,*

$$\alpha f(x) + (1 - \alpha) f(y) \in f(\alpha x + (1 - \alpha) y) + P \quad \forall\, \alpha \in \ ]0, 1[.$$

(ii) *(see* [PT]*) P-lower semicontinuous (P-lsc) at $x_0 \in X$ if $f(x_0) = -\infty$ or if for any open set $V \subset Y$ such that $f(x_0) \in V \cup \{+\infty\}$ there exists an open neighborhood $U \subset X$ of x_0 such that $f(U) \subset V + (P \cup \{+\infty\})$. We shall say that f is P-lsc if it is P-lsc at every point $x_0 \in X$.*

(iii) *P-upper semicontinuous (P-usc) at $x_0 \in X$ if $-f$ is P-lsc at x_0. We shall say that f is P-usc if it is P-usc at every point $x_0 \in X$.*

The following proposition will be used in what follows and in the next sections.

PROPOSITION 2.6. *Let $W \subset Y$ be any nonempty set, $W \neq Y$, and P be a convex cone. Then,*

(a) *$\lambda W \subset W$ for all $\lambda > 0 \Longrightarrow \lambda(Y \setminus W) \subset \mathbb{R}^m \setminus W$ for all $\lambda > 0$ and $0 \in \overline{Y \setminus W}$. Thus also $0 \in \overline{W} \cap (-\overline{W}) \cap \overline{Y \setminus -W}$.*

(b) $W + P \subseteq W \iff (Y \setminus -W) + P \subset Y \setminus -W \iff (Y \setminus W) - P \subset Y \setminus W$.

Proof. Part (a) is obvious since for any fixed $a \in Y \setminus W$, $ta \in Y \setminus W$ for all $t > 0$. Then let $t \to 0$. Part (b): let $u \notin -W$, $p \in P$ and suppose that $u + p \in -W$. Then $u \in -W - P \subset -W$ by assumption on W. Hence $(Y \setminus -W) + P \subseteq Y \setminus -W$. The reverse implication follows by symmetry. The other equivalence follows directly since, for any sets A and B, $A \subset B$ if and only if $-A \subset -B$. ∎

Regarding the notions in Definition 2.5 we have the following lemma. Part (c) can be found in [BHS], and part (d) in [L1] for nonextended vector-valued functions. We present the proof for sake of completeness in the general case.

LEMMA 2.7. *Let P be a closed convex cone; let $W \subset Y$ be a closed set such that $W + P \subset W$, $W \neq Y$; and let $f : X \to \overline{Y}$ be any function. Then we have the following:*

(a) epi f *is convex if and only if f is P-convex.*

(b) *If f is a P-lsc function, then the set $A = \{x \in X : f(x) - \lambda \in -(W \cup \{+\infty\})\}$ is closed for all $\lambda \in Y$.*

(c) *Assume* int $P \neq \emptyset$; *f is P-lsc if and only if $\{x \in X : f(x) - \lambda \notin (\text{int } P) \cup \{+\infty\}\}$ is closed for all $\lambda \in Y$.*

(d) *Assume* int $P \neq \emptyset$; epi f *is closed if and only if $\{x \in X : f(x) - \lambda \in -(P \cup \{+\infty\})\}$ is closed for all $\lambda \in Y$.*

(e) *If f is P-convex, then the set $\{x' \in X : f(x') \in f(x) - P\}$ is convex for all $x \in X$, $f(x) \in Y$.*

Proof. The first assertion follows by definition. Let (x_n), $n \in \mathbb{N}$, be any sequence in A such that $x_n \to x$. We will prove that $x \in A$. If, on the contrary, $x \notin A$, we could have $f(x) = +\infty$ or $f(x) \notin \lambda - W$. Thus $f(x) \in (\lambda + Y \setminus (-W)) \cup \{+\infty\}$. Thus by the P-lower semicontinuity of f at x, there is an open neighborhood U of x satisfying $f(U) \subset \lambda + Y \setminus (-W) + (P \cup \{+\infty\})$. Since $x_n \in U$ for n sufficiently large, the previous inclusion implies that (see Proposition 2.6) $f(x_n) \in \lambda + (Y \setminus (-W)) \cup \{+\infty\}$ for n large enough, which contradicts the choice of $x_n \in A$, proving the second assertion. The "only if" part of (c) follows from (b) by taking $W = P \setminus (-\text{int } P)$. Let us prove the "if" part. If $f(\bar{x}) = -\infty$, there is nothing to prove. Take any open set V such that $f(\bar{x}) \in V \cup \{+\infty\}$. In case $f(\bar{x}) \in V$, since V is open, we can choose $y_0 \in \text{int } P$ such that $f(\bar{x}) - y_0 \in V$. Set $\lambda = f(\bar{x}) - y_0$. Thus $f(\bar{x}) - \lambda \in (\text{int } P) \cup \{+\infty\}$. By assumption, there exists an open set U, $\bar{x} \in U$, such that $f(x) \in (\lambda + \text{int } P) \cup \{+\infty\} \subset V + (P \cup \{+\infty\})$ for all $x \in U$, which means that f is P-lsc at $\bar{x}$. Part (d) is similar to that in [L1], and part (e) is straightforward. ∎

When the function f does not take the value $-\infty$, more precise formulations can be stated. We single out these results in the next proposition.

PROPOSITION 2.8. *Let $f : X \to Y \cup \{+\infty\}$, and let W be as in the previous lemma. Then,*

(a) *if f is a P-lsc function, then $\{x \in X : f(x) \in \lambda - W\}$ is closed for all $\lambda \in Y$;*

(b) *assuming* int $P \neq \emptyset$, *f is P-lsc if and only if $\{x \in X : f(x) - \lambda \notin \text{int } P\}$ is closed for all $\lambda \in Y$;*

(c) *assuming* int $P \neq \emptyset$, epi f *is closed if and only if $\{x \in X : f(x) - \lambda \in -P\}$ is closed for all $\lambda \in Y$;*

(d) *if f is P-lsc, then* epi f *is closed.*

3. Radial epiderivatives and main results. Throughout this paper (Y, P) will denote an ordered real normed vector space, $Y \neq \{0\}$, with P satisfying the following hypothesis.

Hypothesis (H0). $P \subset Y$ is a closed convex pointed cone such that (Y, P) is

order-complete. Notice that if P is pointed and $Y \neq \{0\}$, then $P \neq Y$.

Sometimes we shall require int $P \neq \emptyset$.

Let us consider any vector-valued function $f : X \to Y \cup \{+\infty\}$. For a given $\bar{x} \in \operatorname{dom} f$, the lower radial epiderivative of f at $\bar{x}$, $\underline{D}_e^R f(\bar{x}; \cdot) : X \to \overline{Y}$, is defined by

$$\underline{D}_e^R f(\bar{x}; u) = \liminf_{u' \to u} \inf_{t>0} \frac{f(\bar{x} + tu') - f(\bar{x})}{t} \doteq \sup_{\varepsilon > 0} \inf_{||u'-u|| < \varepsilon} \inf_{t>0} \frac{f(\bar{x} + tu') - f(\bar{x})}{t}.$$

The upper radial epiderivative of f at $\bar{x}$, $\overline{D}_e^R f(\bar{x}; \cdot) : X \to \overline{Y}$, is defined by

$$\overline{D}_e^R f(\bar{x}; u) = \limsup_{u' \to u} \sup_{t>0} \frac{f(\bar{x} + tu') - f(\bar{x})}{t} \doteq \inf_{\varepsilon > 0} \sup_{||u'-u|| < \varepsilon} \sup_{t>0} \frac{f(\bar{x} + tu') - f(\bar{x})}{t}.$$

As a simple example take $f(x) = Ax + b$, $x \in \mathbb{R}^n$, with $P = \mathbb{R}^m_+$, where A is an $m \times n$-matrix with real entries, and $b \in \mathbb{R}^m$. Then one obtains $\underline{D}_e^R f(\bar{x}; u) = \overline{D}_e^R f(\bar{x}; u) = Au$ for all $u \in \mathbb{R}^n$ and all $\bar{x} \in \mathbb{R}^n$. For other less trivial examples, we refer to Remark 3.6 (iii) and Example 3.19.

The following sets play important roles throughout the paper. Such sets are well defined since $(\bar{x}, \bar{y})$, $\bar{y} = f(\bar{x})$, is a boundary point of epi f, hyp f and of their complements as well, provided $P \neq \{0\}$.

$$\begin{aligned}
H_+(u) &= \Big\{ v \in Y : \ (u, v) \in R(\text{epi } f; (\bar{x}, \bar{y})) \Big\}, \\
H_-(u) &= \Big\{ v \in Y : \ (u, v) \in R(\text{hyp } f; (\bar{x}, \bar{y})) \Big\}, \\
H_i(u) &= \Big\{ v \in Y : \ (u, v) \in R^i(\text{hyp } f; (\bar{x}, \bar{y})) \Big\}, \\
H^i(u) &= \Big\{ v \in Y : \ (u, v) \in R^i(\text{epi } f; (\bar{x}, \bar{y})) \Big\}.
\end{aligned}$$

PROPOSITION 3.1. *Assume that P satisfies Hypothesis* (H0). *Let $f : X \to Y \cup \{+\infty\}$, $\bar{x} \in \operatorname{dom} f$, $\bar{y} = f(\bar{x})$. Then we have the following:*

(a)
- $\underline{D}_e^R f(\bar{x}; u) \leq \inf\{v \in Y : \ (u, v) \in R(\text{epi } f; (\bar{x}, \bar{y}))\}$. *Thus,* $R(\text{epi } f; (\bar{x}, \bar{y})) \subset \text{epi}(\underline{D}_e^R f(\bar{x}; \cdot))$.
- $\underline{D}_e^R f(\bar{x}; u) \geq \sup\{v \in Y : \ (u, v) \in R^i(\text{hyp } f; (\bar{x}, \bar{y}))\}$. *Thus,* $R^i(\text{hyp } f; (\bar{x}, \bar{y})) \subset \text{hyp}(\underline{D}_e^R(\bar{x}; \cdot))$.

(b)
- $\overline{D}_e^R f(\bar{x}; u) \geq \sup\{v \in Y : \ (u, v) \in R(\text{hyp } f; (\bar{x}, \bar{y}))\}$. *Thus,* $R(\text{hyp } f; (\bar{x}, \bar{y})) \subset \text{hyp}(\overline{D}_e^R f(\bar{x}; \cdot))$.
- $\overline{D}_e^R f(\bar{x}; u) \leq \inf\{v \in Y : \ (u, v) \in R^i(\text{epi } f; (\bar{x}, \bar{y}))\}$. *Thus,* $R^i(\text{epi } f; (\bar{x}, \bar{y})) \subset \text{epi}(\overline{D}_e^R(\bar{x}; \cdot))$.

Proof. We prove only part (a), the other part being entirely similar. Obviously the first inequality is trivially satisfied if $H_+(u)$ is empty. Otherwise, take any $v \in H_+(u)$. Then, there exist sequences $t_n > 0$, $u_n \to u$, and $v_n \to v$ such that

$$v_n \geq \frac{f(\bar{x} + t_n u_n) - f(\bar{x})}{t_n} \geq \inf_{t>0} \frac{f(\bar{x} + tu_n) - f(\bar{x})}{t} \quad \forall \ n \in \mathbb{N}.$$

Given $\varepsilon > 0$, there exists $n_0 \in \mathbb{N}$ such that $||u_n - u|| < \varepsilon$ for all $n \geq n_0$. Therefore, for all $n \geq n_0$

$$v_n \geq \inf_{t>0} \frac{f(\bar{x} + tu_n) - f(\bar{x})}{t} \geq \inf_{||u'-u|| < \varepsilon} \inf_{t>0} \frac{f(\bar{x} + tu') - f(\bar{x})}{t}.$$

Letting $n \to +\infty$, we obtain

$$v \geq \inf_{||u'-u||<\varepsilon} \inf_{t>0} \frac{f(\bar{x}+tu')-f(\bar{x})}{t}.$$

Since the latter holds for arbitrary $\varepsilon > 0$, we conclude $v \geq \underline{D}_e^R f(\bar{x};u)$.

For the second part of (a), if $H_i(u)$ is empty, the inequality is trivially satisfied. Take any $v \in H_i(u)$; then, by definition, there exists $\bar{\varepsilon} > 0$ such that

$$v' \leq \frac{f(\bar{x}+tu')-f(\bar{x})}{t} \quad \forall\, t>0, \qquad ||u'-u|| < \bar{\varepsilon}, \quad ||v'-v|| < \bar{\varepsilon}.$$

In particular, for $v' = v$, the previous inequality reduces to

$$v \leq \frac{f(\bar{x}+tu')-f(\bar{x})}{t} \quad \forall\, t>0, \qquad ||u'-u|| < \bar{\varepsilon}.$$

Thus

$$\begin{aligned} v &\leq \inf_{||u'-u||<\bar{\varepsilon}} \inf_{t>0} \frac{f(\bar{x}+tu')-f(\bar{x})}{t} \\ &\leq \sup_{\varepsilon>0} \inf_{||u'-u||<\varepsilon} \inf_{t>0} \frac{f(\bar{x}+tu')-f(\bar{x})}{t} = \underline{D}_e^R f(\bar{x};u). \qquad \square \end{aligned}$$

From the preceding proof one can see that, if

$$D_e^R f(\bar{x};u) \doteq \inf\left\{v \in Y :\ (u,v) \in R(\text{epi } f;(\bar{x},\bar{y}))\right\} \in Y \quad \forall\, u \in X,$$

then

$$\text{epi}(D_e^R f(\bar{x};\cdot)) = R(\text{epi } f;(\bar{x},\bar{y})).$$

Thus, $D_e^R f(\bar{x};\cdot)$ could be a good candidate for a notion of epiderivative (it might be with respect to a different cone like the contingent, or Clarke's), but as we said in the introduction, to give conditions guaranteeing the existence of such an epiderivative is not an easy task. Hence, we consider the lower and upper radial epiderivatives as defined above, which may take the values $\pm\infty$.

If, in the definition of $D_e^R f(\bar{x};\cdot)$ above, we consider the Clarke tangent cone instead of the closed radial one, we get the directional subderivative of f at $\bar{x}$, $f^{\uparrow}(\bar{x};\cdot)$, considered and developed in [Th3].

THEOREM 3.2. *Let P satisfy Hypothesis* (H0). *Let $f : X \to Y \cup \{+\infty\}$, $\bar{x} \in \text{dom } f$. Then*

(a) *$f(\bar{x}+tu)-f(\bar{x}) \in t\underline{D}_e^R f(\bar{x};u)+(P\cup\{+\infty\})$ for all $t>0$ and all $u \in X$ such that $\underline{D}_e^R f(\bar{x};u) \neq -\infty$. Hence, $\underline{D}_e^R f(\bar{x};u) = +\infty$ implies $f(\bar{x}+tu) = +\infty$ for all $t>0$. Consequently*

$$f(x)-f(\bar{x}) \in \underline{D}_e^R f(\bar{x};x-\bar{x})+(P\cup\{+\infty\}) \quad \forall\, x \in X,\ \underline{D}_e^R f(\bar{x};x-\bar{x}) \neq -\infty.$$

(b) *$\overline{D}_e^R f(\bar{x};u) \neq -\infty$ for all $u \in X$. Moreover, if $\overline{D}_e^R f(\bar{x};u) \neq +\infty$, then $f(\bar{x}+tu) \in Y$ for all $t>0$ and $\overline{D}_e^R f(\bar{x};u) \in Y$. Hence, for such u, $f(\bar{x}+tu)-f(\bar{x}) \in t\overline{D}_e^R f(\bar{x};u)-P$ for all $t>0$. Consequently*

$$f(x)-f(\bar{x}) \in \overline{D}_e^R f(\bar{x};x-\bar{x})-P \quad \forall\, x \in X,\ \overline{D}_e^R f(\bar{x};x-\bar{x}) \neq +\infty.$$

Proof. First, we recall that $(\bar{x},\bar{y})$ is a boundary point of hyp f (if $P\neq\{0\}$) and thus also of its complement.

(a) Let us fix $u\in X$ such that $\underline{D}_e^R f(\bar{x};u)\neq-\infty$. Thus

$$A=\left\{\inf_{||u'-u||<\varepsilon}\inf_{t>0}\frac{f(\bar{x}+tu')-f(\bar{x})}{t}\in Y\cup\{+\infty\}:\varepsilon>0\right\}\neq\emptyset.$$

For such $\varepsilon>0$ we have for all $t>0$

$$\inf_{||u'-u||<\varepsilon}\inf_{t'>0}\frac{f(\bar{x}+t'u')-f(\bar{x})}{t'}\leq\frac{f(\bar{x}+tu)-f(\bar{x})}{t}.$$

Hence

$$f(\bar{x}+tu)-f(\bar{x})\in t\underline{D}_e^R f(\bar{x};u)+P\cup\{+\infty\}\quad\forall\ t>0.$$

(b) If, on the contrary $\overline{D}_e^R f(\bar{x};u)=-\infty$, we have, by Proposition 3.1,

$$(u,v)\notin R(\text{hyp } f;(\bar{x},\bar{y}))\ \ \forall\ v\in Y.$$

By Remark 2.2, $(u,v)\in R^i(X\times Y\setminus\text{hyp } f;(\bar{x},\bar{y}))$ for all $v\in Y$. In particular, $(\bar{x}+tu,\bar{y}+tv)\notin\text{hyp } f$ for all $v\in Y$ and all $t>0$. In case $f(\bar{x}+tu)=+\infty$, then obviously $(\bar{x}+tu,\bar{y}+tv)\in\text{hyp } f$, a contradiction; if $f(\bar{x}+tu)\in Y$, then $f(\bar{x}+tu)-f(\bar{x})\in tv+(Y\setminus P)$ for all $v\in Y$, which cannot happen if we take $v=(f(\bar{x}+tu)-f(\bar{x}))/t$. This completes the proof of the first part.

Assume that $\overline{D}_e^R f(\bar{x};u)\neq+\infty$. Therefore, since $f(x)\neq-\infty$ for all $x\in X$,

$$B=\left\{\sup_{||u'-u||<\varepsilon}\sup_{t>0}\frac{f(\bar{x}+tu')-f(\bar{x})}{t}\in Y:\varepsilon>0\right\}\neq\emptyset.$$

For such $\varepsilon>0$ we have for all $t>0$

$$\sup_{||u'-u||<\varepsilon}\sup_{t'>0}\frac{f(\bar{x}+t'u')-f(\bar{x})}{t'}\geq\frac{f(\bar{x}+tu)-f(\bar{x})}{t}.$$

Hence $f(\bar{x}+tu)\in Y$ for all $t>0$, and thus $\overline{D}_e^R f(\bar{x};u)\neq-\infty$. In addition,

$$f(\bar{x}+tu)-f(\bar{x})\in t\overline{D}_e^R f(\bar{x};u)-P\quad\forall\ t>0.\qquad\square$$

THEOREM 3.3. *In addition to Hypothesis* (H0)*, assume that* int $P\neq\emptyset$. *Let* $f:X\to Y\cup\{+\infty\}$, $\bar{x}\in\text{dom } f$. *Then for every* $u\in X$ *we have*

$$\underline{D}_e^R f(\bar{x};u)=\sup\left\{v\in Y:\ (u,v)\in R^i(\text{hyp } f;(\bar{x},\bar{y}))\right\},\tag{3.1}$$

$$\overline{D}_e^R f(\bar{x};u)=\inf\left\{v\in Y:\ (u,v)\in R^i(\text{epi } f;(\bar{x},\bar{y}))\right\}.\tag{3.2}$$

Proof. As before, we shall prove only the first equality. From Proposition 3.1 such an equality is true if $\sup H_i(u)=+\infty$. We shall prove now that $\underline{D}_e^R f(\bar{x};u)=+\infty$ implies $\sup H_i(u)=+\infty$. If $\underline{D}_e^R f(\bar{x};u)=\sup A=+\infty$, then A has no upper bound, where

$$A=\left\{a(\varepsilon)\doteq\inf_{||u'-u||<\varepsilon}\inf_{t>0}\frac{f(\bar{x}+tu')-f(\bar{x})}{t}:\varepsilon>0\right\}.$$

In the case in which there exists $\bar{\varepsilon} > 0$ such that $a(\bar{\varepsilon}) = +\infty$, we obtain $f(\bar{x} + tu') = +\infty$ for all $t > 0$ and all u' such that $||u' - u|| < \bar{\varepsilon}$. This implies that $(u, v) \in R^i(\text{hyp } f; (\bar{x}, \bar{y}))$ for all $v \in Y$, that is, $v \in H_i(u)$ for all $v \in Y$. Thus $\sup H_i(u) = +\infty$. Therefore, we may assume that $A \subset Y$. Suppose that $H_i(u)$ has upper bounds. Take any of these, say $v \in Y$. Since A has no upper bound, there exists $v_0 \in A$ such that $v - v_0 \notin P$; this v_0 is of the form

$$v_0 = \inf_{||u'-u||<\varepsilon} \inf_{t>0} \frac{f(\bar{x} + tu') - f(\bar{x})}{t} \quad \text{for some} \quad \varepsilon > 0.$$

The closedness of P implies the existence of $p \in \text{int } P$ such that $-v + v_0 - p \notin -P$. For $||u' - u|| < \varepsilon$, $t > 0$, we have

$$\frac{f(\bar{x} + tu') - f(\bar{x})}{t} \geq v_0 \geq v_0 - \xi \quad \forall\, \xi \in P. \tag{3.3}$$

Since $p \in \text{int } P$, it follows that (possibly for a smaller ε)

$$(\bar{x} + tu', \bar{y} + t(v_0 - \xi)) = (\bar{x}, \bar{y}) + t(u', v') \in \text{hyp } f$$

$$\forall\, t > 0, \quad ||u' - u|| < \varepsilon, \quad ||v' - (v_0 - p)|| < \varepsilon.$$

Thus $(u, v_0 - p) \in R^i(\text{hyp } f; (\bar{x}, \bar{y}))$; that is, $v_0 - p \in H_i(u)$. As v is chosen as an upper bound of $H_i(u)$, we obtain $v \geq v_0 - p$, i.e., $v - v_0 + p \in P$, which contradicts the choice of p, proving that $H_i(u)$ has no upper bound and therefore $\sup H_i(u) = +\infty$. This ends the proof of (3.1) in case any of the sides is equal to $+\infty$. Otherwise, take any upper bound $v \in Y$ of $H_i(u) (\neq \emptyset)$; we shall prove that $v \geq \underline{D}_e^R f(\bar{x}; u)$. From Proposition 3.1, $\underline{D}_e^R f(\bar{x}; u) \neq -\infty$, and because of the previous reasoning, $\underline{D}_e^R f(\bar{x}; u) \neq +\infty$; thus $\underline{D}_e^R f(\bar{x}; u) \in Y$. We assume that $\underline{D}_e^R f(\bar{x}; u) \not\leq v$. Then, there is $v_0 \in A$, $v_0 \in Y$, such that $v - v_0 \notin P$. Thus we can proceed as in the previous case to get a contradiction, proving that $v \geq \underline{D}_e^R f(\bar{x}; u)$. It follows that $\sup\{v \in Y : v \in H_i(u)\} \geq \underline{D}_e^R f(\bar{x}; u)$, which together with Proposition 3.1 gives the equality in (3.1). It remains only to check the equality when any of the sides takes the value $-\infty$. Clearly if $\underline{D}_e^R f(\bar{x}; u) = -\infty$, then $\sup H_i(u) = -\infty$ because of Proposition 3.1 again. If $H_i(u) = \emptyset$, then $\underline{D}_e^R f(\bar{x}; u) \in Y \cup \{-\infty\}$. The case $\underline{D}_e^R f(\bar{x}; u) = +\infty$ cannot happen because of the previous reasoning. In case $\underline{D}_e^R f(\bar{x}; u) \in Y$, the set

$$\tilde{A} = \left\{ \varepsilon > 0 : \inf_{||u'-u||<\varepsilon} \inf_{t>0} \frac{f(\bar{x} + tu') - f(\bar{x})}{t} \in Y \right\} \neq \emptyset.$$

Take any $\varepsilon \in \tilde{A}$ and set

$$v_0 \doteq \inf_{||u'-u||<\varepsilon} \inf_{t>0} \frac{f(\bar{x} + tu') - f(\bar{x})}{t} \in Y.$$

Then, we proceed as before (see (3.3)) to conclude that $H_i(u) \neq \emptyset$, a contradiction. This completes the proof of the theorem. □

COROLLARY 3.4. *Assume* $Y = \mathbb{R}$, $P = [0, +\infty[$. *Then* $\text{epi}(\underline{D}_e^R f(\bar{x}; \cdot)) =$

$R(\text{epi } f;(\bar{x},\bar{y}))$, $\text{hyp}(\overline{D}_e^R f(\bar{x};\cdot)) = R(\text{hyp } f;(\bar{x},\bar{y}))$, *and for every* $u \in X$,

$$\begin{aligned}\underline{D}_e^R f(\bar{x};u) &= \inf\left\{v \in \mathbb{R} : (u,v) \in R(\text{epi } f;(\bar{x},\bar{y}))\right\}\\ &= \sup\left\{v \in \mathbb{R} : (u,v) \in R^i(\text{hyp } f;(\bar{x},\bar{y}))\right\},\\ \overline{D}_e^R f(\bar{x};u) &= \sup\left\{v \in \mathbb{R} : (u,v) \in R(\text{hyp } f;(\bar{x},\bar{y}))\right\}\\ &= \inf\left\{v \in \mathbb{R} : (u,v) \in R^i(\text{epi } f;(\bar{x},\bar{y}))\right\}.\end{aligned}$$

Proof. By part (a) of the previous proposition, it remains only to prove that $\text{epi}(\underline{D}_e^R f(\bar{x};\cdot)) \subset R(\text{epi } f(\bar{x},\bar{y}))$. Let $(u,v) \in \text{epi}(\underline{D}_e^R f(\bar{x};\cdot))$; then $\underline{D}_e^R f(\bar{x};u) = -\infty$, $v \in \mathbb{R}$ or $\underline{D}_e^R f(\bar{x};u) \in \mathbb{R}$ and $v = \underline{D}_e^R f(\bar{x};u) + p$, $p \geq 0$. We consider only the second case, the other being similar. By definition, there is $u_n \to u$ such that

$$\lim_{n\to+\infty} \inf_{t>0} \frac{f(\bar{x}+tu_n) - f(\bar{x})}{t} = \underline{D}_e^R f(\bar{x};u).$$

Thus (up to a subsequence), there is also $t_n > 0$ satisfying

$$\lim_{n\to+\infty} \frac{f(\bar{x}+t_nu_n) - f(\bar{x})}{t_n} = \underline{D}_e^R f(\bar{x};u).$$

By setting

$$v_n = \frac{f(\bar{x}+t_nu_n) - f(\bar{x})}{t_n},$$

we have $v_n \to \underline{D}_e^R f(\bar{x};u)$, $f(\bar{x}) + t_nv_n = f(\bar{x}+t_nu_n)$. Then

$$f(\bar{x}) + t_n(v_n+p) = f(\bar{x}+t_nu_n) + t_np \in f(\bar{x}+t_nu_n) + P.$$

Hence $(u,v) \in R(\text{epi } f;(\bar{x},\bar{y}))$. □

The next corollary is a direct consequence of Proposition 3.1 and Theorem 3.2. Notice that the assumption $\text{int } P \neq \emptyset$ is not needed. The case $Y = \mathbb{R}$ is discussed in Remark 3.6.

COROLLARY 3.5. *Let* P *satisfy Hypothesis* (H0)*; let* $W \subset Y$ *be any nonempty set such that* $W + P \subset W$ *and* $\lambda W \subset W$ *for all* $\lambda > 0$ (*for example,* $W = P$, $W = Y \setminus (-P \setminus \{0\})$, $W = Y \setminus (-\text{int } P)$)*; and let* $f : X \to Y \cup \{+\infty\}$, $\bar{x} \in \text{dom } f$. *Then*

(a) $f(x) - f(\bar{x}) \in W \cup \{+\infty\}$ *for all* $x \in X \Longrightarrow \overline{D}_e^R f(\bar{x};x-\bar{x}) \in W \cup \{+\infty\}$ *for all* $x \in X$;

(b) $\underline{D}_e^R f(\bar{x};x-\bar{x}) \in W \cup \{+\infty\}$ *for all* $x \in X \Longrightarrow f(x) - f(\bar{x}) \in W \cup \{+\infty\}$ *for all* $x \in X$;

(c) $f(x) - f(\bar{x}) \in P \cup \{+\infty\}$ *for all* $x \in X \Longleftrightarrow \underline{D}_e^R f(\bar{x};x-\bar{x}) \in P \cup \{+\infty\}$ *for all* $x \in X$.

Proof. (a) If, on the contrary, $\overline{D}_e^R f(\bar{x};x-\bar{x}) \in (Y \setminus W) \cup \{-\infty\}$, we easily get a contradiction by virtue of Theorem 3.2, and Proposition 2.6 in case $\overline{D}_e^R f(\bar{x};x-\bar{x}) \in Y \setminus W$. The case $\overline{D}_e^R f(\bar{x};x-\bar{x}) = -\infty$ never happens. Part (b) follows from Theorem 3.2, and part (c) is easily obtained. □

Remark 3.6. (i) We provide an example showing that the reverse implication of part (a) of the previous corollary does not hold. Take $f(x) = \sqrt{|x|}$, $x \in \mathbb{R}$, $\bar{x} = 1$. Then we have $\overline{D}_e^R f(1;x-1) \geq 0$ for all $x \in \mathbb{R}$, and $\bar{x} = 1$ is not a minimum for f.

(ii) Those $\bar{x}$ satisfying (c) are called ideal or strong minimizers for f. This equivalence, which expresses an optimality condition for $\bar{x}$ to be a strong minimizer for f, seems to be new in this general framework. Such a condition could be termed the Fermat rule for the problem

$$f(x) - f(\bar{x}) \in P \cup \{+\infty\} \quad \forall\, x \in X. \tag{3.4}$$

Notice that no convexity assumption is required in order that such a condition be also sufficient, as occurs in [AE, Proposition 4, Chapter 4, section 1] or in [AF, section 6.1.3] when $Y = \mathbb{R}$. A similar equivalence when f is a set-valued map is established in [F3] under additional assumptions, which reduces to (c) in the real case and when f is single-valued. In such a case the equivalence can be written as

$$f(x) - f(\bar{x}) \geq 0 \ \ \forall\, x \in X \iff \underline{D}_e^R f(\bar{x}; x - \bar{x}) \geq 0 \ \ \forall\, x \in X. \tag{3.5}$$

Indeed, the notion of a radial epiderivative for set-valued maps is introduced in [F3]. This corresponds to our lower radial epiderivative $\underline{D}_e^R f(\bar{x}; \cdot)$ in the single-valued case. In that paper, (3.5) is proved, among other results (see Theorem 3.9 in [F3] particularized to $Y = \mathbb{R}$), by using the equality $\operatorname{epi}(\underline{D}_e^R f(\bar{x}; \cdot)) = R(\operatorname{epi} f; (\bar{x}, \bar{y}))$, which is true by Corollary 3.4.

To the best of our knowledge, condition (3.5) appears for the first time in [F3], although similar necessary optimality conditions (with different cones) may be found in [P2]; see also [HU1, HU2]. In order to get an idea of the applicability of (3.5), simply take $f(x) = \sqrt{|x|}$, $x \in \mathbb{R}$, considered in [L2]. Clearly the theory of Clarke's derivatives is not applicable in this case; see also Theorem 3.1 in [BZ]. Neither Theorem 6 in [CJ] nor Corollary 2 in [JR] can be used, since the function involved is not convex (see also [Y2]). Moreover, Theorem 2.1 in [L2] asserts only that 0 is a local minimum of f. However, an easy computation shows that $\underline{D}_e^R f(0; x) = 0$ for all $x \in \mathbb{R}$. Thus, (3.5) says that 0 is a global minimum.

Additional sufficient and necessary optimality conditions for the problem (3.4), in a different setting, may be found in [FO].

(iii) Contrary to the real-valued case ($Y = \mathbb{R}$), or when $W = P$, as discussed in (ii), it could happen that the reverse implication in part (b) is not valid in the vector case when $W \neq P$. To see this, let us consider the example in [P2]: $X = \mathbb{R}$, $Y = \mathbb{R}^2$, $P = \mathbb{R}^2_+$, $W = \mathbb{R}^2 \setminus (-\operatorname{int} \mathbb{R}^2_+)$, $f(0) = 0$,

$$f(x) = \begin{cases} (x, -x) & \text{if } \ |x| \in \ ]1/2n, 1/(2n-1)[, \\ (-x, x) & \text{if } \ |x| \in [1/(2n-1), 1/(2n-2)]. \end{cases}$$

It is not hard to check that $\underline{D}_e^R f(0; 1) \in -\operatorname{int} \mathbb{R}^2_+$ and $f(x) - f(0) \in \mathbb{R}^2 \setminus (-\operatorname{int} \mathbb{R}^2_+)$ for all $x \in \mathbb{R}$.

PROPOSITION 3.7. *Assume that P satisfies Hypothesis* (H0) *and* $\operatorname{int} P \neq \emptyset$. *For a vector-valued function $f : X \to Y \cup \{+\infty\}$, $\bar{x} \in \operatorname{dom} f$, we have*

(a) $\underline{D}_e^R f(\bar{x}; \lambda u) = \lambda \underline{D}_e^R f(\bar{x}; u)$ *for all $\lambda > 0$, for all $u \in X$;*

(b) $\underline{D}_e^R f(\bar{x}; 0) = 0$ *if and only if* $\underline{D}_e^R f(\bar{x}; u) \neq -\infty$ *for all $u \in X$;*

(c) $\overline{D}_e^R f(\bar{x}; \lambda u) = \lambda \overline{D}_e^R f(\bar{x}; u)$ *for all $\lambda > 0$, for all $u \in X$;*

(d) $\overline{D}_e^R f(\bar{x}; 0) = 0$ *if and only if* $\overline{D}_e^R f(\bar{x}; u) \in Y$ *for all $u \in X$.*

Proof. (a) follows from Theorem 3.3. (b) One implication is as follows. If there exists $u \in X$ such that $\underline{D}_e^R f(\bar{x}; u) = -\infty$, by part (a), $\underline{D}_e^R f(\bar{x}; \lambda u) = -\infty$ for all $\lambda > 0$. This means that $(\lambda u, v) \notin R^i(\operatorname{hyp} f; (\bar{x}, \bar{y}))$ for all $v \in Y$ for all $\lambda > 0$. Since

the interiorly radial cone is open, $(0,v) \notin R^i(\text{hyp } f;(\bar{x},\bar{y}))$ for all $v \in Y$. Hence $\underline{D}_e^R f(\bar{x};0) = -\infty$, which cannot happen. The other implication is proved by noticing that $(0,0) \in R(\text{epi } f;(\bar{x},\bar{y}))$ and then applying Proposition 3.1 together with the previous part. The other parts are similar. □

Before establishing the next proposition, we recall that $(\bar{x},\bar{y})$, $\bar{y} = f(\bar{x})$, is a boundary point of epi f (if $P \neq \{0\}$) and of hyp f, and therefore of their complements as well.

PROPOSITION 3.8. *Assume that P satisfies hypothesis* (H0). *Given any vector-valued function $f : X \to Y \cup \{+\infty\}$, $\bar{x} \in \text{dom } f$, $u \in X$, we have*

(a)

$$R^i(\text{hyp } f;(\bar{x},\bar{y})) \subset \text{hyp } (\underline{D}_e^R f(\bar{x};\cdot)) \subset R^i(X\times Y\setminus \text{epi } f;(\bar{x},\bar{y})) \cup \bigcap_{t>0} t(\text{hyp } f-(\bar{x},\bar{y})),$$

(b)

$$\text{int}\left(\bigcap_{t>0} t(\text{epi } f-(\bar{x},\bar{y}))\right) = R^i(\text{epi } f;(\bar{x},\bar{y})) \subset \text{epi } (\overline{D}_e^R f(\bar{x};\cdot)) \subset \bigcap_{t>0} t(\text{epi } f-(\bar{x},\bar{y})),$$

(c) *if* int $P \neq \emptyset$, *then* $R(\text{epi } f;(\bar{x},\bar{y})) \subset \text{epi } (\underline{D}_e^R f(\bar{x};\cdot)) \subset R(X\times Y\setminus \text{hyp } f;(\bar{x},\bar{y}))$,

(d) *if* int $P \neq \emptyset$, *then* $R(\text{hyp } f;(\bar{x},\bar{y})) \subset \text{hyp } (\overline{D}_e^R f(\bar{x};\cdot)) \subset R(X\times Y\setminus \text{epi } f;(\bar{x},\bar{y}))$.

Proof. The first inclusion of (a) has been proved in Proposition 3.1. For the second inclusion, we reason as follows. Let $(u,v) \in \text{hyp}(\underline{D}_e^R f(\bar{x};\cdot))$. If $\underline{D}_e^R f(\bar{x};u) = +\infty$, then $(u,v) \notin R(\text{epi } f;(\bar{x},\bar{y}))$ by Proposition 3.1. Hence $(u,v) \in R^i(X \times Y \setminus \text{epi } f;(\bar{x},\bar{y}))$. We consider the case $\underline{D}_e^R f(\bar{x};u) \in Y$. Thus $v \in \underline{D}_e^R f(\bar{x};u) - P$. Therefore, from (a) of Theorem 3.2, we obtain

$$f(\bar{x}+tu) - f(\bar{x}) \in tv + (P \cup \{+\infty\}) \quad \forall\, t > 0.$$

For those $t > 0$ such that $f(\bar{x}+tv) = +\infty$, we clearly have $(\bar{x},\bar{y}) + t(u,v) \in \text{hyp } f$. If, on the contrary, $f(\bar{x}+tv) \in Y$, we also obtain $(\bar{x},\bar{y}) + t(u,v) \in \text{hyp } f$. Consequently,

$$(u,v) \in \bigcap_{t>0} t(\text{hyp } f - (\bar{x},\bar{y})),$$

which completes the proof of part (a). The first inclusion of (b) was already obtained in Proposition 3.1. The other inclusion follows in a similar way as in part (a) by using (b) of Theorem 3.2. Let us prove (c). The first inclusion also follows from Proposition 3.1. Take any $(u,v) \in \text{epi}(\underline{D}_e^R f(\bar{x};\cdot))$. If, on the contrary, $(u,v) \notin R(X \times Y \setminus \text{hyp } f;(\bar{x},\bar{y}))$, then $(u,v) \in R^i(\text{hyp } f;(\bar{x},\bar{y}))$ by Remark 2.2. This immediately leads us to a contradiction if $\underline{D}_e^R f(\bar{x};u) \in Y$, since the interiorly radial cone is open and int $P \neq \emptyset$. In case $\underline{D}_e^R f(\bar{x};u) = -\infty$, Proposition 3.1 implies $(u,v) \notin R^i(\text{hyp } f;(\bar{x},\bar{y}))$, which is also a contradiction. It turns out that $(u,v) \in R(X\times Y\setminus \text{hyp } f;(\bar{x},\bar{y}))$, proving part (c). The proof of part (d) is similar to that of (c). □

The next two propositions establish some relationship between the upper/lower radial epiderivatives and the radial derivative of f at $\bar{x} \in \text{dom } f$, $D^R f(\bar{x};\cdot) : X \to 2^Y$, defined by

$$\text{graph}(D^R f(\bar{x};\cdot)) = R(\text{graph } f;(\bar{x},\bar{y})),$$

with $\bar{y} = f(\bar{x})$, where graph $f = \{(x,y) \in X \times Y :\ y = f(x)\}$; a similar definition for the graph of set-valued maps is given, for instance, in [F3].

PROPOSITION 3.9. *Let P satisfy Hypothesis* (H0)*, and let $f : X \to Y \cup \{+\infty\}$, $\bar{x} \in \operatorname{dom} f$. Given $u \in \operatorname{dom}(\underline{D}_e^R f(\bar{x};\cdot)) \cap \operatorname{dom}(\overline{D}_e^R f(\bar{x};\cdot))$, we have*

(a) $D^R f(\bar{x};u) \subset (\underline{D}_e^R f(\bar{x};u) + P) \cap (\overline{D}_e^R f(\bar{x};u) - P)$;

(b) *if $Y = \mathbb{R}$ and $f : X \to \mathbb{R}$ is continuous, the previous inclusion becomes the equality*

$$D^R f(\bar{x};u) = \left[\underline{D}_e^R f(\bar{x};u), \overline{D}_e^R f(\bar{x};u)\right].$$

Proof. (a) This result follows from the inclusions (see Proposition 3.1)

$$R(\operatorname{graph} f;(\bar{x},\bar{y})) \subset R(\operatorname{epi} f;(\bar{x},\bar{y})) \cap R(\operatorname{hyp} f;(\bar{x},\bar{y})) \subset \operatorname{epi}(\underline{D}_e^R f(\bar{x};\cdot)) \cap \operatorname{hyp}(\overline{D}_e^R f(\bar{x};\cdot)).$$

Part (b) requires some noninvolved computations. □

We point out that the inclusion in the previous proposition may be strict, as the following function shows: take $f(x) = 0$ if $x \in\ ]-\infty, 0] \cup [1, +\infty[$, $f(x) = 1$ if $x \in\ ]0, 1[$. Then $\underline{D}_e^R f(0;u) = 0$ for all $u \in \mathbb{R}$, $\overline{D}_e^R f(0;u) = 0$ if $u < 0$, and $\overline{D}_e^R(0;u) = +\infty$ if $u \geq 0$. However, $D^R f(0;u) = \{0\}$ if $u < 0$, and $D^R f(0;u) = \{0\} \cup [u, +\infty[$ if $u \geq 0$.

THEOREM 3.10. *Let P satisfy Hypothesis* (H0)*; let $f : X \to Y \cup \{+\infty\}$ be any function such that* epi f *is closed, and let $\bar{x} \in \operatorname{dom} f$ with $R^i(\operatorname{epi} f;(\bar{x},\bar{y})) \neq \emptyset$, $\bar{y} = f(\bar{x})$. If* epi f *is radiant at $(\bar{x},\bar{y})$ and* $(\operatorname{epi} f)^\infty$ *is convex, then*

$$\overline{\operatorname{epi}(\overline{D}_e^R f(\bar{x};\cdot))} = (\operatorname{epi} f)^\infty. \tag{3.6}$$

Therefore, if in addition to the previous assumptions $\operatorname{epi}(\overline{D}_e^R f(\bar{x};\cdot))$ *is also closed, then*

$$\operatorname{epi}(\overline{D}_e^R f(\bar{x};\cdot)) = (\operatorname{epi} f)^\infty. \tag{3.7}$$

Consequently, if epi f *is convex besides the closedness of* $\operatorname{epi}(\overline{D}_e^R f(\bar{x};\cdot))$*, then*

$$\overline{D}_e^R f(\bar{x};\cdot) = \overline{D}_e^R f(x;\cdot) \quad \forall\ x \in \operatorname{dom} f.$$

Proof. Since $(\operatorname{epi} f)^\infty$ is convex with nonempty interior by assumption, we obtain

$$\overline{\operatorname{int}((\operatorname{epi} f)^\infty)} = (\operatorname{epi} f)^\infty,$$

since the asymptotic cone of any set is always closed. Thus (3.6) follows from Proposition 3.8(b) and (2.2). Taking into account the additional assumption, (3.7) is a consequence of (3.6). In the case in which epi f is convex as well, the previous equality holds independently of $\bar{x}$. □

Conditions guaranteeing the assumptions of the preceding theorem are exhibited in Lemma 2.7 or Proposition 2.8.

Remark 3.11. One can recognize that (3.7) is the condition to be satisfied for the asymptotic function in the case of $Y = \mathbb{R}$; here $\overline{D}_e^R f(\bar{x};\cdot) = f^\infty$, where f^∞ denotes the asymptotic function of f. This satisfies epi $f^\infty = (\operatorname{epi} f)^\infty$; see, for instance, [BHU] for some formulae of f^∞. Thus, we have found an expression for a candidate recession function for a class of vector-valued functions.

Notice that $(\operatorname{epi} f)^\infty$ may be convex without being epi f. In fact, let us consider $f(x) = \sqrt{|x|}$, $x \in \mathbb{R}$. Then $f^\infty(u) = 0$ for all $u \in \mathbb{R}$ and

$$\bigcap_{t>0} t(\operatorname{epi} f) = \{0\} \times \mathbb{R}_+;$$

thus (2.1) gives $R^i(\text{epi } f;(0,0)) = \emptyset$, and therefore $\overline{D}_e^R f(0;u) = +\infty$ for all $u \in \mathbb{R}$, while $\underline{D}_e^R f(0;u) = 0$ for all $u \in \mathbb{R}$. Notice that 0 is a minimum for f in $\mathbb{R}$. Moreover, $R^i(\text{epi } f;(1,1))$ is a nonempty convex set.

PROPOSITION 3.12. *Assume that P satisfies Hypothesis* (H0). *Let $W \subset Y$ be any nonempty set satisfying $W + P \subset W$ and $\lambda W \subset W$ for all $\lambda > 0$; let $f : X \to Y \cup \{+\infty\}$, $x \in \text{dom } f$. Then*

$$\Big\{u \in X : \overline{D}_e^R f(x;u) \in -W\Big\} \subset \Big\{u \in X :\ f(x+tu) - f(x) \in -W \ \ \forall\, t > 0\Big\}$$
$$\subset \Big\{x' \in X : f(x') \in f(x) - W\Big\}^\infty \cap \Big\{u \in X :\ \underline{D}_e^R f(x;u) \in -(W \cup \{+\infty\})\Big\}.$$

As a consequence, if $\{x' \in X :\ f(x') \in f(x) - P\}$ is convex and closed, then

$$\Big\{u \in X :\ f(x+tu) - f(x) \in -P \ \ \forall\, t > 0\Big\} = \Big\{x' \in X : f(x') \in f(x) - P\Big\}^\infty.$$

Proof. Let u be such that $\overline{D}_e^R f(\bar{x};u) \in -W$; then the first inclusion follows from Theorem 3.2. Set

$$K(x) \doteq \Big\{x' \in X : f(x') \in f(x) - W\Big\}.$$

If $u \in X$ is such that $f(x+tu) - f(x) \in -W$ for all $t > 0$, then $x + tu \in K(x)$ for all $t > 0$. It turns out that $u \in (K(x))^\infty$, proving part of the second inclusion. On the other hand, let $u \in X$ be such that $\underline{D}_e^R f(x;u) \in (Y \setminus (-W)) \cup \{+\infty\}$. Then, by Theorem 3.2 and Proposition 2.6,

$$f(x+tu) - f(x) \in (Y \setminus (-W)) \cup \{+\infty\} \quad \forall\, t > 0.$$

This completes the proof of the second inclusion. In case $K(x)$ is convex and closed for $W = P$, we have $u \in (K(x))^\infty$ if and only if $x + tu \in K(x)$ for all $t > 0$, which proves the reverse inclusion. □

Conditions under which $K(x)$ is closed and convex are given in Lemma 2.7.

For $f : X \to Y \cup \{+\infty\}$, given $\bar{x} \in \text{dom } f$, we set

$$f'_-(\bar{x};u) \doteq \inf_{t>0} \frac{f(\bar{x}+tu) - f(\bar{x})}{t}, \qquad f'_+(\bar{x};u) \doteq \sup_{t>0} \frac{f(\bar{x}+tu) - f(\bar{x})}{t}.$$

We immediately obtain

$$f'_-(\bar{x};0) = f'_+(\bar{x};0) = 0, \qquad f'_+(\bar{x};u) \neq -\infty \ \ \forall\, u \in X, \tag{3.8}$$

$$\underline{D}_e^R f(\bar{x};u) \leq f'_-(\bar{x};u), \qquad \overline{D}_e^R f(\bar{x};u) \geq f'_+(\bar{x};u) \quad \forall\, u \in X. \tag{3.9}$$

When f is P-convex, the directional derivative $f'_-(\bar{x};\cdot)$ is considered in [V]. We return to this point in the next section.

THEOREM 3.13. *Assume that P satisfies Hypothesis* (H0) *and* int $P \neq \emptyset$. *Let $f : X \to Y \cup \{+\infty\}$, $\bar{x} \in \text{dom } f$. Then,*

(a) *if $f'_-(\bar{x};\cdot)$ is P-lsc at u, then $f'_-(\bar{x};u) = \underline{D}_e^R f(\bar{x};u)$;*

(b) *if $f'_+(\bar{x};\cdot)$ is P-usc at u, then $f'_+(\bar{x};u) = \overline{D}_e^R f(\bar{x};u)$.*

Proof. Let us prove only part (a), the other one being similar. Clearly if $f'_-(\bar{x};u) = -\infty$, then $\underline{D}_e^R f(\bar{x};u) = -\infty$ by (3.9). We consider the case $f'_-(\bar{x};u) =$

$+\infty$. For every $y_0 \in Y$, the P-lower semicontinuity at u (see Definition 2.5) implies the existence of $\varepsilon_0 > 0$ such that for $||u' - u|| < \varepsilon_0$ we have

$$f'_-(\bar{x}; u') \in y_0 + \text{int } P + (P \cup \{+\infty\}) \subset y_0 + (P \cup \{+\infty\}).$$

Then

$$f'_-(\bar{x}; u') \geq y_0 \quad \text{whenever} \quad ||u' - u|| < \varepsilon_0.$$

Thus by definition of $\underline{D}_e^R f(\bar{x}; \cdot)$ one obtains

$$\underline{D}_e^R f(\bar{x}; u) \geq \inf_{||u'-u|| < \varepsilon_0} f'_-(\bar{x}; u') \geq y_0.$$

Since the latter holds for every $y_0 \in Y$, we obtain $\underline{D}_e^R f(\bar{x}; u) = +\infty$. The case $f'_-(\bar{x}; u) \in Y$ is treated by taking $y_0 = f'_-(\bar{x}; u) - \lambda x_0$ in the previous reasoning, where $x_0 \in \text{int } P$ is fixed and $\lambda > 0$ is arbitrary, since in this case $f'_-(\bar{x}; u) - \lambda x_0 + \text{int } P$ is an open set containing $f'_-(\bar{x}; u)$. Then, we let λ go to 0 and use the closedness of P to conclude with $\underline{D}_e^R f(\bar{x}; u) \geq f'_-(\bar{x}; u)$; the other inclusion follows from (3.9). □

Remark 3.14. When $Y = \mathbb{R}$, $\overline{D}_e^R f(\bar{x}; \cdot)$ is already usc in the usual sense (see Corollary 3.4), but it could happen that $f'_+(\bar{x}; \cdot)$ is lsc, as the function $f(x) = x^3$, $x \in \mathbb{R}$, shows. Here $f'_+(0; u) = 0$ if $u \leq 0$, $f'_+(0; u) = +\infty$ if $u > 0$, whereas $\overline{D}_e^R f(0; u) = 0$ if $u < 0$, $\overline{D}_e^R f(0; u) = +\infty$ if $u \geq 0$.

As a consequence of Proposition 3.12 and Theorem 3.13, we have the following result.

COROLLARY 3.15. *Assume that P satisfies Hypothesis* (H0) *and* int $P \neq \emptyset$. *Let $f : X \to Y \cup \{+\infty\}$, $\bar{x} \in$ dom f, such that $f'_+(\bar{x}; \cdot) : X \to Y \cup \{+\infty\}$ is P-usc on X and $\{x' \in X : f(x') \in f(\bar{x}) - P\}$ is convex and closed. Then*

$$\begin{aligned}\left\{u \in X : \ \overline{D}_e^R f(\bar{x}; u) \in -P\right\} &= \left\{u \in X : \ f(\bar{x} + tu) - f(\bar{x}) \in -P \ \ \forall\, t > 0\right\} \\ &= \left\{x' \in X : f(x') \in f(\bar{x}) - P\right\}^{\infty}.\end{aligned}$$

Proof. One inclusion of the first equality follows from Proposition 3.12 as well as the second equality. The other inclusion is implied by Theorem 3.13. □

Remark 3.16. One cannot drop the assumption that $f'_+(\bar{x}; \cdot)$ is P-usc on X, as the function $f(x) = \sqrt{|x|}$, $x \in \mathbb{R}$, with $\bar{x} = 0$ shows, since in this case $\overline{D}_e^R f(0; u) = +\infty$, $u \in \mathbb{R}$, whereas $f'_+(0; 0) = 0$, $f'_+(0; u) = +\infty$ if $u \neq 0$. Notice that this function is not convex. However, the result expressed in the previous corollary is the analogue to that satisfied in the scalar case when f is convex and lsc, by substituting $\overline{D}_e^R f(\bar{x}; \cdot)$ by the asymptotic function f^∞ and $P = [0, +\infty[$.

Given $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$, $x \in$ dom f, set

$$S_f(x) = \left\{y \in \mathbb{R}^n : \ f(y) \leq f(x)\right\}.$$

By specializing $Y = \mathbb{R}$ in the previous corollary, we obtain the following.

COROLLARY 3.17. *Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be lsc and quasi-convex. Let $\bar{x} \in$ dom f. Assume that $f'_+(\bar{x}; \cdot) : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is usc. Then*

$$S_f(\bar{x}) \ \text{ is bounded } \iff \overline{D}_e^R f(\bar{x}; u) > 0 \ \ \forall\, u \in \mathbb{R}^n, \ u \neq 0.$$

Example 3.18. Take the function $f(x) = \frac{x^2}{1+x^2}$, $x \in \mathbb{R}$. For $0 \leq x < \frac{1}{\sqrt{3}}$, easy computations show that

$$f'_+(x;u) = \begin{cases} \dfrac{\sqrt{1+x^2}+x}{2(1+x^2)}u & \text{if } \ u > 0, \\ -\dfrac{\sqrt{1+x^2}-x}{2(1+x^2)}u & \text{if } \ u \leq 0. \end{cases}$$

Thus, because of Theorem 3.13, $\overline{D}_e^R f(x;u) = f'_+(x;u)$ for $u \in \mathbb{R}$, $0 \leq x < \frac{1}{\sqrt{3}}$.

Example 3.19 (The classical convex case). Let us consider $X = \mathbb{R}^n$, $Y = \mathbb{R}^m$, $P = \mathbb{R}^m_+$, and $f = (f_1, \ldots, f_m) : \mathbb{R}^n \to \mathbb{R}^m$ an $\mathbb{R}^m_+$-convex function; that is, each $f_i : \mathbb{R}^n \to \mathbb{R}$ is a convex (hence continuous) function for $i = 1, \ldots, m$. For a fixed $\bar{x} \in \mathbb{R}^n$, it is assumed that

$$f'_+(\bar{x};u) = \sup_{t>0} \frac{f(\bar{x}+tu) - f(\bar{x})}{t}$$

is continuous as a function of $u \in \mathbb{R}^n$. Theorems 3.10 and 3.13 imply

$$\overline{D}_e^R f(\bar{x};u) = f'_+(\bar{x};u) \quad \forall\ u \in \mathbb{R}^n.$$

The latter is independent of $\bar{x}$ due to the convexity of f_i (see Theorem 3.10).

We identify any element in $\mathbb{R}^m$ having at least one component equal to $+\infty$, the extended element in $\mathbb{R}$, with $+\infty$ (the artificial element added to $\mathbb{R}^m$), to obtain

$$\overline{D}_e^R f(\bar{x};u) = (f_1^\infty(u), \ldots, f_m^\infty(u)), \quad u \in \mathbb{R}^n.$$

Here, as before, f_i^∞ stands for the asymptotic function of f_i, defined as usual.

The following proposition, whose proof is straightforward, arises in the case in which we have a constrained minimization problem.

PROPOSITION 3.20. *Let $f : X \to Y \cup \{+\infty\}$; $K \subset X$, $\bar{x} \in X$ such that $\bar{x} \in$ dom $f \cap K$. If $f(x) - f(\bar{x}) \in P \cup \{+\infty\}$ for all $x \in K$, then*

(a) $\underline{D}_e^R f(\bar{x};u) \in P \cup \{+\infty\}$ *for all* $u \in R^i(K;\bar{x})$;

(b) $\overline{D}_e^R f(\bar{x};u) \in P \cup \{+\infty\}$ *for all* $u \in R(K;\bar{x})$.

Example 3.21. Let $f : X \to \mathbb{R}$ be a real-valued locally Lipschitz function. By [HU1] and as can be verified, the condition $f^0(\bar{x};u) \geq 0$ for all $u \in X$ gives a necessary condition for $\bar{x}$ to be a local minimizer for f. Here $f^0(\bar{x};\cdot)$ denotes the Clarke directional derivative of f at $\bar{x}$ (see [Cl]). As $f^0(\bar{x};u) \geq \underline{D}_e^R f(\bar{x};u)$, our optimality condition given by (c) of Corollary 3.5 yields more information. On the other hand, the sufficient condition imposed in Theorem 3.3 of [BZ] certainly implies our condition (see also section 4 in [F3]).

4. The vector minimization problem. In order to define the vector minimization problem, we are given two cones P, W in Y, where P defines the underlying preference relation on Y. The basic assumptions on P, W are listed in the following hypothesis.

Hypothesis (H1). $P \subset Y$ is a convex closed pointed cone, and $W \subset Y$ is a closed cone such that $W \neq Y$ and $W + P \subset W$.

In some results the pointedness of P is not needed. Examples for W are $W = P$, $W = Y \setminus (-\text{int } P)$.

We now consider the problem of finding

$$\bar{x} \in X : \ f(x) - f(\bar{x}) \in W \cup \{+\infty\} \ \ \forall\, x \in X. \tag{4.1}$$

The solution set to problem (4.1), that is, the set of $\bar{x} \in X$ satisfying (4.1), is denoted by E_W. The problem of the existence of solutions to (4.1) was discussed in [F2] and for more general problems in [FF] in the finite dimensional setting; see also [O]. Thus, our main concern in this section is the asymptotic description of the solution set, together with some optimality conditions. Some of them were already given in Corollary 3.5.

We introduce the following cones:

$$R_P \doteq \bigcap_{x \in \operatorname{dom} f} \Big\{u \in X : \ f(x+tu) - f(x) \in -P \ \ \forall\, t > 0\Big\},$$
$$R_W \doteq \bigcap_{x \in \operatorname{dom} f} \Big\{u \in X : \ f(x+tu) - f(x) \in -W \ \ \forall\, t > 0\Big\},$$

which are nonempty (since $0 \in W$) closed cones (under the P-lsc on f) not necessarily convex. Clearly $R_P \subset R_W$. These sets were introduced in [F1] when $Y = \mathbb{R}$, and in the case of finite dimensional spaces in [F2, FF].

The importance of such cones lies in the next theorem, but before proceeding we introduce the following notion, which captures the usual P-convexity and the properly P-quasiconvexity, as introduced in [Fe].

Definition 4.1. *Let P be a convex closed cone, W a closed cone. We say $f : X \to Y \cup \{+\infty\}$ is (P,W)-convex if, for every $\lambda \in Y$, every $x, y \in \operatorname{dom} f$, $f(x) \in \lambda - P$, $f(y) \in \lambda - W$, we have*

$$f(\alpha x + (1-\alpha)y) \in \lambda - W \quad \forall\, \alpha \in \ ]0,1[.$$

One can see that every P-convex function is (P,W)-convex for all W satisfying $W + P \subset W$. The same is true if f is properly P-quasi-convex (see [Fe]), that is, for every $x, y \in \operatorname{dom} f$, for all $\alpha \in \ ]0,1[$,

$$f(\alpha x + (1-\alpha)y) \in f(x) - P \quad \text{or} \quad f(\alpha x + (1-\alpha)y) \in f(y) - P.$$

Theorem 4.2. *Let P, W satisfy Hypothesis* (H1)*. Assume that the vector-valued function $f : X \to Y \cup \{+\infty\}$ is (P,W)-convex, and also assume that for every $x \in \operatorname{dom} f$ the set $\{y \in X : \ f(y) - f(x) \in -W\}$ is closed. If $E_W \neq \emptyset$, then*

$$R_P \subset (E_W)^\infty \subset R_W = \bigcap_{y \in \operatorname{dom} f} \Big\{x \in \operatorname{dom} f : \ f(x) - f(y) \in -W\Big\}^\infty, \tag{4.2}$$

where E_W is as before. If, in addition, the set of ideal solutions E_P (some people call them strong solutions), i.e., solutions to (4.1) *with $W = P$, is nonempty, then $(E_W)^\infty = R_W$.*

Proof. We prove the first inclusion. Let $\bar{x} \in E_W$ and $u \in R$. In particular, $f(\bar{x}) - f(\bar{x} + tu) \in P$ for all $t > 0$. On the other hand, $f(x) - f(\bar{x}) \in W$ for all $x \in \operatorname{dom} f$. Hence $f(x) - f(\bar{x} + tu) \in W + P \subset W$ for all $x \in \operatorname{dom} f$, proving that $\bar{x} + tu \in E_W$ for all $t > 0$, i.e., $u \in (E_W)^\infty$. The second inclusion is as follows. Let $u \in (E_W)^\infty$. Then there exist $t_n \downarrow 0$, $u_n \in E_W$ such that $t_n u_n \to u$. For

$x \in \operatorname{dom} f$ arbitrary we have $f(u_n) - f(x) \in -W$ for all $n \in \mathbb{N}$. Fixing any $t > 0$, the (P, W)-convexity of f yields for all n sufficiently large

$$f((1 - tt_n)x + tt_n u_n) - f(x) \in -W.$$

By assumption, it follows that $f(x + tu) - f(x) \in -W$. This proves $u \in R_W$.

We now prove the inclusion "$\subseteq$" for the equality in (4.2); the other is left as an exercise. Let $u \in X$ such that $f(x) - f(x + tu) \in W$ for all $t > 0$ and all $x \in \operatorname{dom} f$. Given any $y \in \operatorname{dom} f$, set $x_n \doteq y + nu \in \operatorname{dom} f$, $n \in \mathbb{N}$. Then $f(y) - f(x_n) \in W$ for all $n \in \mathbb{N}$. By choosing $t_n = \frac{1}{n}$, we have $t_n x_n = \frac{y}{n} + u \to u$, i.e., $u \in \{x \in \operatorname{dom} f : \ f(y) - f(x) \in W\}^\infty$. Since $y \in \operatorname{dom} f$ was arbitrary, the proof of the last inclusion is complete. To prove the last part of the theorem, we need to show $R_W \subset (E_W)^\infty$. Take any $u \in R_W$ and $z \in E_P$. Then $f(y) - f(z + tu) = f(y) - f(z) + f(z) - f(z + tu) \in P + W \subset W$ for all $t > 0$ and all $y \in \operatorname{dom} f$. This implies $z + tu \in E_W$ for all $t > 0$. Hence $u \in (E_W)^\infty$. □

The inclusions in (4.2) may be strict, as Examples 5.3 and 5.6 in [F2] show. Such inclusions are used in the same paper to obtain the existence of solutions to problem (4.1) under the P-convexity condition. In fact, if X, Y are finite dimensional spaces, any condition implying $R_W = \{0\}$ will guarantee the nonemptiness and compactness of the solution set as established in [F2]. Thus one recovers some of the results in [D, CC2]. In infinite dimensional spaces we need additional assumptions. For other results concerning the finite dimensional setting, we refer to [F2] (see also [FF]), as well as [F4], where a unified approach for convex/nonconvex vector minimization problems is proposed.

A vector function $f : X \to Y \cup \{+\infty\}$ is said to be P-quasi-convex [Fe, L1] if it is (P, P)-convex in the sense of Definition 4.1.

We single out the result of Theorem 4.2 in the case when $W = P$.

THEOREM 4.3. *Let P be a closed convex pointed cone; let $f : X \to Y \cup \{+\infty\}$ be P-quasi-convex such that, for every $x \in \operatorname{dom} f$, the set $\{y \in X : \ f(y) - f(x) \in -P\}$ is closed. If the set of ideal solutions, i.e., solutions to* (4.1) *with $W = P$, is nonempty, then*

$$\begin{aligned}(E_P)^\infty &= \bigcap_{y \in \operatorname{dom} f} \Big\{u \in X : \ f(y + tu) - f(y) \in -P \ \ \forall\, t > 0\Big\} \\ &= \bigcap_{y \in \operatorname{dom} f} \Big\{x \in X : \ f(x) - f(y) \in -P\Big\}^\infty.\end{aligned}$$

Proof. The first equality is a direct consequence of (4.2), setting $W = P$. The second equality follows from Proposition 3.1. □

We have immediately the following theorem, whose proof is a consequence of Corollary 3.15 together with Theorems 3.10, 3.13, 4.3 and Proposition 2.8.

THEOREM 4.4. *Assume that P satisfies Hypothesis* (H0) *with* int $P \neq \emptyset$*; let $f : X \to Y \cup \{+\infty\}$ be a function such that* epi f *is a closed and convex set; and let $\bar{x} \in \operatorname{dom} f$ with* $\operatorname{epi}(\overline{D}_e^R f(\bar{x}; \cdot))$ *being a closed set. Furthermore, assume that* $\operatorname{int}((\operatorname{epi} f)^\infty) \neq \emptyset$ *and $u \mapsto f'_+(\bar{x}; u) \in Y \cup \{+\infty\}$ is P-usc on X. If $E_P \neq \emptyset$, then*

$$\begin{aligned}(E_P)^\infty &= \Big\{u \in X : \ \overline{D}_e^R f(\bar{x}; u) \in -P\Big\} = \Big\{u \in X : \ f(\bar{x} + tu) - f(\bar{x}) \in -P \ \ \forall\, t > 0\Big\} \\ &= \Big\{y \in X : \ f(y) - f(\bar{x}) \in -P\Big\}^\infty.\end{aligned}$$

The preceding theorem is a vector version of the similar result found in scalar minimization problems for convex lsc functions.

Having defined some kinds of derivatives, we can now introduce some notions of subdifferentials. To that purpose, $L(X,Y)$ will denote the set of continuous linear mappings from X into Y.

DEFINITION 4.5. *For a function $f : X \to Y \cup \{+\infty\}$, $\bar{x} \in \operatorname{dom} f$,*

(i) *the strong subdifferential of f at $\bar{x}$ is defined by*

$$\partial_P f(\bar{x}) = \Big\{A \in L(X,Y) : \ f(x) - f(\bar{x}) \in A(x-\bar{x}) + (P \cup \{+\infty\}) \quad \forall\ x \in X\Big\};$$

(ii) *the radial lower strong subdifferential of f at $\bar{x}$ is defined by*

$$\underline{\partial}_P^R f(\bar{x}) = \Big\{A \in L(X,Y) : \ \underline{D}_e^R f(\bar{x};u) \in A(u) + (P \cup \{+\infty\}) \quad \forall\ u \in X\Big\};$$

(iii) *the radial upper strong subdifferential of f at $\bar{x}$ is defined by*

$$\overline{\partial}_P^R f(\bar{x}) = \Big\{A \in L(X,Y) : \ \overline{D}_e^R f(\bar{x};u) \in A(u) + (P \cup \{+\infty\}) \quad \forall\ u \in X\Big\}.$$

Similarly, by replacing P by W (both satisfying Hypothesis (H1)) in the definitions above, we introduce, respectively, the weak subdifferential, radial lower weak subdifferential, and the radial upper weak subdifferential of f at $\bar{x} \in \operatorname{dom} f$, $\partial_W f(\bar{x})$, $\underline{\partial}_W^R f(\bar{x})$, $\overline{\partial}_W^R f(\bar{x})$.

Conditions ensuring the nonemptiness of $\partial_P f$ may be found in [V, Bo, Z], and for $\partial_W f$ in [CC1, CJ, Y1] under convexity assumptions.

By using Proposition 3.1, Theorem 3.3, and the definitions above, one can easily obtain the following result, whose second part is analogous to that in [P2].

PROPOSITION 4.6. *Assume that (Y,P) is order-complete, with P satisfying Hypothesis* (H1). *Let $f : X \to Y \cup \{+\infty\}$ be any function, $\bar{x} \in \operatorname{dom} f$. Then*

(a) $\underline{\partial}_P^R f(\bar{x}) \subset \partial_P f(\bar{x}) \subset \overline{\partial}_P^R f(\bar{x})$, $\underline{\partial}_W^R f(\bar{x}) \subset \partial_W f(\bar{x}) \subset \overline{\partial}_W^R f(\bar{x})$;

(b) $\underline{\partial}_P^R f(\bar{x}) \subset \{A \in L(X,Y) : \ A(u) \le v \quad \forall\ (u,v) \in R(\operatorname{epi} f; (\bar{x},\bar{y}))\}$;

(c) *if, in addition,* $\operatorname{int} P \neq \emptyset$, *we have*

$$\overline{\partial}_P^R f(\bar{x}) = \Big\{A \in L(X,Y) : \ A(u) \le v \quad \forall\ (u,v) \in R^i(\operatorname{epi}\ f; (\bar{x},\bar{y}))\Big\}.$$

In [V], when f is P-convex, the subdifferential is defined

$$\partial_c f(\bar{x}) = \Big\{A \in L(X,Y) : \ f'_-(\bar{x};u) \in A(u) + (P \cup \{+\infty\}) \ \ \forall\ u \in X\Big\}.$$

It follows from [Th3] that

$$f'_-(\bar{x};u) = \sup\Big\{A(u) : \ A \in \partial_c f(\bar{x})\Big\}$$

whenever $f'_-(\bar{x};\cdot)$ is continuous. Moreover, under this assumption, Theorem 3.13 implies $f'_-(\bar{x};u) = \underline{D}_e^R f(\bar{x};u)$, and therefore

$$\partial_c f(\bar{x}) = \underline{\partial}_P^R f(\bar{x}).$$

We close this section by writing the optimality conditions expressed in Corollary 3.5 in terms of the subdifferentials just defined.

PROPOSITION 4.7. *Assume that (Y,P) is order-complete, with P satisfying Hypothesis* (H1). *Given any $f : X \to Y \cup \{+\infty\}$, $\bar{x} \in \operatorname{dom} f$, we have*

(a) $0 \in \partial_P f(\bar{x}) \iff f(x) - f(\bar{x}) \in P \cup \{+\infty\}$ *for all* $x \in X \iff 0 \in \underline{\partial}_P^R f(\bar{x})$;

(b) $0 \in \underline{\partial}_W^R f(\bar{x}) \Longrightarrow f(x) - f(\bar{x}) \in W \cup \{+\infty\}$ *for all* $x \in X \Longleftrightarrow 0 \in \partial_W f(\bar{x}) \Longrightarrow 0 \in \overline{\partial}_W^R f(\bar{x})$.

5. Conclusions. In this paper the concepts of lower/upper radial epiderivatives for vector-valued functions are introduced, and they are shown to be useful in the study of nonconvex vector optimization problems. In particular, the strong (ideal) solutions are completely characterized in terms of the lower radial epiderivative. Such a characterization appears new in the literature even in the real-valued case, except in [F3], where it is derived in a different way.

On the other hand, some optimality conditions required for a point to be a weakly efficient (weak-Pareto) solution are also derived by means of these radial epiderivatives.

Moreover, the upper radial epiderivative is used to propose a notion of asymptotic function for a class of vector-valued functions. In order to describe the asymptotic behavior of the solution set for a vector optimization problem, some cones of asymptotic directions are introduced.

We believe that the approach developed in this paper, when applied to real-valued quasi-convex functions, yields new results.

A unified approach to dealing with the existence of solutions to nonconvex vector minimization problems in finite dimensional spaces is presented in [F4].

Acknowledgments. The author wishes to express his gratitude to two anonymous referees for their valuable remarks, which led to the present improved version of this article.

REFERENCES

[A] J.P. Aubin, *Contingent derivatives of set-valued maps and existence of solutions in nonlinear inclusions and differential inclusions*, in Advances in Mathematics, Supplementary Studies 7A, L. Nachbin, ed., Academic Press, New York, 1981, pp. 160–232.

[AE] J.P. Aubin and I. Ekeland, *Applied Nonlinear Analysis*, John Wiley & Sons, New York, 1984.

[AF] J.P. Aubin and H. Frankowska, *Set-Valued Analysis*, Birkhäuser Boston, Basel, Berlin, 1990.

[BHU] J. Benoist and J.-B. Hiriart-Urruty, *What is the subdifferential of the closed convex hull of a function?*, SIAM J. Math. Anal., 27 (1996), pp. 1661–1679.

[BHS] M. Bianchi, N. Hadjisavvas, and S. Schaible, *Vector equilibrium problems with monotone bifunctions*, J. Optim. Theory Appl., 92 (1997), pp. 527–542.

[BZ] A. Ben-Tal and J. Zowe, *Directional derivatives in nonsmooth optimization*, J. Optim. Theory Appl., 47 (1985), pp. 483–490.

[Bo] J.M. Borwein, *A Lagrange multiplier theorem and a sandwich theorem for convex relation*, Math. Scand., 48 (1981), pp. 189–204.

[Cl] F.H. Clarke, *Optimization and Nonsmooth Analysis*, Classics Appl. Math. 5, SIAM, Philadelphia, 1990.

[CC1] G.Y. Chen and B.D. Craven, *A vector variational inequality and optimization over an efficient set*, ZOR—Math. Methods Oper. Res., 34 (1990), pp. 1–12.

[CC2] G.Y. Chen and B.D. Craven, *Existence and continuity of solutions for vector optimization*, J. Optim. Theory Appl., 81 (1994), pp. 459–468.

[CJ] G.Y. Chen and J. Jahn, *Optimality conditions for set-valued optimization*, Math. Methods Oper. Res., 48 (1998), pp. 187–200.

[C] H.W. Corley, *Optimality conditions for maximizations of set-valued functions*, J. Optim. Theory Appl., 58 (1988), pp. 1–10.

[D] S. Deng, *Characterizations of the nonemptiness and compactness of solutions sets in convex vector optimization*, J. Optim. Theory Appl., 96 (1998), pp. 123–131.

[D1] J.-P. Dedieu, *Cône asymptote d'un ensemble non convexe, application à l'optimization*, C. R. Acad. Sci. Paris Sér. A-B, 285 (1977), pp. 501–503.

[D2] J.-P. Dedieu, *Cônes asymptotes d'ensembles non convexes*, Bull. Soc. Math. France Mém., 60 (1979), pp. 31–44.

[Fe] F. Ferro, *Minimax type theorems for n-valued functions*, Ann. Mat. Pura Appl., 32 (1982), pp. 113–130.

[F1] F. Flores-Bazán, *Existence theorems for generalized noncoercive equilibrium problems: The quasi-convex case*, SIAM J. Optim., 11 (2000), pp. 675–690.

[F2] F. Flores-Bazán, *Ideal, weakly efficient solutions for vector optimization problems*, Math. Program., 93 (2002), pp. 453–475.

[F3] F. Flores-Bazán, *Optimality conditions in non-convex set-valued optimization*, Math. Methods Oper. Res., 53 (2001), pp. 403–417.

[F4] F. Flores-Bazán, *Semi-strictly quasi-convex vector functions and nonconvex vector optimization*, preprint DIM 2003-03, Universidad de Concepción, Octava Región, Chile, 2003.

[FF] F. Flores-Bazán and F. Flores-Bazán, *Vector equilibrium problems under asymptotic analysis*, J. Global Optim., 26 (2003), pp. 141–166.

[FO] F. Flores-Bazán and W. Oettli, *Simplified optimality conditions for minimizing a difference of vector-valued functions*, J. Optim. Theory Appl., 108 (2001), pp. 571–586.

[GJ] A. Götz and J. Jahn, *The Lagrange multiplier rule in set-valued optimization*, SIAM J. Optim., 10 (1999), pp. 331–344.

[HU1] J.B. Hiriart-Urruty, *On optimality conditions in nondifferentiable programming*, Math. Programming, 14 (1978), pp. 73–86.

[HU2] J.B. Hiriart-Urruty, *Tangent cones, generalized gradients and mathematical programming in Banach spaces*, Math. Oper. Res., 4 (1979), pp. 79–97.

[J] J. Jahn, *Introduction to the Theory of Nonlinear Optimization*, Springer-Verlag, Berlin, 1996.

[JR] J. Jahn and R. Rauh, *Contingent epiderivatives and set-valued optimization*, Math. Methods Oper. Res., 46 (1997), pp. 193–211.

[L1] D.T. Luc, *Theory of Vector Optimization*, Lecture Notes in Econom. and Math. Systems 319, Springer-Verlag, New York, Berlin, 1989.

[L2] D.T. Luc, *Contingent derivatives of set-valued maps and applications to vector optimization*, Math. Programming, 50 (1991), pp. 99–111.

[O] W. Oettli, *A remark on vector-valued equilibria and generalized monotonicity*, Acta Math. Vietnamica, 22 (1997), pp. 213–221.

[P1] J.-P. Penot, *Sous-différentiels de fonctions numériques non convexes*, C. R. Acad. Sci. Paris Sér. A, 278 (1974), pp. 1553–1555.

[P2] J.-P. Penot, *Calcul sous-différentiel et optimisation*, J. Funct. Anal., 27 (1978), pp. 248–276.

[P3] J.-P. Penot, *Differentiability of relations and differential stability of perturbed optimization problems*, SIAM J. Control Optim., 22 (1984), pp. 529–551.

[PT] J.-P. Penot and M. Thera, *Polarité des applications convexes à valeurs vectorielles*, C. R. Acad. Sci. Paris Sér. A-B, 288 (1979), pp. 419–422.

[T] A. Taa, *Set-valued derivatives of multifunctions and optimality conditions*, Numer. Funct. Anal. Optim., 19 (1998), pp. 121–140.

[Th1] L. Thibault, *Epidifférentiels de fonctions vectorielles*, C. R. Acad. Sci. Paris Sér. A-B, 290 (1980), pp. 87–90.

[Th2] L. Thibault, *On generalized differentials and subdifferentials of Lipschitz vector-valued functions*, Nonlinear Anal., 6 (1982), pp. 1037–1053.

[Th3] L. Thibault, *Subdifferentials of non-convex vector-valued functions*, J. Math. Anal. Appl., 86 (1982), pp. 319–344.

[V] M. Valadier, *Sous-différential de fonctions convexes à valeurs dans un espace vectoriel ordonné*, Math. Scand., 30 (1972), pp. 65–72.

[Y1] X.Q. Yang, *A Hahn–Banach theorem in ordered linear spaces and its applications*, Optimization, 25 (1992), pp. 1–9.

[Y2] X.Q. Yang, *Directional derivatives for set-valued mappings and applications*, Math. Methods Oper. Res., 48 (1998), pp. 273–285.

[Z] J. Zowe, *Subdifferentiability of convex functions with values in ordered vector spaces*, Math. Scand., 34 (1974), pp. 69–83.

SIAM J. OPTIM.
Vol. 14, No. 2, pp. 307–333

COMPUTATIONAL EXPERIENCE AND THE EXPLANATORY VALUE OF CONDITION MEASURES FOR LINEAR OPTIMIZATION*

FERNANDO ORDÓÑEZ† AND ROBERT M. FREUND‡

Abstract. The modern theory of condition measures for convex optimization problems was initially developed for convex problems in the conic format

$$(\mathrm{CP}_d) \qquad z_* := \min_x \{c^t x \mid Ax - b \in C_Y, x \in C_X\},$$

and several aspects of the theory have now been extended to handle nonconic formats as well. In this theory, the (Renegar) condition measure $C(d)$ for (CP_d) has been shown to be connected to bounds on a wide variety of behavioral and computational characteristics of (CP_d), from sizes of optimal solutions to the complexity of algorithms for solving (CP_d). Herein we test the practical relevance of the condition measure theory, as applied to linear optimization problems that one might typically encounter in practice. Using the NETLIB suite of linear optimization problems as a test bed, we found that 71% of the NETLIB suite problem instances have infinite condition measure. In order to examine condition measures of the problems that are the actual input to a modern interior-point-method (IPM) solver, we also computed condition measures for the NETLIB suite problems after preprocessing by CPLEX 7.1. Here we found that 19% of the postprocessed problem instances in the NETLIB suite have infinite condition measure, and that $\log C(d)$ of the postprocessed problems is fairly nicely distributed. Furthermore, among those problem instances with finite condition measure after preprocessing, there is a positive linear relationship between IPM iterations and $\log C(d)$ of the postprocessed problem instances (significant at the 95% confidence level), and 42% of the variation in IPM iterations among these NETLIB suite problem instances is accounted for by $\log C(d)$ of the postprocessed problem instances.

Key words. condition measure, interior-point method, linear programming, computation, preprocessing

AMS subject classifications. 90-04, 90C05, 90C60, 90C51

DOI. 10.1137/S1052623402401804

1. Introduction.

The modern theory of condition measures for convex optimization problems was initially developed in [24] for problems in the following conic format:

$$(\mathrm{CP}_d) \qquad \begin{array}{ll} z_* := \min & c^t x \\ \text{s.t.} & Ax - b \in C_Y, \\ & x \in C_X, \end{array} \tag{1.1}$$

where, for concreteness, we consider A to be an $m \times n$ real matrix; $b \in \mathbb{R}^m$, $c \in \mathbb{R}^n$; $C_X \subseteq \mathbb{R}^n$, $C_Y \subseteq \mathbb{R}^m$ are closed convex cones; and the data of the problem is the array $d = (A, b, c)$. We assume that we are given norms $\|x\|$ and $\|y\|$ on $\mathbb{R}^n$ and $\mathbb{R}^m$, respectively, and let $\|A\|$ denote the usual operator norm; let $\|v\|_*$ denote the dual norm associated with the norm $\|w\|$ on $\mathbb{R}^n$ or $\mathbb{R}^m$. We define the norm of the data instance $d = (A, b, c)$ by $\|d\| := \max\{\|A\|, \|b\|, \|c\|_*\}$.

The theory of condition measures for (CP_d) focuses on three measures, $\rho_P(d)$, $\rho_D(d)$, and $C(d)$, to bound various behavioral and computational quantities pertaining

*Received by the editors January 31, 2002; accepted for publication (in revised form) February 10, 2003; published electronically August 22, 2003. This research was partially supported through the Singapore-MIT Alliance.

http://www.siam.org/journals/siopt/14-2/40180.html

†Industrial and Systems Engineering, University of Southern California, GER-247, Los Angeles, CA 90089-0193 (fordon@usc.edu).

‡MIT Sloan School of Management, 50 Memorial Drive, Cambridge, MA 02142 (rfreund@mit.edu).

to (CP_d). The quantity $\rho_P(d)$ is called the "distance to primal infeasibility" and is defined as

$$\rho_P(d) := \inf\{\|\Delta d\| \mid X_{d+\Delta d} = \emptyset\},$$

where X_d denotes the feasible region of (CP_d):

$$X_d := \{x \in \mathbb{R}^n \mid Ax - b \in C_Y, x \in C_X\}.$$

The quantity $\rho_D(d)$ is called the "distance to dual infeasibility" for the conic dual (CD_d) of (CP_d),

$$(\mathrm{CD}_d) \qquad \begin{array}{rl} z^* := \max & b^t y \\ \text{s.t.} & c - A^t y \in C_X^*, \\ & y \in C_Y^*, \end{array} \tag{1.2}$$

and is defined similarly to $\rho_P(d)$ but using the dual problem instead. The quantity $C(d)$ is called the "condition measure" or the "condition number" of the problem instance d and is a (positively) scale-invariant reciprocal of the smallest data perturbation Δd that will render the perturbed data instance either primal or dual infeasible:

$$C(d) := \frac{\|d\|}{\min\{\rho_P(d), \rho_D(d)\}}; \tag{1.3}$$

a problem is called "ill-posed" if $\min\{\rho_P(d), \rho_D(d)\} = 0$, equivalently, $C(d) = \infty$. These three condition measure quantities have been shown in theory to be connected to a wide variety of bounds on behavioral characteristics of (CP_d) and its dual, including bounds on sizes of feasible solutions, bounds on sizes of optimal solutions, bounds on optimal objective values, bounds on the sizes and aspect ratios of inscribed balls in the feasible region, bounds on the rate of deformation of the feasible region under perturbation, bounds on changes in optimal objective values under perturbation, and numerical bounds related to the linear algebra computations of certain algorithms; see [24], [5], [4], [8], [9], [10], [29], [27], [30], [28], [20], [22]. In the context of interior-point methods for linear and semidefinite optimization, these same three condition measures have also been shown to be connected to various quantities of interest regarding the central trajectory; see [16] and [17]. The connection of these condition measures to the complexity of algorithms has been shown in [8], [9], [25], [2], [3], and some of the references contained therein. While this literature has focused almost exclusively on the conic format of (1.1), there have been some attempts to extend the theory to convex problems in structured nonconic formats; see Filipowski [4], Peña [21] and [19], and [18].

Given the theoretical importance of these many results, it is natural to ask what typical values of these condition measures might arise in practice. Are such problems typically well-posed or ill-posed? How are the condition measures of such problems distributed? We begin to answer these questions in this paper, where we compute and analyze these three condition measures for the NETLIB suite of industrial and academic linear programming (LP) problems. We present computational results that indicate that 71% of the NETLIB suite of linear optimization problem instances are ill-posed, i.e., have infinite condition measure; see section 4.1.

In the case of modern interior-point-method (IPM) algorithms for linear optimization, the number of IPM iterations needed to solve a linear optimization instance has been observed to vary from 10 to 100, over a huge range of problem sizes; see [13], for example. Using the condition-measure model for complexity analysis, one can bound the IPM iterations by $O(\sqrt{n}\log(C(d)+\cdots))$ for linear optimization in standard form, where the other terms in the bound are of a more technical nature; see [25] for details.

(Of course, the IPM algorithms that are used in practice are different from the IPM algorithms that are used in the development of the complexity theory.) A natural question to ask then is to what extent the observed variation in the number of IPM iterations (already small) can be accounted for by the condition measures of the LP problems that are solved. In order to answer this question, first note that typical IPM solvers perform routine preprocessing to modify the LP problem prior to solving. In order to examine condition measures of the problems that are the actual input to a modern IPM solver, we computed condition measures for the NETLIB suite problems after prepreprocessing by CPLEX 7.1. We found that 19% of the postprocessed problem instances in the NETLIB suite have infinite condition measure, and that $\log C(d)$ of the postprocessed problems is fairly nicely distributed; see section 4.2. In section 4.3, we show that, among the 72 postprocessed problem instances in the NETLIB suite with finite condition measure, the number of IPM iterations needed to solve these problems varies roughly linearly (and monotonically) with $\log C(d)$ of the postprocessed problem instances. A simple linear regression model of IPM iterations as the dependent variable and $\log C(d)$ as the independent variable yields a positive linear relationship between IPM iterations and $\log C(d)$ for the postprocessed problem instances, significant at the 95% confidence level, with $R^2 = 0.4160$. Therefore, in the sample of 72 NETLIB suite problem instances whose postprocessed condition measure is finite, about 42% of the variation in IPM iterations among these problems is accounted for by $\log C(d)$ of the problem instances after preprocessing. Additionally, $\log C(d)$ correlates with IPM iterations better than any other problem measure; see section 4.3.

The organization of this paper is as follows. In section 2, we lay the groundwork for the computation of condition measures for the NETLIB suite. Section 3 describes our methodology for computing condition measures, and section 4 contains the computational results. Section 5 contains some discussion and open questions.

2. Linear programming, conic format, and ground-set format. In order to attempt to address the issues raised in the previous section about practical computational experience and the relevance of condition measures, one can start by computing the condition measures for a suitably representative set of linear optimization instances that arise in practice, such as the NETLIB suite of industrial and academic linear optimization problems; see [15]. Practical methods for computing (or approximately computing) condition measures for convex optimization problems in conic format (CP_d) have been developed in [9] and [20], and such methods are relatively easy to implement. It would then seem to be a simple task to compute condition measures for the NETLIB suite. However, it turns out that there is a subtle catch that gets in the way of this simple strategy and in fact necessitates using an extension of the condition measure theory just a bit, as we now explain.

Linear optimization problems arising in practice are typically conveyed in the following format:

$$\begin{array}{lll} \min\limits_{x} & c^t x & \\ \text{s.t.} & A_i x \le b_i, & i \in L, \\ & A_i x = b_i, & i \in E, \\ & A_i x \ge b_i, & i \in G, \\ & x_j \ge l_j, & j \in L_B, \\ & x_j \le u_j, & j \in U_B, \end{array} \tag{2.1}$$

where the first three sets of inequalities/equalities are the "constraints" and the last

two sets of inequalities are the lower and upper bound conditions, and where $L_B, U_B \subset \{1, \ldots, n\}$. (LP problems in practice might also contain range constraints of the form "$b_{i,l} \leq A_i x \leq b_{i,u}$." We ignore this for now.) By defining C_Y to be an appropriate Cartesian product of nonnegative halflines $\mathbb{R}_+$, nonpositive halflines $-\mathbb{R}_+$, and the origin $\{0\}$, we can naturally consider the constraints to be in the conic format "$Ax-b \in C_Y$," where $C_Y \subset \mathbb{R}^m$ and $m = |L| + |E| + |G|$. However, for the lower and upper bounds on the variables, there are different ways to convert the problem into the required conic format for computation and analysis of condition measures. One way is to convert the lower and upper bound constraints into ordinary constraints, whose conversion of (2.1) to conic format is

$$\begin{array}{rll} P_1 : \min\limits_{x} & c^t x & \\ \text{s.t.} & Ax - b \in C_Y, & \\ & Ix - l \geq 0, & \\ & Ix - u \leq 0, & \end{array}$$

whose data for this now-conic format is

$$\bar{A} := \begin{pmatrix} A \\ I \\ I \end{pmatrix}, \quad \bar{b} := \begin{pmatrix} b \\ l \\ u \end{pmatrix}, \quad \bar{c} := c,$$

with cones

$$\bar{C}_Y := C_Y \times \mathbb{R}^n_+ \times -\mathbb{R}^n_+ \quad \text{and} \quad \bar{C}_X := \mathbb{R}^n.$$

Another way to convert the problem to conic format is to replace the variables x with nonnegative variables $s := x - l$ and $t := u - x$, yielding

$$\begin{array}{rll} P_2 : \min\limits_{s,t} & c^t s + c^t l & \\ \text{s.t.} & As - (b - Al) \in C_Y, & \\ & Is + It - (u - l) = 0, & \\ & s, t \geq 0, & \end{array}$$

whose data for this now-conic format is

$$\tilde{A} := \begin{pmatrix} A & 0 \\ I & I \end{pmatrix}, \quad \tilde{b} := \begin{pmatrix} b - Al \\ u - l \end{pmatrix}, \quad \tilde{c} := c,$$

with cones

$$\tilde{C}_Y := C_Y \times \{0\}^n \quad \text{and} \quad \tilde{C}_X := \mathbb{R}^n_+ \times \mathbb{R}^n_+.$$

These two different conic versions of the same original problem have different data and different cones, and so will generically have different condition measures. This is illustrated on the following elementary example:

$$\begin{array}{rll} P : \min\limits_{x_1, x_2} & x_1 & \\ \text{s.t.} & x_1 + x_2 \geq 1, & \\ & 400x_1 + x_2 \leq 420, & \\ & 1 \leq x_1 \leq 5, & \\ & -1 \leq x_2. & \end{array}$$

TABLE 2.1
Condition measures for two different conic conversions of the same problem, using the L_∞-norm in the space of the variables and the L_1-norm in the space of the right-hand-side vector.

	P_1	P_2
$\|d\|$	428	405
$\rho_P(d)$	0.24450	0.90909
$\rho_D(d)$	0.00250	1.00000
$C(d)$	171,200	445

Table 2.1 shows condition measures for problem P under the two different conversion strategies of P_1 and P_2, using the L_∞-norm in the space of the variables and the L_1-norm in the space of the right-hand-side vector. (The method for computing these condition measures is described in Remark 6 of [10].) As Table 2.1 shows, the choice of conversion strategy can have a very large impact on the resulting condition measures, thereby calling into question the practical significance of performing such conversions to conic format.

2.1. Structured formats for optimization. The analysis presented above indicates a need for extending condition-measure concepts to problems with structured nonconic formats, and indeed there has been some research along this line. Filipowski [4] examines the efficiency of solving symmetric-form linear programs whose sparsity pattern is not subject to modification, and Peña [21] develops condition measures for conic problems where certain rows and columns of data are not subject to modification; the latter can be used directly or indirectly to construct condition measures for many types of structured nonconic problems. More recently, in [18], the theory of condition measures and their properties has been extended from the conic format to handle more general structured convex optimization in the following "ground-set" format:

$$z_*(d) = \begin{array}{ll} \min & c^t x \\ \text{s.t.} & Ax - b \in C_Y, \\ & x \in P, \end{array} \qquad (\mathrm{GP}_d) \tag{2.2}$$

where P is called the ground set; P is no longer required to be a cone, but instead can be any closed convex set. In practical applications, P could be chosen to be the solution of lower and upper bound constraints $l \le x \le u$, or P could be a convex cone C_X, or P could perhaps be the solution to network flow constraints of the form $Nx = b$, $x \ge 0$, etc. The set P (and the cone C_Y) remains fixed as part of the definition of the problem, and the description of P is not part of the data $d = (A, b, c)$. Many aspects of the theory of condition measures for conic convex optimization have been extended to the more general ground-set model format (2.2); see [18]. We will use this ground-set format in our computation and evaluation of condition measures for linear programs that arise in practice.

In treating linear programs (2.1) as instances in the format (2.2), there is some leeway as to what structure to place in the ground set P. One strategy is to define P simply by the lower and upper bounds,

$$P := \{x \mid x_j \ge l_j \text{ for } j \in L_B,\ x_j \le u_j \text{ for } j \in U_B\}, \tag{2.3}$$

and then rewrite the other constraints in conic format as described earlier. In this approach the lower and upper bounds are handled conveniently, although the data d does not then include the lower and upper bound data $l_j, j \in L_B$ and $u_j, j \in U_B$.

This is somewhat advantageous since in many settings of linear optimization the lower and/or upper bounds on most variables are 0 or 1 or other scalars that are not generally thought of as subject to data modification. Of course, there are other settings where keeping the lower and upper bounds fixed independent of the other constraints is not as natural.

Another strategy would be to try to examine the individual constraints of the LP instance in order to identify specific structures to include in P. For example, in addition to lower and upper bound constraints, a particular LP instance might also have some network constraints, or might have generalized upper bound (GUB) constraints of the form $\sum_{j\in J} x_j \leq M$, variable upper bound constraints $x_j \leq x_k$, etc. Constraints of this type have no inherent data that one would think of as subject to possible modification; therefore they could be included in the set P.

In order to develop some computational experience with condition measures for the NETLIB suite, we chose the more straightforward strategy of defining P only by the upper and lower bounds of the LP instance (2.3). We chose this approach because (2.3) best reflects the types of LP structures that are explicitly treated algorithmically in modern simplex and IPM software, and because we had minimal foreknowledge of any explicit structures of individual linear prgrams in the NETLIB suite.

(In the related area of robust optimization, Ben-Tal and Nemirovski test robust optimization methodologies on the NETLIB suite by attempting to identify individual data entries of linear inequalities (but not equalities) in constraints of NETLIB suite linear programs that might be subject to data modification or data error; see [1].)

2.2. Definition of $C(d)$ for ground-set format. The general set-up for the development of condition-measure theory for the ground-set model format is developed in [18]. We review this material briefly here for completeness.

Let X_d denote the feasible region of (GP_d),

$$X_d := \{x \in \mathbb{R}^n \mid Ax - b \in C_Y, x \in P\},$$

and define the primal distance to infeasibility $\rho_P(d)$ as

$$\rho_P(d) := \inf\{\|\Delta d\| \mid X_{d+\Delta d} = \emptyset\},$$

similar to the conic case. In order to state the Lagrange dual of (GP_d) we use the following definitions, which depend on the ground set P.

Let R denote the recession cone of P, namely,

$$R := \{v \mid \text{there exists } x \in P \text{ for which } x + \theta v \in P \text{ for all } \theta \geq 0\}. \tag{2.4}$$

Since P is a closed convex set, the recession cone R is a closed convex cone.

Define

$$C_P := \{(x,t) \mid x \in tP, t > 0\},$$

and let C denote the closed convex cone

$$C := \mathrm{cl}C_P,$$

where "cl$S$" denotes the closure of a set S. Then it is straightforward to show that

$$C = C_P \cup \{(r,0) \mid r \in R\}$$

and that

$$\begin{aligned} C^* &:= \{(s,v) \mid s^t x + v \geq 0 \text{ for any } x \in P\} \\ &= \left\{(s,v) \mid \inf_{x \in P} s^t x \geq -v\right\}. \end{aligned}$$

The Lagrange dual of (GP_d) is

$$(\mathrm{GD}_d) \qquad \begin{array}{rll} z^*(d) = & \max\limits_{y,v} & b^t y - v \\ & \text{s.t.} & (c - A^t y, v) \in C^*, \\ & & y \in C_Y^*. \end{array} \tag{2.5}$$

Let Y_d denote the feasible region of the dual problem (GD_d),

$$Y_d := \{(y,v) \in \mathbb{R}^m \times \mathbb{R} \mid (c - A^t y, v) \in C^*,\ y \in C_Y^*\},$$

and define the dual distance to infeasibility $\rho_D(d)$:

$$\rho_D(d) := \inf\{\|\Delta d\| \mid Y_{d+\Delta d} = \emptyset\}.$$

The condition measures $\rho_P(d), \rho_D(d)$ are shown in [18] to be connected to a variety of behavioral characteristics of (GP_d) and its dual, including sizes of feasible solutions, sizes of optimal solutions, optimal objective values, aspect ratios of inscribed balls, deformation of the feasible region under perturbation, and the complexity of interior-point algorithms.

Let $\mathcal{F}$ denote the set of data instances d for which both (GP_d) and (GD_d) are feasible:

$$\mathcal{F} = \{d \mid X_d \neq \emptyset \text{ and } Y_d \neq \emptyset\}.$$

For $d \in \mathcal{F}$, the definition of the condition measure in the ground set model is identical to the definition in the conic case,

$$C(d) := \frac{\|d\|}{\min\{\rho_P(d), \rho_D(d)\}};$$

it is the (positive) scale invariant reciprocal of the distance to the set of data instances that are either primal or dual infeasible, and $\rho(d) := \min\{\rho_P(d), \rho_D(d)\}$ is the distance to ill-posedness.

3. Computation of $\rho_P(d)$, $\rho_D(d)$, and $C(d)$ via convex optimization. In this section we show how to compute $\rho_P(d)$ and $\rho_D(d)$ for linear optimization data instances $d = (A, b, c)$ of the ground-set model format, as well as how to estimate $\|d\|$ and $C(d)$. The methodology presented herein is an extension of the methodology for computing $\rho_P(d)$ and $\rho_D(d)$ developed in [9]. We will make the following choice of norms throughout this section and the rest of this paper.

ASSUMPTION 1. *The norm on the space of the x variables in $\mathbb{R}^n$ is the L_∞-norm, and the norm on the space of the right-hand-side vector in $\mathbb{R}^m$ is the L_1-norm.*

Using this choice of norms, we will show in this section how to compute $\rho(d)$ for linear optimization problems by solving $2n + 2m$ linear programs of size roughly that of the original problem. As is discussed in [9], the complexity of computing $\rho(d)$ very much depends on the chosen norms, with the norms given in Assumption 1 being

particularly appropriate for efficient computation of $\rho_P(d)$ and $\rho_D(d)$. We begin our analysis with a seemingly innocuous proposition which will prove to be very useful.

PROPOSITION 3.1. *Consider the problem*

$$\begin{array}{rll} z_1 = & \min\limits_{v,w} & f(v,w) \\ & \text{s.t.} & \|v\|_\infty = 1, \\ & & (v,w) \in K, \end{array} \tag{3.1}$$

where $v \in \mathbb{R}^k$, $w \in \mathbb{R}^l$, K is a closed convex cone in $\mathbb{R}^{k+l}$, and $f(\cdot) : \mathbb{R}^{k+l} \mapsto \mathbb{R}_+$ is positively homogeneous of degree one ($f(\alpha(v,w)) = |\alpha| f(v,w)$ for any $\alpha \in \mathbb{R}$ and $(v,w) \in \mathbb{R}^{k+l}$). Then problems (3.1) *and* (3.2) *have the same optimal values, i.e., $z_1 = z_2$, where*

$$\begin{array}{rll} z_2 = & \min\limits_{i \in \{1,\ldots,n\}, j \in \{-1,1\}} & \min\limits_{v,w} f(v,w) \\ & & v_i = j, \\ & & (v,w) \in K. \end{array} \tag{3.2}$$

Proof. Let (v^*, w^*) be an optimal solution of (3.1). Since $\|v^*\|_\infty = 1$, there exist $i^* \in \{1,\ldots,n\}$ and $j^* \in \{-1,1\}$ such that $v^*_{i^*} = j^*$. Therefore (v^*, w^*) is feasible for the inner problem in (3.2) for $i = i^*$ and $j = j^*$, and so $z_2 \le z_1$.

If (v^*, w^*) is an optimal solution of (3.2) with $i = i^*$ and $j = j^*$, then $\|v^*\|_\infty \ge 1$. If $\|v^*\|_\infty = 1$, the point (v^*, w^*) is feasible for (3.1), which means that $z_1 \le z_2$, completing the proof. Therefore, assume that $\|v^*\|_\infty > 1$, and consider the new point $(\tilde{v}, \tilde{w}) := \frac{1}{\|v^*\|_\infty}(v^*, w^*) \in K$. Then $(\tilde{v}, \tilde{w})$ is feasible for an inner problem in (3.2) for some $i = \hat{i} \neq i^*$ and $j = \hat{j}$, and so $z_2 \le f(\tilde{v}, \tilde{w}) = f(\frac{1}{\|v^*\|_\infty}(v^*, w^*)) = \frac{1}{\|v^*\|_\infty} f(v^*, w^*) \le z_2$, which now implies that $(\tilde{v}, \tilde{w})$ is also an optimal solution of (3.2). Since $\|\tilde{v}\|_\infty = 1$, the previous argument implies that $z_1 \le z_2$, completing the proof. ∎

3.1. Computing $\rho_P(d)$ and $\rho_D(d)$. The following theorem, which is proved in [18], characterizes $\rho_P(d)$ and $\rho_D(d)$ as the optimal solution values of certain optimization problems. In this theorem, recall from (2.4) that R denotes the recession cone of the ground set P.

THEOREM 3.2 (Theorems 5 and 6 of [18]). *Suppose $d \in \mathcal{F}$, and that the norms are chosen as in Assumption* 1. *Then $\rho_P(d) = j_P(d)$ and $\rho_D(d) = j_D(d)$, where*

$$\begin{array}{rll} j_P(d) = & \min\limits_{y,s,v} & \max\{\|A^t y + s\|_1, |b^t y - v|\} \\ & \text{s.t.} & \|y\|_\infty = 1, \\ & & y \in C_Y^*, \\ & & (s,v) \in C^*, \end{array} \tag{3.3}$$

and

$$\begin{array}{rll} j_D(d) = & \min\limits_{x,p,g} & \max\{\|Ax - p\|_1, |c^t x + g|\} \\ & \text{s.t.} & \|x\|_\infty = 1, \\ & & x \in R, \\ & & p \in C_Y, \\ & & g \ge 0. \end{array} \tag{3.4}$$

Neither (3.3) nor (3.4) are convex problems. However, both (3.3) and (3.4) are of the form (3.1), and so we can invoke Proposition 3.1 and solve (3.3) and (3.4) using

problem (3.2). From Proposition 3.1, we have

$$\begin{array}{rlll} \rho_P(d) = & \min\limits_{i\in\{1,\ldots,m\},j\in\{-1,1\}} & \min\limits_{y,s,v} & \max\{\|A^t y + s\|_1, |b^t y - v|\} \\ & & \text{s.t.} & y_i = j, \\ & & & y \in C_Y^*, \\ & & & (s,v) \in C^*, \end{array} \tag{3.5}$$

and

$$\begin{array}{rlll} \rho_D(d) = & \min\limits_{i\in\{1,\ldots,n\},j\in\{-1,1\}} & \min\limits_{x,p,g} & \max\{\|Ax - p\|_1, |c^t x + g|\} \\ & & \text{s.t.} & x_i = j, \\ & & & x \in R, \\ & & & p \in C_Y, \\ & & & g \geq 0. \end{array} \tag{3.6}$$

Taken together, (3.5) and (3.6) show that we can compute $\rho_P(d)$ by solving $2m$ convex optimization problems, and we can compute $\rho_D(d)$ by solving $2n$ convex optimization problems. In conclusion, we can compute $\rho(d)$ by solving $2n + 2m$ convex optimization problems, where all of the optimization problems involved are of roughly the same size as the original problem (GP_d).

Of course, each of the $2n + 2m$ convex problems in (3.5) and (3.6) will be computationally tractable only if we can conveniently work with the cones involved; we now show that for the special case of linear optimization models (2.1) there are convenient linear inequality characterizations of all of the cones involved in (3.5) and (3.6). The cone C_Y is easily seen to be

$$C_Y = \{p \in \mathbb{R}^m \mid p_i \leq 0 \text{ for } i \in L, p_i = 0 \text{ for } i \in E, p_i \geq 0 \text{ for } i \in G\}, \tag{3.7}$$

and so

$$C_Y^* = \{y \in \mathbb{R}^m \mid y_i \leq 0 \text{ for } i \in L, y_i \in \mathbb{R} \text{ for } i \in E, y_i \geq 0 \text{ for } i \in G\}. \tag{3.8}$$

With the ground set P defined in (2.3), we have

$$R = \{x \in \mathbb{R}^n \mid x_j \geq 0 \text{ for } j \in L_B,\ x_j \leq 0 \text{ for } j \in U_B\} \tag{3.9}$$

and also

$$C = \{(x,t) \in \mathbb{R}^n \times \mathbb{R} \mid t \geq 0, x_j \geq l_j t \text{ for } j \in L_B,\ x_j \leq u_j t \text{ for } j \in U_B\}. \tag{3.10}$$

The only cone whose characterization is less than obvious is C^*, which we now characterize. Consider the following system of linear inequalities in the variables $(s, v, s^+, s^-) \in \mathbb{R}^n \times \mathbb{R} \times \mathbb{R}^n \times \mathbb{R}^n$:

$$\begin{array}{rcll} s - s^+ + s^- & = & 0, & \\ s^+ & \geq & 0, & \\ s^- & \geq & 0, & \\ s_j^- & = & 0 & \text{for } j \in N \setminus U_B, \\ s_j^+ & = & 0 & \text{for } j \in N \setminus L_B, \\ v + \sum\limits_{j\in L_B} l_j s_j^+ - \sum\limits_{j\in U_B} u_j s_j^- & \geq & 0, & \end{array} \tag{3.11}$$

where we use the notation $N := \{1, \ldots, n\}$ and $S \setminus T$ is the set difference $\{k \mid k \in S, k \notin T\}$.

PROPOSITION 3.3. *For the ground set P defined in* (2.3)*, the cone C^* is characterized by*

$$C^* = \{(s, v) \in \mathbb{R}^n \times \mathbb{R} \mid (s, v, s^+, s^-) \text{ satisfies (3.11) for some } s^+, s^- \in \mathbb{R}^n\}.$$

Proof. Suppose first that (s, v) together with some s^+, s^- satisfies (3.11). Then for all $(x, t) \in C$ we have

$$\begin{aligned} (x,t)^t(s,v) &= \sum_{j \in L_B} s_j^+ x_j - \sum_{j \in U_B} s_j^- x_j + tv \\ &\geq \sum_{j \in L_B} s_j^+ l_j t - \sum_{j \in U_B} s_j^- u_j t + tv \qquad (3.12) \\ &\geq 0, \end{aligned}$$

and so $(s, v) \in C^*$. Conversely, suppose that $(s, v) \in C^*$. Then

$$\begin{aligned} -\infty < -v \leq \min_{x \in P} s^t x = \min \quad & \sum_{j=1}^{n} s_j x_j \\ \text{s.t.} \quad & x_j \geq l_j \text{ for } j \in L_B, \qquad (3.13) \\ & x_j \leq u_j \text{ for } j \in U_B, \end{aligned}$$

and define s^+ and s^- to be the positive and negative parts of s, respectively. Then $s = s^+ - s^-$, $s^+ \geq 0$, and $s^- \geq 0$, and (3.13) implies $s_j^+ = 0$ for $j \in N \setminus L_B$, $s_j^- = 0$ for $j \in N \setminus U_B$, as well as the last inequality of (3.11), whereby (s, v, s^+, s^-) satisfies all inequalities of (3.11). ∎

Taken together, we can use (3.7), (3.8), (3.9), (3.10), and Proposition 3.3 to rewrite the right-most minimization problems of (3.5) and (3.6) and obtain

$$\begin{aligned} \rho_P(d) = \min_{\substack{i \in \{1,\ldots,m\} \\ j \in \{-1,1\}}} \min_{y, s^+, s^-, v} \quad & \max\{\|A^t y + s^+ - s^-\|_1, |b^t y - v|\} \\ \text{s.t.} \quad & y_i = j, \\ & y_l \leq 0 \quad \text{for } l \in L, \\ & y_l \geq 0 \quad \text{for } l \in G, \\ & s_k^- = 0 \quad \text{for } k \in N \setminus U_B, \qquad (3.14) \\ & s_k^+ = 0 \quad \text{for } k \in N \setminus L_B, \\ & v + \sum_{k \in L_B} l_k s_k^+ - \sum_{k \in U_B} u_k s_k^- \geq 0, \\ & s^+, s^- \geq 0, \end{aligned}$$

and

$$\begin{aligned} \rho_D(d) = \min_{\substack{i \in \{1,\ldots,n\} \\ j \in \{-1,1\}}} \min_{x, p, g} \quad & \max\{\|Ax - p\|_1, |c^t x + g|\} \\ \text{s.t.} \quad & x_i = j, \\ & x_k \geq 0 \quad \text{if } k \in L_B, \\ & x_k \leq 0 \quad \text{for } k \in U_B, \qquad (3.15) \\ & p_l \leq 0 \quad \text{for } l \in L, \\ & p_l = 0 \quad \text{for } l \in E, \\ & p_l \geq 0 \quad \text{for } l \in G, \\ & g \geq 0, \end{aligned}$$

whose right-most objective functions can then easily be converted to linear optimization problems by standard techniques. This then shows that we can indeed compute $\rho_P(d)$, $\rho_D(d)$, and $\rho(d)$ by solving $2n+2m$ linear programs, under the choice of norms given in Assumption 1.

3.2. Computing $\|d\|$. In order to compute the condition measure $C(d) := \|d\|/\rho(d)$, we must also compute $\|d\| = \max\{\|A\|, \|b\|, \|c\|_*\}$. Under Assumption 1 we have $\|b\| = \|b\|_1$ and $\|c\|_* = \|c\|_1$, which are both easy to compute. However, $\|A\|$ is the operator norm, and so $\|A\| := \|A\|_{\infty,1} := \max\{\|Ax\|_1 \mid \|x\|_\infty = 1\}$, whose computation is NP-hard (one can easily pose MAXCUT as a special case). We therefore will bound $\|A\|_{\infty,1}$ and hence $\|d\|$ from below and above, using the following elementary norm inequalities:

$$\max\{\|A\|_{1,1}, \|A\|_{2,2}, \|A\|_F, \|Ae\|_1, \|A\hat{x}\|_1\} \leq \|A\|_{\infty,1} \leq \min\{\|A\|_{L_1}, \sqrt{nm}\|A\|_{2,2}\},$$

where

$$\begin{aligned}
\|A\|_{1,1} &= \max_{j=1,\ldots,n} \|A_{\bullet j}\|_1, \\
\|A\|_{2,2} &= \sqrt{\lambda_{\max}(A^t A)}, \\
\|A\|_F &= \sqrt{\textstyle\sum_{i=1}^m \sum_{j=1}^n (A_{i,j})^2}, \\
\|A\|_{L_1} &= \textstyle\sum_{i=1}^m \sum_{j=1}^n |A_{i,j}|,
\end{aligned}$$

$e := (1, \ldots, 1)^t$, and $\hat{x}$ is defined using $\hat{x}_j = \text{sign}(A_{i^*,j})$, where $i^* = \text{argmax}_{i=1,\ldots,m}\|A_{i\bullet}\|_1$.

4. Computational results on the NETLIB suite of linear optimization problems.

4.1. Condition measures for the NETLIB suite prior to preprocessing. We chose the NETLIB suite of linear optimization problem instances as a representative suite of LP problems encountered in practice, and we computed the condition measures $\rho_P(d), \rho_D(d)$, and $C(d)$ for problem instances in this suite using the methodology developed in section 3. The NETLIB suite is comprised of 98 linear optimization problems from diverse application areas, collected over a period of many years. While this suite does not contain any truly large problems by today's standards, it is arguably the best publicly available collection of practical LP problems, and the sizes and diversity of the problems contained therein seem to be representative of general practice. The sizes of the problem instances in the NETLIB suite range from 32 variables and 28 constraints to problems with roughly 9,000 variables and 3,000 constraints. 44 of the 98 problems in the suite have nonzero lower bound constraints and/or upper bound constraints on the variables, and five problems have range constraints. We omitted the five problems with range constraints (boeing1, boeing2, forplan, nesm, seba) for the purposes of our analysis (range constraints do not naturally fit into either the conic model or the ground-set model format). On four of the remaining problems (dfl001, qap12, qap15, stocfor3) our methodology has not yet exhibited convergence to a solution, and these four problems were omitted as well, yielding a final sample set of 89 linear optimization problems. The burden of computing the distances to ill-posedness for the NETLIB suite via the solution of $2n + 2m$ linear programs obviously grows with the dimensions of the problem instances. On `afiro`, which is a small problem instance ($n = 28$, $m = 32$), the total computation time amounted to only 0.28 seconds of machine time, whereas for `maros-r7` ($n = 9,408$ and $m = 3,136$), the total computation time was $240,627.59$ seconds of machine time (66.84 hours).

Table 4.1 shows the distances to ill-posedness and the condition measure estimates for the 89 problems, using the methodology for computing $\rho_P(d)$ and $\rho_D(d)$ and for estimating $\|d\|$ presented in section 3. All LP computation was performed using CPLEX 7.1 (function *primopt*).

Table 4.2 presents some summary statistics of the condition measure computations from Table 4.1. As the table shows, 71% (63/89) of the problems in the NETLIB suite are ill-posed due to either $\rho_P(d) = 0$ or $\rho_D(d) = 0$ or both. Furthermore, notice that, among these 63 ill-posed problems, almost all (61 out of 63) have $\rho_P(d) = 0$. This means that for 69% (61/89) of the problems in the NETLIB suite, arbitrarily small changes in the data will render the primal problem infeasible.

Notice from Table 4.1 that there are three problems for which $\rho_D(d) = \infty$, namely, `fit1d`, `fit2d`, and `sierra`. This can happen only when the ground set P is bounded, which for linear optimization means that all variables have finite lower and upper bounds.

The computational results in Tables 4.1 and 4.2 have shown that 61 of the 89 linear programs in the NETLIB suite are primal ill-posed, i.e., $\rho_P(d) = 0$, and so arbitrarily small changes in the data will render the primal problem infeasible. For feasible linear programs, $\rho_P(d) = 0$ can happen only if (i) there are linear dependencies among the equations of the problem instance (2.1), or (ii) there is an implied reverse inequality among the inequalities and lower and upper bounds of the problem instance. Furthermore, it is easy to show that if $s = 0$ in an optimal solution of (3.3), then there are linear dependencies in the equations (and possibly implied reverse inequalities as well), whereas if $s \neq 0$ in an optimal solution of (3.3), then there is an implied reverse inequality (and possibly linear dependencies as well). This then can be used to evaluate the causes of the ill-posedness of the 61 primal ill-posed instances. We examined the optimal solutions of (3.3) for the 61 primal ill-posed linear programs in the NETLIB suite in order to evaluate the causes of the ill-posedness among these problems. Table 4.3 summarizes our findings, which show that for at least 34% of the primal ill-posed problem instances there are linear dependencies among the equations of (2.1).

4.2. Condition measures for the NETLIB suite after preprocessing. Most commercial software packages for solving linear optimization problems perform preprocessing heuristics prior to solving a problem instance. These heuristics typically include checks for eliminating linearly dependent equations, heuristics for identifying and eliminating redundant variable lower and upper bounds, and rules for row and/or column rescaling, etc. The purposes of the preprocessing are to reduce the size of the problem instance by eliminating dependent equations and redundant inequalities, and to improve numerical computation and enhance iteration performance by rescaling of rows and/or columns. The original problem instance is converted to a postprocessed instance by the processing heuristics, and it is this postprocessed instance that is used as input to solution software. In CPLEX 7.1, the postprocessed problem can be accessed using function *prslvwrite*. This function writes the postprocessed problem to disk, whence it can be read.

In order to examine condition measures of the problems that are the actual input to a modern IPM solver, we computed condition measures for the NETLIB suite problems after preprocessing by CPLEX 7.1. The processing used was the default CPLEX preprocessing with the linear dependency check option activated. Table 4.4 shows the condition measures in detail for the postprocessed versions of the problems, and Table 4.5 presents some summary statistics of these condition measures. Notice

TABLE 4.1

Condition measures for the NETLIB suite LP problem instances (prior to preprocessing by CPLEX 7.1).

			$\|d\|$		$\log C(d)$	
Problem	$\rho_P(d)$	$\rho_D(d)$	Lower bound	Upper bound	Lower bound	Upper bound
25fv47	0.000000	0.000000	30,778	55,056	∞	∞
80bau3b	0.000000	0.000000	142,228	142,228	∞	∞
adlittle	0.000000	0.051651	68,721	68,721	∞	∞
afiro	0.397390	1.000000	1,814	1,814	3.7	3.7
agg	0.000000	0.771400	5.51E+07	5.51E+07	∞	∞
agg2	0.000000	0.771400	1.73E+07	1.73E+07	∞	∞
agg3	0.000000	0.771400	1.72E+07	1.72E+07	∞	∞
bandm	0.000000	0.000418	10,200	17,367	∞	∞
beaconfd	0.000000	0.000000	15,322	19,330	∞	∞
blend	0.003541	0.040726	1,020	1,255	5.5	5.5
bnl1	0.000000	0.106400	8,386	9,887	∞	∞
bnl2	0.000000	0.000000	36,729	36,729	∞	∞
bore3d	0.000000	0.003539	11,912	12,284	∞	∞
brandy	0.000000	0.000000	7,254	10,936	∞	∞
capri	0.000252	0.095510	33,326	33,326	8.1	8.1
cycle	0.000000	0.000000	365,572	391,214	∞	∞
czprob	0.000000	0.008807	328,374	328,374	∞	∞
d2q06c	0.000000	0.000000	171,033	381,438	∞	∞
d6cube	0.000000	2.000000	47,258	65,574	∞	∞
degen2	0.000000	1.000000	3,737	3,978	∞	∞
degen3	0.000000	1.000000	4,016	24,646	∞	∞
e226	0.000000	0.000000	22,743	37,344	∞	∞
etamacro	0.000000	0.200000	31,249	63,473	∞	∞
fffff800	0.000000	0.033046	1.55E+06	1.55E+06	∞	∞
finnis	0.000000	0.000000	31,978	31,978	∞	∞
fit1d	3.500000	∞	493,023	618,065	5.1	5.2
fit1p	1.271887	0.437500	218,080	384,121	5.7	5.9
fit2d	317.000000	∞	1.90E+06	2.25E+06	3.8	3.9
fit2p	1.057333	1.000000	621,470	658,700	5.8	5.8
ganges	0.000000	1.000000	1.29E+06	1.29E+06	∞	∞
gfrd-pnc	0.000000	0.347032	1.63E+07	1.63E+07	∞	∞
greenbea	0.000000	0.000000	21,295	26,452	∞	∞
greenbeb	0.000000	0.000000	21,295	26,452	∞	∞
grow15	0.572842	0.968073	209	977	2.6	3.2
grow22	0.572842	0.968073	303	1,443	2.7	3.4
grow7	0.572842	0.968073	102	445	2.3	2.9
israel	0.027248	0.166850	2.22E+06	2.22E+06	7.9	7.9
kb2	0.000201	0.018802	10,999	11,544	7.7	7.8
lotfi	0.000306	0.000000	166,757	166,757	∞	∞
maros	0.000000	0.000000	2.51E+06	2.55E+06	∞	∞
maros-r7	1.000000	0.628096	1.02E+07	1.02E+07	7.2	7.2
modszk1	0.000000	0.108469	1.03E+06	1.03E+06	∞	∞
perold	0.000000	0.000943	703,824	2.64E+06	∞	∞
pilot	0.000000	0.000290	26,633	30,427	∞	∞
pilot.ja	0.000000	0.000750	2.67E+07	1.40E+08	∞	∞
pilot.we	0.000000	0.044874	5.71E+06	5.71E+06	∞	∞
pilot4	0.000000	0.000075	763,677	1.09E+06	∞	∞
pilot87	0.000000	0.000000	111,163	138,736	∞	∞
pilotnov	0.000000	0.000750	2.36E+07	1.35E+08	∞	∞
qap8	0.000000	4.000000	17,248	17,248	∞	∞
recipe	0.000000	0.000000	14,881	19,445	∞	∞
sc105	0.000000	0.133484	3,000	3,000	∞	∞
sc205	0.000000	0.010023	5,700	5,700	∞	∞
sc50a	0.000000	0.562500	1,500	1,500	∞	∞
sc50b	0.000000	0.421875	1,500	1,500	∞	∞

TABLE 4.1 (*cont.*)

Problem	$\rho_P(d)$	$\rho_D(d)$	$\|d\|$ Lower bound	$\|d\|$ Upper bound	$\log C(d)$ Lower bound	$\log C(d)$ Upper bound
scagr25	0.021077	0.034646	430,977	430,977	7.3	7.3
scagr7	0.022644	0.034646	120,177	120,177	6.7	6.7
scfxm1	0.000000	0.000000	21,425	22,816	∞	∞
scfxm2	0.000000	0.000000	44,153	45,638	∞	∞
scfxm3	0.000000	0.000000	66,882	68,459	∞	∞
scorpion	0.000000	0.949393	5,622	5,622	∞	∞
scrs8	0.000000	0.000000	68,630	69,449	∞	∞
scsd1	5.037757	1.000000	1,752	1,752	3.2	3.2
scsd6	1.603351	1.000000	2,973	2,973	3.5	3.5
scsd8	0.268363	1.000000	5,549	5,549	4.3	4.3
sctap1	0.032258	1.000000	8,240	17,042	5.4	5.7
sctap2	0.586563	1.000000	32,982	72,870	4.7	5.1
sctap3	0.381250	1.000000	38,637	87,615	5.0	5.4
share1b	0.000015	0.000751	60,851	87,988	9.6	9.8
share2b	0.001747	0.287893	19,413	23,885	7.0	7.1
shell	0.000000	1.777778	253,434	253,434	∞	∞
ship04l	0.000000	13.146000	811,956	811,956	∞	∞
ship04s	0.000000	13.146000	515,186	515,186	∞	∞
ship08l	0.000000	21.210000	1.91E+06	1.91E+06	∞	∞
ship08s	0.000000	21.210000	1.05E+06	1.05E+06	∞	∞
ship12l	0.000000	7.434000	794,932	794,932	∞	∞
ship12s	0.000000	7.434000	381,506	381,506	∞	∞
sierra	0.000000	∞	6.60E+06	6.61E+06	∞	∞
stair	0.000580	0.000000	976	1,679	∞	∞
standata	0.000000	1.000000	21,428	23,176	∞	∞
standgub	0.000000	0.000000	21,487	23,235	∞	∞
standmps	0.000000	1.000000	22,074	23,824	∞	∞
stocfor1	0.001203	0.011936	23,212	23,441	7.3	7.3
stocfor2	0.000437	0.000064	462,821	467,413	9.9	9.9
truss	0.518928	10.000000	154,676	154,676	5.5	5.5
tuff	0.000000	0.017485	136,770	145,448	∞	∞
vtp.base	0.000000	0.500000	530,416	534,652	∞	∞
wood1p	0.000000	1.000000	3.66E+06	5.04E+06	∞	∞
woodw	0.000000	1.000000	9.86E+06	1.35E+07	∞	∞

TABLE 4.2

Summary statistics of distances to ill-posedness for the NETLIB suite (prior to preprocessing by CPLEX 7.1*).*

		$\rho_D(d)$ 0	$\rho_D(d)$ Finite	$\rho_D(d)$ ∞	Totals
$\rho_P(d)$	0	19	41	1	61
	Finite	2	24	2	28
	∞	0	0	0	0
	Totals	21	65	3	89

TABLE 4.3

Evaluation of ill-posedness of the 61 *primal ill-posed instances in the NETLIB suite (prior to preprocessing by CPLEX* 7.1*).*

Indication	Number of instances
Dependent equations	21
Implied reverse inequalities	40
Total	61

TABLE 4.4
Condition measures for the NETLIB suite after preprocessing by CPLEX 7.1.

			$\|d\|$		$\log C(d)$	
Problem	$\rho_P(d)$	$\rho_D(d)$	Lower bound	Upper bound	Lower bound	Upper bound
25fv47	0.000707	0.000111	35,101	54,700	8.5	8.7
80bau3b	0.000000	0.000058	126,355	126,355	∞	∞
adlittle	0.004202	1.000488	68,627	68,627	7.2	7.2
afiro	0.397390	1.000000	424	424	3.0	3.0
agg	0.000000	0.031728	3.04E+07	3.04E+07	∞	∞
agg2	0.000643	1.005710	1.57E+07	1.57E+07	10.4	10.4
agg3	0.000687	1.005734	1.56E+07	1.56E+07	10.4	10.4
bandm	0.001716	0.000418	7,283	12,364	7.2	7.5
beaconfd	0.004222	1.000000	6,632	6,632	6.2	6.2
blend	0.011327	0.041390	872	1,052	4.9	5.0
bnl1	0.000016	0.159015	8,140	9,544	8.7	8.8
bnl2	0.000021	0.000088	18,421	20,843	8.9	9.0
bore3d	0.000180	0.012354	8,306	8,306	7.7	7.7
brandy	0.000342	0.364322	4,342	7,553	7.1	7.3
capri	0.000375	0.314398	30,323	30,323	7.9	7.9
cycle	0.000021	0.009666	309,894	336,316	10.2	10.2
czprob	0.000000	0.001570	206,138	206,138	∞	∞
d2q06c	0.000000	0.003925	172,131	378,209	∞	∞
d6cube	0.945491	2.000000	43,629	60,623	4.7	4.8
degen2	0.000000	1.000000	2,613	3,839	∞	∞
degen3	0.000000	1.000000	4,526	24,090	∞	∞
e226	0.000737	0.021294	21,673	35,518	7.5	7.7
etamacro	0.001292	0.200000	55,527	87,767	7.6	7.8
fffff800	0.000000	0.033046	696,788	696,788	∞	∞
finnis	0.000000	0.000000	74,386	74,386	∞	∞
fit1d	3.500000	∞	493,023	617,867	5.1	5.2
fit1p	1.389864	1.000000	218,242	383,871	5.3	5.6
fit2d	317.000000	∞	1.90E+06	2.24E+06	3.8	3.8
fit2p	1.057333	1.000000	621,470	658,700	5.8	5.8
ganges	0.000310	1.000000	143,913	143,913	8.7	8.7
gfrd-pnc	0.015645	0.347032	1.22E+07	1.22E+07	8.9	8.9
greenbea	0.000033	0.000004	65,526	65,526	10.2	10.2
greenbeb	0.000034	0.000007	43,820	43,820	9.8	9.8
grow15	0.572842	0.968073	209	977	2.6	3.2
grow22	0.572842	0.968073	303	1,443	2.7	3.4
grow7	0.572842	0.968073	102	445	2.3	2.9
israel	0.135433	0.166846	2.22E+06	2.22E+06	7.2	7.2
kb2	0.000201	0.026835	10,914	11,054	7.7	7.7
lotfi	0.000849	0.001590	170,422	170,422	8.3	8.3
maros	0.000000	0.006534	1.76E+06	1.80E+06	∞	∞
maros-r7	1.000131	0.846743	9.39E+06	9.39E+06	7.0	7.0
modszk1	0.016030	0.114866	1.03E+06	1.03E+06	7.8	7.8
perold	0.000000	0.002212	1.56E+06	2.35E+06	∞	∞
pilot	0.000002	0.000290	35,379	35,379	10.2	10.2
pilot.ja	0.000000	0.001100	2.36E+07	1.36E+08	∞	∞
pilot.we	0.000000	0.044874	5.71E+06	5.71E+06	∞	∞
pilot4	0.000399	0.002600	696,761	1.03E+06	9.2	9.4
pilot87	0.000000	0.000199	100,187	125,426	∞	∞
pilotnov	0.000000	0.001146	2.36E+07	1.32E+08	∞	∞
qap8	0.022222	2.000000	17,248	17,248	5.9	5.9
recipe	0.063414	0.000000	13,356	15,815	∞	∞
sc105	0.778739	0.400452	3,000	3,000	3.9	3.9
sc205	0.778739	0.030068	5,700	5,700	5.3	5.3
sc50a	0.780744	1.000000	1,500	1,500	3.3	3.3
sc50b	0.695364	1.000000	1,500	1,500	3.3	3.3

TABLE 4.4 (*cont.*)

Problem	$\rho_P(d)$	$\rho_D(d)$	$\|d\|$ Lower bound	$\|d\|$ Upper bound	$\log C(d)$ Lower bound	$\log C(d)$ Upper bound
scagr25	0.021191	0.049075	199,859	199,859	7.0	7.0
scagr7	0.022786	0.049075	61,259	61,259	6.4	6.4
scfxm1	0.000010	0.002439	20,426	21,811	9.3	9.3
scfxm2	0.000010	0.002439	38,863	43,630	9.6	9.6
scfxm3	0.000010	0.002439	57,300	65,449	9.8	9.8
scorpion	0.059731	0.995879	123,769	123,769	6.3	6.3
scrs8	0.009005	0.004389	66,362	68,659	7.2	7.2
scsd1	5.037757	1.000000	1,752	1,752	3.2	3.2
scsd6	1.603351	1.000000	2,973	2,973	3.5	3.5
scsd8	0.268363	1.000000	5,549	5,549	4.3	4.3
sctap1	0.032258	1.000000	7,204	15,186	5.3	5.7
sctap2	0.669540	1.000000	27,738	64,662	4.6	5.0
sctap3	0.500000	1.000000	32,697	78,415	4.8	5.2
share1b	0.000015	0.000751	1.67E+06	1.67E+06	11.0	11.0
share2b	0.001747	0.287893	19,410	23,882	7.0	7.1
shell	0.000263	0.253968	874,800	874,800	9.5	9.5
ship04l	0.000386	25.746000	881,005	881,005	9.4	9.4
ship04s	0.000557	25.746000	545,306	545,306	9.0	9.0
ship08l	0.000000	22.890000	1.57E+06	1.57E+06	∞	∞
ship08s	0.000000	22.890000	816,531	816,531	∞	∞
ship12l	0.000124	7.434000	748,238	748,238	9.8	9.8
ship12s	0.000149	7.434000	340,238	340,238	9.4	9.4
sierra	0.001039	47.190000	6.60E+06	6.61E+06	9.8	9.8
stair	0.003800	0.163162	7,071	7,071	6.3	6.3
standata	0.090909	1.000000	4,931	5,368	4.7	4.8
standgub	0.090909	1.000000	4,931	5,368	4.7	4.8
standmps	0.020000	1.000000	12,831	12,831	5.8	5.8
stocfor1	0.002130	0.109062	10,833	29,388	6.7	7.1
stocfor2	0.000811	0.000141	45,458	616,980	8.5	9.6
truss	0.518928	10.000000	154,676	154,676	5.5	5.5
tuff	0.000025	0.047081	131,554	138,783	9.7	9.7
vtp.base	0.005287	3.698630	17,606	17,606	6.5	6.5
wood1p	0.059008	1.442564	2.11E+06	3.25E+06	7.6	7.7
woodw	0.009357	1.000000	5.68E+06	7.26E+06	8.8	8.9

TABLE 4.5

Summary statistics of distances to ill-posedness for the NETLIB suite after preprocessing by CPLEX 7.1.

		$\rho_D(d)$ 0	$\rho_D(d)$ Finite	$\rho_D(d)$ ∞	Totals
$\rho_P(d)$	0	1	15	0	16
	Finite	1	70	2	73
	∞	0	0	0	0
	Totals	2	85	2	89

from Table 4.5 that 19% (17/89) of the postprocessed problems in the NETLIB suite are ill-posed. In contrast to the original problems, the vast majority of postprocessed problems have finite condition measures, as the preprocessing heuristics are very effective at identifying and correcting many instances of implied reverse inequalities in addition to finding and eliminating linearly dependent equations. We also examined the optimal solutions of (3.3) for the 16 primal ill-posed postprocessed problems in the NETLIB suite in order to evaluate the causes of the ill-posedness among these

TABLE 4.6
Evaluation of ill-posedness of the 16 *primal ill-posed instances in the NETLIB suite after preprocessing by CPLEX* 7.1.

Indication	Number of instances
Dependent equations	0
Implied reverse inequalities	16
Total	16

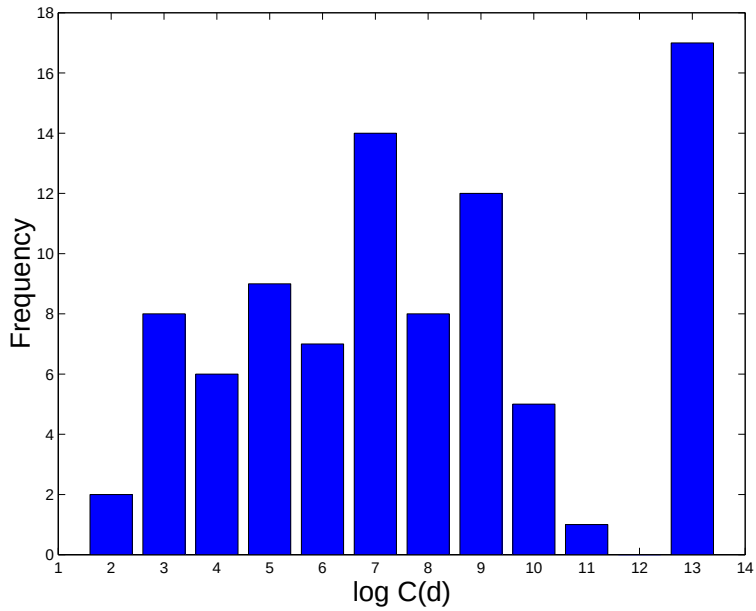


FIG. 4.1. *Histogram of condition measures for the NETLIB suite after preprocessing by CPLEX* 7.1 *(using the geometric mean of the lower and upper bound estimates of* $C(d)$*).*

postprocessed problem instances. Table 4.6 summarizes our findings, which show that all of the ill-posed postprocessed LP instances have implied reverse inequalities among the inequalities and/or lower/upper bounds.

Figure 4.1 presents a histogram of the condition measures of the postprocessed problems taken from Table 4.4. The condition measure of each problem is represented by the geometric mean of the upper and lower bound estimates in this histogram. The right-most column in the figure is used to tally the number of problems for which $C(d) = \infty$, and is shown to give a more complete picture of the data. This histogram shows that of the problems with finite condition measure, $\log C(d)$ is fairly nicely distributed between 2.6 and 11.0. Of course, when $C(d) = 10^{11}$, it is increasingly difficult to distinguish between a finite and nonfinite condition measure.

4.3. Condition measures and the observed performance of interior-point methods on the NETLIB suite. In the case of modern IPM algorithms for linear optimization, the number of IPM iterations needed to solve a linear optimization instance has been observed to be fairly constant over a huge range of problem sizes; for the NETLIB suite the number of iterations varies between 8 and 48 using CPLEX 7.1 *baropt*; for other codes the numbers are a bit different. Extensive computational experience over the past 15 years has shown that the IPM iterations needed

to solve a linear optimization problem instance vary in the range between 10–100 iterations. There is some evidence that the number of IPM iterations grows roughly as $\log n$ on a particular class of structured problem instances; see, for example, [12].

The observed performance of modern IPM algorithms is fortunately superior to the worst-case bounds on IPM iterations that arise via theoretical complexity analysis. Depending on the complexity model used, one can bound the number of IPM iterations from above by $\sqrt{\vartheta}\tilde{L}$, where ϑ is the number of inequalities plus the number of variables with at least one bound in the problem instance,

$$\vartheta := |L| + |G| + |L_B| + |U_B| - |L_B \cap U_B|, \tag{4.1}$$

and $\tilde{L}$ is the bit-size of a binary encoding of the problem instance data; see [23]. (Subtraction of the final term of (4.1) is shown in [7].) The bit-size model was a motivating force for modern polynomial-time LP algorithms, but is viewed today as somewhat outdated in the context of linear and nonlinear optimization. Using instead the condition-measure model for complexity analysis, one can bound the IPM iterations by $O(\sqrt{\vartheta}\log(C(d)+\cdots))$, where the other terms in the bound are of a more technical nature; see [25] for details. Of course, even here one must bear in mind that the IPM algorithms that are used in practice are different from the IPM algorithms that are used in the development of the complexity theory.

A natural question to ask is whether the observed variation in the number of IPM iterations (already small) can be accounted for by the condition measures of the problem instances that are the input to the IPM algorithm. The finite condition measures of the 72 postprocessed problems from the NETLIB suite shown in Table 4.4 provide a rich set of data that can be used to address this question. Here the goal is to assess whether or not condition measures are relevant for understanding the practical performance of IPM algorithms (we do *not* aim at validating the complexity theory).

In order to assess any relationship between condition measures and IPM iterations for the NETLIB suite, we first solved and recorded the IPM iterations for the 89 problems from the NETLIB suite. The problems were preprocessed with the linear dependency check option and solved with CPLEX 7.1 function *baropt* with default parameters. The default settings use the standard barrier algorithm, include a starting heuristic that sets the initial dual solution to zero, and a convergence criteria of a relative complementarity smaller than 10^{-8}. The iteration counts are shown in Table 4.7. Notice that these iteration counts vary between 8 and 48.

Figure 4.2 shows a scatter plot of the number of IPM iterations taken by CPLEX 7.1 to solve the 89 problems in the NETLIB suite after preprocessing (from Table 4.7) and $\log C(d)$ of the postprocessed problems (using the $\log C(d)$ estimates from columns 6 and 7 of Table 4.4). In the figure, the horizontal lines represent the range for $\log C(d)$ due to the lower and upper estimates of $C(d)$ from the last two columns of Table 4.4. Also, similarly to Figure 4.1, problems with infinite condition measure are shown in the figure on the far right as a visual aid.

Figure 4.2 shows that as $\log C(d)$ increases, so does the number of IPM iterations needed to solve the problem (with exceptions, of course). Perhaps a more accurate summary of the figure is that if the number of IPM iterations is large, then the problem will tend to have a large value of $\log C(d)$. The converse of this statement is not supported by the scatter plot: if a problem has a large value of $\log C(d)$, one cannot state in general that the problem will take a large number of IPM iterations to solve.

TABLE 4.7
IPM iterations for the NETLIB suite using CPLEX 7.1 *function* baropt.

Problem	IPM iterations	Problem	IPM iterations	Problem	IPM iterations
25fv47	22	gfrd-pnc	18	scorpion	13
80bau3b	30	greenbea	38	scrs8	20
adlittle	12	greenbeb	33	scsd1	10
afiro	9	grow15	12	scsd6	11
agg	22	grow22	12	scsd8	9
agg2	18	grow7	10	sctap1	13
agg3	21	israel	23	sctap2	15
bandm	16	kb2	17	sctap3	15
beaconfd	8	lotfi	14	share1b	22
blend	11	maros	27	share2b	14
bnl1	25	maros-r7	9	shell	16
bnl2	28	modszk1	23	ship04l	13
bore3d	16	perold	42	ship04s	17
brandy	19	pilot	22	ship08l	14
capri	19	pilot.ja	46	ship08s	14
cycle	25	pilot.we	48	ship12l	19
czprob	32	pilot4	35	ship12s	17
d2q06c	28	pilot87	26	sierra	16
d6cube	22	pilotnov	19	stair	16
degen2	13	qap8	9	standata	9
degen3	19	recipe	9	standgub	9
e226	18	sc105	10	standmps	13
etamacro	24	sc205	11	stocfor1	10
fffff800	30	sc50a	10	stocfor2	16
finnis	19	sc50b	9	truss	17
fit1d	14	scagr25	14	tuff	21
fit1p	13	scagr7	13	vtp.base	10
fit2d	18	scfxm1	18	wood1p	13
fit2p	18	scfxm2	20	woodw	21
ganges	13	scfxm3	20		

In order to be a bit more definitive, we ran a simple linear regression with the IPM iterations of the postprocessed problem as the dependent variable and $\log C(d)$ as the independent variable, for the 72 NETLIB problems which have a finite condition measure after preprocessing. For the purposes of the regression computation we used the geometric mean of the lower and upper estimates of the condition measure from the last two columns of Table 4.4. The resulting linear regression equation is

$$\text{IPM Iterations} = 4.1223 + 1.7490 \log C(d),$$

with $R^2 = 0.4160$. This indicates that in the sample of 72 NETLIB suite problem instances whose postprocessed condition measure is finite, about 42% of the variation in IPM iterations among these problems is accounted for by $\log C(d)$ of the postprocessed problem instance. A plot of this regression line is shown in Figure 4.3, where once again the 17 problems that are ill-posed are shown in the figure on the far right as a visual aid. Both coefficients of this simple linear regression are significant at the 95% confidence level; see the regression statistics shown in Table 4.8.

The above regression analysis indicates that $\log C(d)$ accounts for 42% of the variation in IPM iteration counts among those NETLIB suite problem instances with finite postprocessed condition measure. However, recall that the complexity theory of interior-point methods bounds the number of IPM iterations by $O(\sqrt{\vartheta} \log(C(d)+\cdots))$. The factor $\sqrt{\vartheta}$ in the complexity bound seems to be a fixture of the theory of self-concordant barrier functions (see [14]), despite the belief that such dependence is not

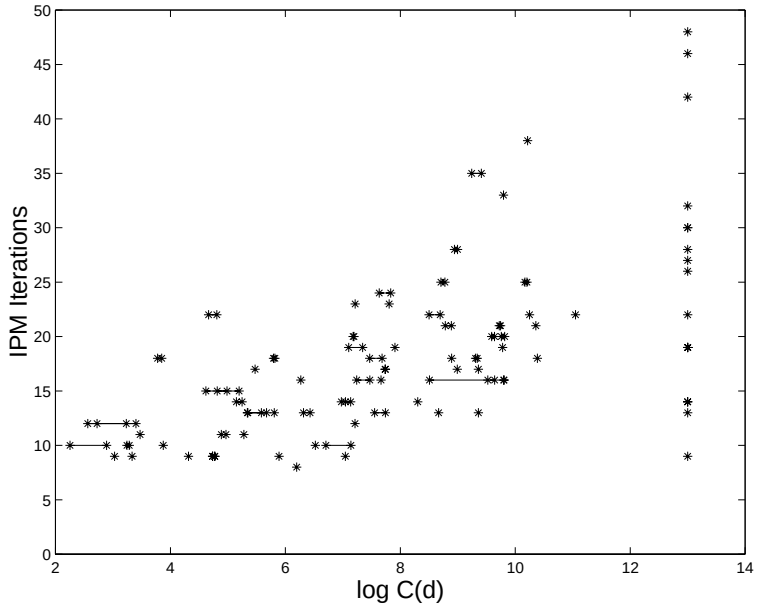


FIG. 4.2. *Scatter plot of IPM iterations and* $\log C(d)$ *for* 89 *NETLIB problems after preprocessing, using CPLEX* 7.1.

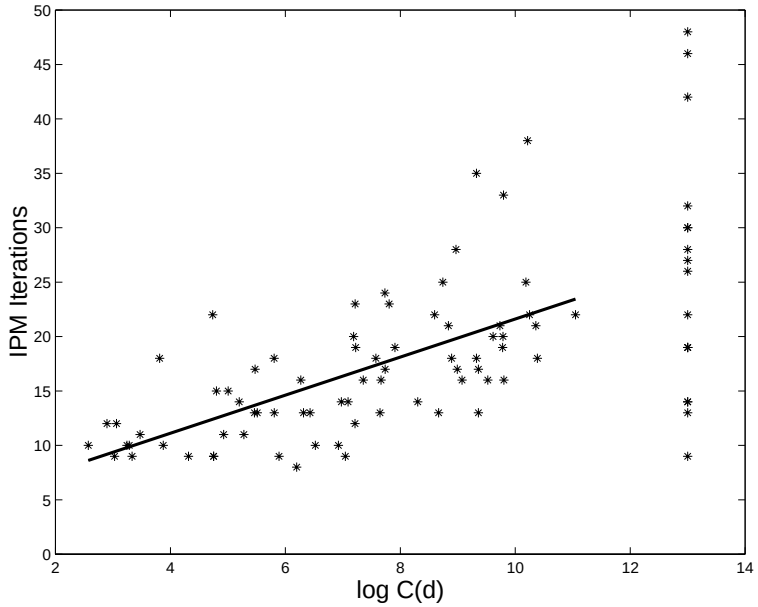


FIG. 4.3. *Linear regression of IPM iterations and* $\log C(d)$ *for* 72 *NETLIB problems with finite condition measure after preprocessing, using CPLEX* 7.1 *(using the geometric mean of the lower and upper bound estimates of* $C(d)$*).*

TABLE 4.8
Statistics for the linear regression of IPM iterations and $\log C(d)$.

Coefficient	Value	t-statistic	95% Confidence interval
β_0	4.1223	2.2480	[0.4650 , 7.7796]
β_1	1.7490	7.0620	[1.2551 , 2.2430]

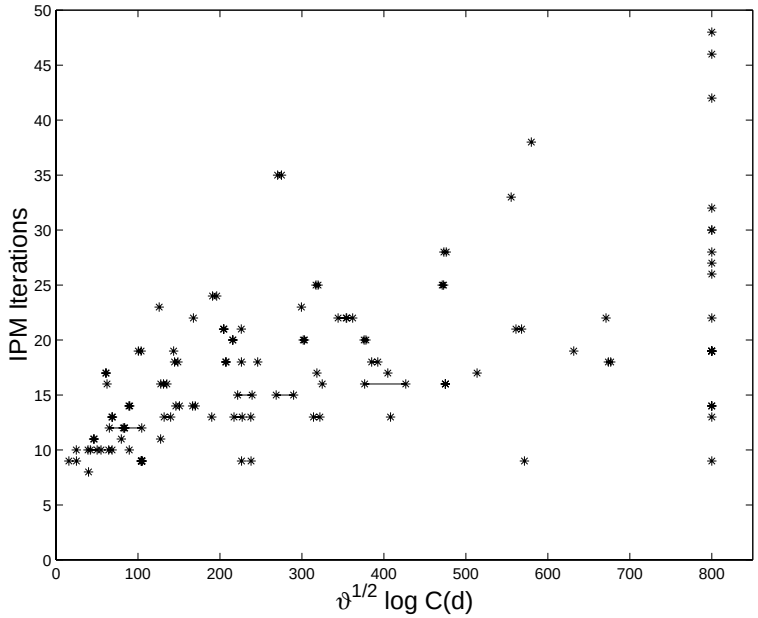


FIG. 4.4. *Scatter plot of IPM iterations and* $\sqrt{\vartheta}\log C(d)$ *for* 89 *NETLIB problems after preprocessing, using CPLEX* 7.1.

borne out in practice. Nevertheless, one can also ask whether $\sqrt{\vartheta}\log C(d)$ as opposed to $\log C(d)$ might better account for the variation in IPM iteration counts among the NETLIB suite problems. We now address this question. Figure 4.4 shows a scatter plot of the number of IPM iterations taken by CPLEX 7.1 to solve the 89 problems in the NETLIB suite after preprocessing and $\sqrt{\vartheta}\log C(d)$ of the postprocessed problems. (The horizontal lines refer to the range of the lower and upper estimates of $C(d)$ from the last two columns of Table 4.4; also, problems with infinite condition measure are shown in the figure on the far right as a visual aid.) We also ran a simple linear regression of IPM iterations of the postprocessed problem as the dependent variable and $\sqrt{\vartheta}\log C(d)$ as the independent variable, again for the 72 NETLIB problems which have a finite condition measure after preprocessing. The resulting linear regression equation is

$$\text{IPM Iterations} = 11.7903 + 0.0195\sqrt{\vartheta}\log C(d),$$

with $R^2 = 0.3021$. A plot of this regression is shown in Figure 4.5, and Table 4.9 shows the regression statistics. Notice that $R^2 = 0.3021$ for the $\sqrt{\vartheta}\log C(d)$ regression model, which is inferior to $R^2 = 0.4160$ for the $\log C(d)$ regression model. These results indicate that among the 72 NETLIB suite postprocessed problem instances with finite condition measure, $\log C(d)$ is better than $\sqrt{\vartheta}\log C(d)$ at accounting for the variation in IPM iterations for these NETLIB suite problems.

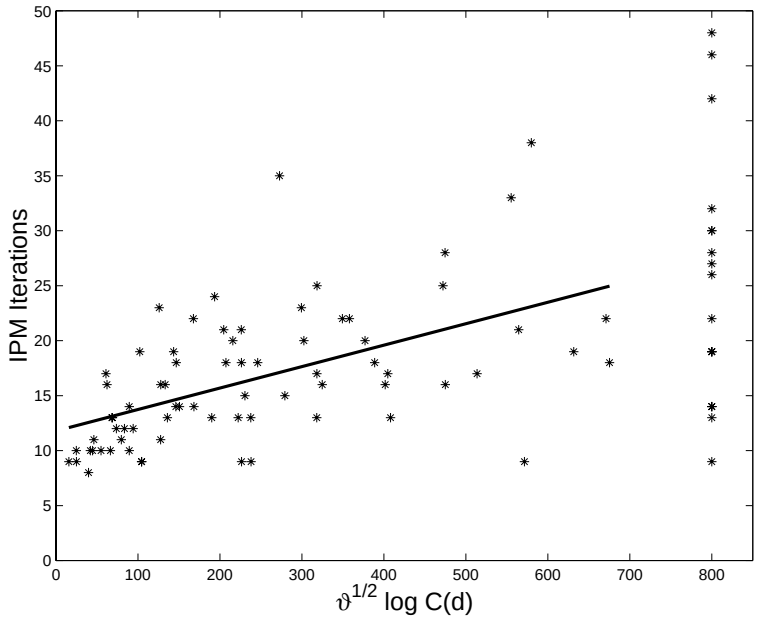


FIG. 4.5. *Linear regression of IPM iterations and* $\sqrt{\vartheta}\log C(d)$ *for* 72 *NETLIB problems with finite condition measure after preprocessing, using CPLEX* 7.1 *(using the geometric mean of the lower and upper bound estimates of* $C(d)$*).*

TABLE 4.9
Statistics for the linear regression of IPM iterations and $\sqrt{\vartheta}\log C(d)$*.*

Coefficient	Value	t-statistic	95% Confidence interval
β_0	11.7903	11.2667	[9.7031 , 13.8774]
β_1	0.0195	5.5046	[0.0124 , 0.0266]

TABLE 4.10
Sample correlations for 72 *NETLIB suite problems after preprocessing by CPLEX* 7.1 *(using the geometric mean of the lower and upper bound estimates of* $C(d)$*).*

	IPM iterations	$\log C(d)$	$\log n$	$\log m$	$\log \vartheta$	$\sqrt{\vartheta}$
IPM iterations	1.000					
$\log C(d)$	0.645	1.000				
$\log n$	0.383	0.217	1.000			
$\log m$	0.432	0.371	0.777	1.000		
$\log \vartheta$	0.398	0.224	0.991	0.808	1.000	
$\sqrt{\vartheta}$	0.311	0.093	0.909	0.669	0.918	1.000

We also computed the sample correlation coefficients of the IPM iterations from Table 4.7 with the following dimensional measures for the 72 problems in the NETLIB suite with finite condition measure of the postprocessed problem instance: $\log m$, $\log n$, $\log \vartheta$, and $\sqrt{\vartheta}$. The resulting sample correlations are shown in Table 4.10. Observe from Table 4.10 that IPM iterations are better correlated with $\log C(d)$ than with any of the other measures. The closest other measure is $\log m$, for which $R = 0.432$, and so a linear regression of IPM iterations as a function of $\log m$ would yield $R^2 = (0.432)^2 = 0.187$, which is decidedly less than $R^2 = 0.4160$ for $\log C(d)$. Also, note from Table 4.10 that both $\log \vartheta$ and $\sqrt{\vartheta}$ by themselves are significantly less correlated with the IPM iterations than $\log C(d)$.

4.4. Controlled perturbations of problems in the NETLIB suite. One potential drawback of the analysis in subsection 4.3 is that in making comparisons of problem instances with different condition measures one necessarily fails to keep the problem size or structure invariant. Herein, we attempt to circumvent this drawback by performing controlled perturbations of linear optimization problems, which allows one to keep the problem size and structure intact.

Consider a problem instance $d = (A, b, c)$ and the computation of the primal and dual distances to ill-posedness $\rho_P(d)$ and $\rho_D(d)$. It is fairly straightforward to show that if $(i^*, j^*, y^*, (s^+)^*, (s^-)^*, v^*)$ is an optimal solution of (3.14), then the rank-1 data perturbation

$$\Delta d = (\Delta A, \Delta b, \Delta c) := \big(-j^* e^{i^*} (A^t y^* + (s^+)^* - (s^-)^*)^t, -j^* e^{i^*} (b^t y^* - v^*), 0\big) \tag{4.2}$$

is a minimum-norm perturbation for which $\rho_P(d + \Delta d) = 0$ (where e^{i^*} denotes the (i^*)th unit vector in $\mathbb{R}^m$). That is, $\|\Delta d\| = \rho_P(d)$, and the data instance $\tilde{d} := d + \Delta d$ is primal ill-posed.

The simple construction shown in (4.2) allows one to construct a controlled perturbation of the data instance d. Consider the family of data instances $d_\alpha := d + \alpha \Delta d$ for $\alpha \in [0, 1]$. Then if $\rho_D(d) \geq \rho_P(d) > 0$, it follows that $\rho(d_\alpha) = (1 - \alpha)\rho(d)$ for $\alpha \in [0, 1]$, and we can bound the condition measure of d_α as follows:

$$C(d_\alpha) = \frac{\|d + \alpha \Delta d\|}{(1 - \alpha)\rho(d)} \geq \frac{\|d\| - \alpha \rho(d)}{(1 - \alpha)\rho(d)},$$

where the numerator satisfies $\|d\| - \alpha\rho(d) \geq 0$ for $\alpha \in [0, 1]$. In the case when $\|d\| > \rho(d)$ (satisfied by all problem instances in the NETLIB suite) we can create a family of data instances for which $C(d_\alpha) \to \infty$ as $\alpha \to 1$ by varying α in the range $[0, 1]$, all the while keeping the problem dimensions, the structure of the cone C_Y, and the ground set P invariant.

To illustrate, consider the problem `scagr25` from the NETLIB suite, and let $\bar{d}$ denote the data for this problem instance after preprocessing. According to Table 4.4, $\rho_D(\bar{d}) = 0.049075 \geq 0.021191 = \rho_P(\bar{d}) > 0$. Now let $\Delta\bar{d}$ be the perturbation of this data instance according to (4.2). If we solve the resulting perturbed problem instances $\bar{d}_\alpha$ for select values of $\alpha \in [0, 1]$ and record the number of IPM iterations, we obtain the results portrayed in Figure 4.6. As the figure shows, the number of IPM iterations grows as the perturbed problem instance becomes more ill-posed.

The pattern of growth in IPM iterations as the perturbed problem becomes more ill-posed is not shared by all problem instances in the NETLIB suite. Figure 4.7 shows the plot of IPM iterations for problem `e226` as the perturbed problem instance becomes more ill-posed. For this problem instance the growth in IPM iterations is not monotone.

Of the 72 postprocessed problems in the NETLIB suite with finite condition measure, 59 of these problems satisfy $\rho_D(d) \geq \rho_P(d) > 0$ and $\|d\| > \rho(d)$, and so are amenable to analysis via the construction described above. For a given problem instance in the NETLIB suite, let k_α denote the number of IPM iterations needed to solve the perturbed postprocessed problem instance $\bar{d}_\alpha$. Then

$$\Delta k := k_1 - k_0$$

is the difference between the IPM iterations needed to solve the unperturbed and fully perturbed problem instances. Table 4.11 shows some summary statistics of the

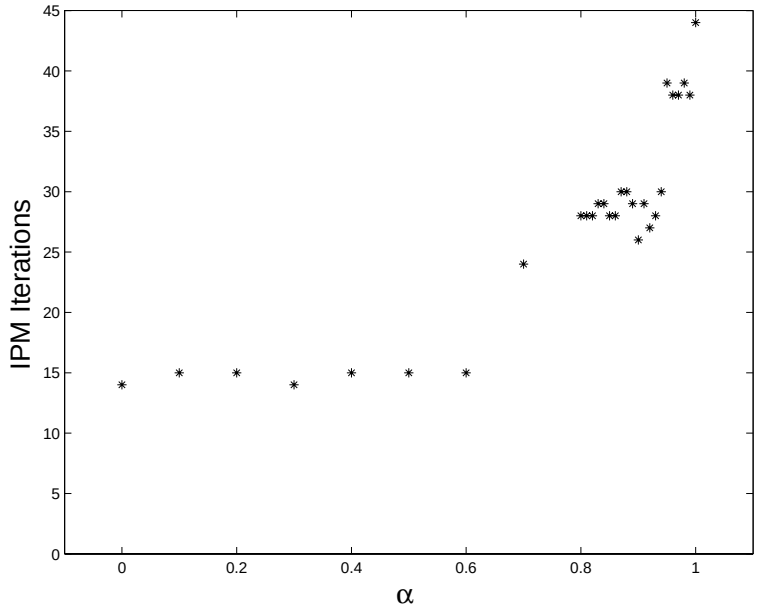


FIG. 4.6. *The number of IPM iterations needed to solve the perturbed postprocessed problem instance* `scagr25`*, as a function of the perturbation scalar* α*.*

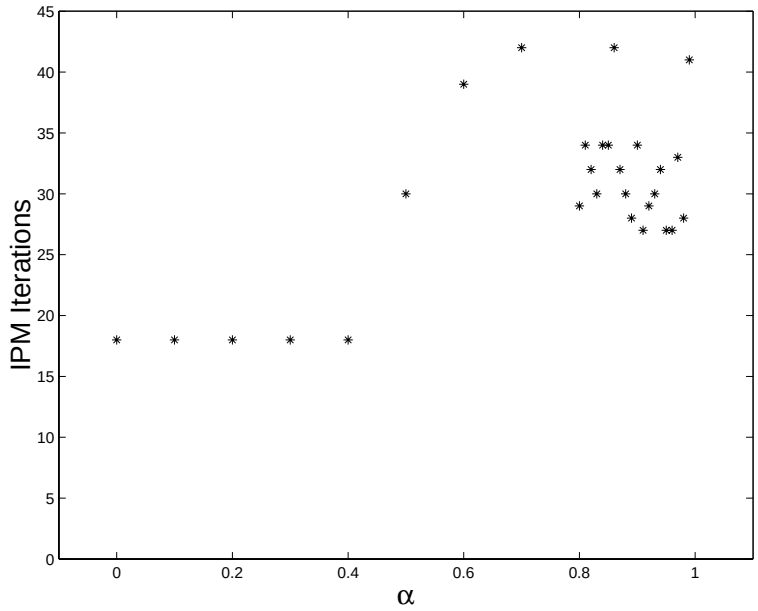


FIG. 4.7. *The number of IPM iterations needed to solve the perturbed postprocessed problem instance* `e226`*, as a function of the perturbation scalar* α*.*

distribution of Δk for the 59 problems in the NETLIB suite that are readily amenable to this analysis. As the table shows, the fully perturbed problem instance has a larger IPM iteration count in 68% (40 out of 59) of the problem instances. Curiously, the

TABLE 4.11
The distribution of the change in IPM iterations needed to solve the unperturbed problem instance and the fully perturbed problem instance for 59 postprocessed problems in the NETLIB suite.

Change in IPM iterations (Δk)	Number of problem instances
−1 or less	11
0	8
1 to 5	13
6 to 10	9
11 or more	18
Total	59

number of IPM iterations is actually less for the fully perturbed problem instance in 19% (11 out of 59) problem instances amenable to this analysis. A rough summary of the results in Table 4.11 is that the number of IPM iterations for the fully perturbed problem increases dramatically (more than 10 iterations) on 31% of the problem instances, increases modestly (1–10 iterations) on 37% of the problem instances, and remains the same or decreases slightly on 32% of problem instances.

5. Discussion and open questions. The purpose of this paper has been to study condition measures for linear optimization on problem instances that one might encounter in practice. We used the NETLIB suite of linear optimization problems as a test bed for condition measure computation and analysis, and we computed condition measures for 89 original NETLIB suite problem instances, as well as for the corresponding problem instances after preprocessing by CPLEX 7.1. We then investigated the extent to which the condition measure provides some explanatory value for the (already small) variance in the observed IPM iterations among problem instances in the NETLIB suite.

Except for certain classes of structured LP problems (see [12]), there is not yet a clear practical understanding (nor a theory) to explain the variation in the iteration counts of IPM algorithms (either theoretical or practical) on different LP instances. Herein we have explored the extent to which condition measures provide explanatory value for this variation. The scatter-plot in Figure 4.2 indicates that problem instances with large IPM iteration counts must have large condition measures. However, the converse of this assertion is not supported by the data; there are problem instances that have a high condition measure and low IPM iteration counts, for example, `agg2`, `recipe`, and some of the controlled-perturbation instances of section 4.4 with large condition measure.

It is easy to construct families of LP instances with ever-larger condition measures, whose IPM iteration counts do not grow excessively (see section 4.4). However, despite much effort, we have been unable to construct a family of problem instances with ever-larger practical IPM iteration counts but whose condition measures remains bounded. The existence of such a family is an open question.

The scatter-plot in Figure 4.2 indicates visually that there is a linear relationship between $\log C(d)$ and IPM iterations, and such a relationship is borne out by simple linear regression, with a resulting $R^2 = 0.4160$. However, in performing the regression analysis, there was no satisfactory way to include the 17 data instances with nonfinite (postprocessed) condition measures, and so these were removed, arguably biasing the results in favor of the explanatory value of the condition measure. A similar criticism can be made for the sample correlation coefficients computed in Table 4.10.

However, we feel that, at least on a relative basis, the results in Table 4.10 point to the conclusion that the condition measure does a better job of explaining the variation in IPM iteration counts than do any of the obvious reasonable alternative measures of problem size: $\log n$, $\log m$, $\log \vartheta$, or $\sqrt{\vartheta}$.

This work is a first attempt at explaining the observed performance of modern IPM solvers using condition measures that arise in the worst-case complexity analysis of interior-point methods. There are a variety of other instance-specific measures that have been used to bound the theoretical complexity of interior-point methods for linear optimization, including the bit-size L (see Karmarkar [11]), $\bar{\chi}_A$ (see Vavasis and Ye [26]), σ (see Ye [31]), and g and D_ϵ [6]. One natural question to ask is whether these or perhaps other measures might further explain the observed performance of IPM solvers.

Finally, the theory of condition measures referenced herein pertains to the very general class of conic convex optimization problems (and some formats for nonconic convex optimization problems as well), including semidefinite programming (SDP). Given the importance of SDP and the continuing development of IPM software for SDP, it is natural to ask to what extent condition measures (or other measures) might explain the observed performance of IPM solvers for SDP.

REFERENCES

[1] A. Ben-Tal and A. Nemirovski, *Robust solutions of linear programming problems contaminated with uncertain data*, Math. Program., 88 (2000), pp. 411–424.

[2] F. Cucker and J. Peña, *A primal-dual algorithm for solving polyhedral conic systems with a finite-precision machine*, SIAM J. Optim., 12 (2001), pp. 522–554.

[3] M. Epelman and R. M. Freund, *A new condition measure, preconditioners, and relations between different measures of conditioning for conic linear systems*, SIAM J. Optim., 12 (2002), pp. 627–655.

[4] S. Filipowski, *On the complexity of solving sparse symmetric linear programs specified with approximate data*, Math. Oper. Res., 22 (1997), pp. 769–792.

[5] S. Filipowski, *On the complexity of solving feasible linear programs specified with approximate data*, SIAM J. Optim., 9 (1999), pp. 1010–1040.

[6] R. M. Freund, *Complexity of convex optimization using geometry-based measures and a reference point*, Math. Program., to appear.

[7] R. M. Freund and M. J. Todd, *Barrier functions and interior-point algorithms for linear programming with zero-, one-, or two-sided bounds on the variables*, Math. Oper. Res., 20 (1995), pp. 415–440.

[8] R. M. Freund and J. R. Vera, *Condition-based complexity of convex optimization in conic linear form via the ellipsoid algorithm*, SIAM J. Optim., 10 (1999), pp. 155–176.

[9] R. M. Freund and J. R. Vera, *On the complexity of computing estimates of condition measures of a conic linear system*, Math. Oper. Res., to appear.

[10] R. M. Freund and J. R. Vera, *Some characterizations and properties of the "distance to ill-posedness" and the condition measure of a conic linear system*, Math. Program., 86 (1999), pp. 225–260.

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

[12] I. J. Lustig, R. E. Marsten, and D. F. Shanno, *The primal-dual interior point method on the Cray supercomputer*, in Large-Scale Numerical Optimization (Workshop held at Cornell University, Ithaca, NY, 1989), SIAM Proc. Appl. Math. 46, T. F. Coleman and Y. Li, eds., SIAM, Philadelphia, 1990, pp. 70–80.

[13] I. Lustig, R. Marsten, and D. Shanno, *Interior point methods: Computational state of the art*, ORSA J. Comput., 6 (1994), pp. 1–14.

[14] Y. Nesterov and A. Nemirovskii, *Interior-Point Polynomial Algorithms in Convex Programming*, SIAM Stud. Appl. Math. 13, SIAM, Philadelphia, 1994.

[15] *NETLIB linear programming library*, online at http://www.netlib.org/lp/.

[16] M. A. Nunez and R. M. Freund, *Condition measures and properties of the central trajectory of a linear program*, Math. Programming, 83 (1998), pp. 1–28.

[17] M. A. Nunez and R. M. Freund, *Condition-measure bounds on the behavior of the central trajectory of a semidefinite program*, SIAM J. Optim., 11 (2001), pp. 818–836.
[18] F. Ordóñez, *On the Explanatory Value of Condition Numbers for Convex Optimization: Theoretical Issues and Computational Experience*, Ph.D. thesis, MIT, Cambridge, MA, 2002.
[19] J. Peña, *A characterization of the distance to infeasibility under structured perturbations*, Linear Algebra Appl., to appear.
[20] J. Peña, *Computing the Distance to Infeasibility: Theoretical and Practical Issues*, Technical report, Center for Applied Mathematics, Cornell University, Ithaca, NY, 1998.
[21] J. Peña, *Understanding the geometry of infeasible perturbations of a conic linear system*, SIAM J. Optim., 10 (2000), pp. 534–550.
[22] J. Peña and J. Renegar, *Computing approximate solutions for convex conic systems of constraints*, Math. Program., 87 (2000), pp. 351–383.
[23] J. Renegar, *A polynomial-time algorithm, based on Newton's method, for linear programming*, Math. Programming, 40 (1988), pp. 59–93.
[24] J. Renegar, *Some perturbation theory for linear programming*, Math. Programming, 65 (1994), pp. 73–91.
[25] J. Renegar, *Linear programming, complexity theory, and elementary functional analysis*, Math. Programming, 70 (1995), pp. 279–351.
[26] S. A. Vavasis and Y. Ye, *A primal-dual interior point method whose running time depends only on the constraint matrix*, Math. Programming, 74 (1996), pp. 79–120.
[27] J. R. Vera, *Ill-Posedness and the Computation of Solutions to Linear Programs with Approximate Data*, Technical report, Cornell University, Ithaca, NY, 1992.
[28] J. R. Vera, *Ill-Posedness in Mathematical Programming and Problem Solving with Approximate Data*, Ph.D. thesis, Cornell University, Ithaca, NY, 1992.
[29] J. R. Vera, *Ill-posedness and the complexity of deciding existence of solutions to linear programs*, SIAM J. Optim., 6 (1996), pp. 549–569.
[30] J. R. Vera, *On the complexity of linear programming under finite precision arithmetic*, Math. Programming, 80 (1998), pp. 91–123.
[31] Y. Ye, *Toward probabilistic analysis of interior-point algorithms for linear programming*, Math. Oper. Res., 19 (1994), pp. 38–52.

SIAM J. OPTIM.
Vol. 14, No. 2, pp. 334–358

OPTIMAL CONTROL PROBLEMS WITH SET-VALUED CONTROL AND STATE CONSTRAINTS*

ZSOLT PÁLES† AND VERA ZEIDAN‡

Abstract. In this paper a general optimal control problem with pure state and mixed control-state constraints is considered. These constraints are of the form of set-inclusions. Second-order necessary optimality conditions for weak local minimum are derived for this problem in terms of the original data. In particular the nonemptiness of the set of critical directions and the evaluation of its support function are expressed in terms of the given functions and set-valued maps. In order that the Lagrange multiplier corresponding to the mixed control-state inclusion constraint be represented via an integrable function, a strong normality condition involving the notion of the critical tangent cone is introduced.

Key words. first- and second-order optimality conditions, critical cone, critical tangent cone, set-valued constraints

AMS subject classifications. 49B27, 49B36

DOI. 10.1137/S1052623401389774

1. Introduction.

Consider the following optimization problem:

$$(\mathcal{P}) \qquad \text{Minimize} \quad F(z) \quad \text{subject to} \quad G(z) \in \mathbf{Q},\ H(z) = 0,$$

where $F : \mathcal{D} \to \mathbb{R}$, $G : \mathcal{D} \to X$, $H : \mathcal{D} \to Y$, and X, Y, Z are Banach spaces, $\mathcal{D} \subset Z$ is nonempty and open, and $\mathbf{Q} \subset X$ is a closed convex set with nonempty interior.

The prototype of such problems arises, for instance, in optimal control theory with control and/or state constraints in the inclusion form $x(t) \in Q(t)$.

Better understanding of optimality conditions is an ongoing topic of research for several researchers. This question is of great value in theory and in applications. Usually, such conditions must be given in terms of the original data of the problem and, in the context of necessity, are expected to be as strong as they can be.

In 1988, Kawasaki [Kaw88a], [Kaw91] discovered, for the problem $(\mathcal{P})$, where $\mathbf{Q}$ is a cone, second-order necessary conditions that contain an extra term manifesting the presence of infinitely many inequalities in the constraint $G(z) \in \mathbf{Q}$. This phenomenon is known as the "envelope-like effect" and extends the results found in [BT80] and [BTZ82]. Such a result was generalized by Cominetti in [Com90]. Both results assumed a Mangasarian–Fromovitz-type condition.

In [PZ94a] the authors generalized the results of [Kaw88a], [Kaw91], [Kaw92], and [Com90] to the nondifferentiable case without assuming a Mangasarian–Fromovitz condition. The second-order admissible variation set used therein (defined first by Dubovitskii and Milyutin in [DM63] and [DM65]) is described in the following definition.

*Received by the editors May 24, 2001; accepted for publication (in revised form) March 7, 2003; published electronically August 22, 2003.
http://www.siam.org/journals/siopt/14-2/38977.html

†Institute of Mathematics and Informatics, University of Debrecen, H-4010 Debrecen, Pf. 12, Hungary (pales@math.klte.hu). Research supported by the Hungarian Scientific Research Fund (OTKA), grant T-038072, and by the Hungarian Higher Education, Research, and Development Fund (FKFP), grant 0215/2001.

‡Department of Mathematics, Michigan State University, East Lansing, MI 48824 (zeidan@math.msu.edu). Research supported partially by the Department of Mathematics at Michigan State University and by the National Science Foundation under grant DMS-0072598.

DEFINITION. *Let X be a normed space, $\mathbf{Q} \subset X$, $x \in \mathbf{Q}$, and $d \in X$. A vector $v \in X$ is called a* second-order admissible variation of $\mathbf{Q}$ at x in the direction d *if there exists $\bar{\varepsilon} > 0$ such that*

$$x + \varepsilon d + \varepsilon^2(v + u) \in \mathbf{Q} \qquad \text{for all} \quad 0 < \varepsilon < \bar{\varepsilon},\ \|u\| < \bar{\varepsilon},\ u \in X.$$

The set of all such variations is denoted by $V(x, d|\mathbf{Q})$. It follows directly from the definition that $V(x, d|\mathbf{Q})$ is an open set. If $\mathbf{Q}$ is also convex, then $V(x, d|\mathbf{Q})$ is convex as well.

In order to derive meaningful second-order optimality conditions, it is necessary to select directions d that guarantee the nonemptiness of $V(x, d|\mathbf{Q})$. Such directions $d \in X$ are labeled as the *critical directions of* $\mathbf{Q}$ *at* x and form a set called the *critical direction cone to* $\mathbf{Q}$ *at* x. Throughout this paper, this cone will be denoted by $C(x|\mathbf{Q})$. It can be easily seen that $C(x|\mathbf{Q})$ is a *convex cone* if $\mathbf{Q}$ is convex.

Define

$$S(x|\mathbf{Q}) := \operatorname{cone}(\mathbf{Q} - x) := \{\lambda(q - x) \mid q \in \mathbf{Q},\ \lambda > 0\}$$

and its closure

$$T(x|\mathbf{Q}) := \overline{\operatorname{cone}}(\mathbf{Q} - x) = \operatorname{cl} S(x|\mathbf{Q}).$$

If $\mathbf{Q}$ is convex, then for the nonemptiness of V it is necessary, but not sufficient, that the interior of $\mathbf{Q}$ be nonempty and that d belong to $T(x|\mathbf{Q})$. However, the nonemptiness of V is assured if $\operatorname{intr} \mathbf{Q} \neq \emptyset$ and $d \in S(x|\mathbf{Q})$. Therefore, for convex $\mathbf{Q}$, we have

$$S(x|\mathbf{Q}) \subset C(x|\mathbf{Q}) \subset T(x|\mathbf{Q}).$$

In the applications, when the inequality-type constraint is expressed in terms of several inclusions and inequalities, it is useful to know the following easy-to-establish *product rules*:

$$C(x|\mathbf{Q}) = \prod_{i=1}^{k} C(x_i|\mathbf{Q}_i) \qquad \text{and} \qquad V(x, d|\mathbf{Q}) = \prod_{i=1}^{k} V(x_i, d_i|\mathbf{Q}_i),$$

where $\mathbf{Q}_1, \ldots, \mathbf{Q}_k$ are subsets of vector spaces, $\mathbf{Q} := \mathbf{Q}_1 \times \cdots \times \mathbf{Q}_k$, $x = (x_1, \ldots, x_k) \in \mathbf{Q}$, $d = (d_1, \ldots, d_k) \in C(x|\mathbf{Q})$.

In order to recall the first- and second-order necessary conditions for $(\mathcal{P})$ obtained in [PZ94a, Corollary 2] and in [PZ96], we need to introduce the following notation and notions.

- A point $\widehat{z} \in \mathcal{D}$ is called an *admissible point* for $(\mathcal{P})$ if $G(\widehat{z}) \in \mathbf{Q}$ and $H(\widehat{z}) = 0$ hold. A point $\widehat{z} \in \mathcal{D}$ is a *solution (local minimum)* of the problem if it is admissible and there exists a neighborhood U of $\widehat{z}$ such that $F(z) \geq F(\widehat{z})$ for all admissible points $z \in U$.
- A point $\widehat{z} \in \mathcal{D}$ is called a *regular point* for $(\mathcal{P})$ if F, G, and H are strictly Fréchet differentiable at $\widehat{z}$ and the range of the linear operator $H'(\widehat{z})$ is a closed subspace of Y.

Let $\widehat{z}$ be an admissible regular point for $(\mathcal{P})$ and $d \in Z$.

- A vector $\delta z \in Z$ is called a *critical direction* at $\widehat{z}$ for $(\mathcal{P})$ if

$$F'(\widehat{z})\delta z \leq 0, \qquad G'(\widehat{z})\delta z \in C(G(\widehat{z})|\mathbf{Q}), \qquad H'(\widehat{z})\delta z = 0.$$

- A vector $\delta z \in Z$ is called a *regular direction* at $\widehat{z}$ for $(\mathcal{P})$ if the second-order directional derivative of $L := (F, G, H)$,

$$L''(\widehat{z}, \delta z) := \lim_{\varepsilon \to 0+} 2\frac{L(\widehat{z} + \varepsilon\delta z) - L(\widehat{z}) - \varepsilon L'(\widehat{z})\delta z}{\varepsilon^2},$$

  exists.

Clearly, the zero vector is always a regular critical direction at $\widehat{z}$ for $(\mathcal{P})$.

Now we are ready to state a particular case of the result of [PZ94a, Corollary 2].

THEOREM 1.1. *Let $\widehat{z}$ be a regular local solution of the above problem $(\mathcal{P})$. Then, for all regular critical directions δz, there correspond Lagrange multipliers $\lambda \geq 0$, $x^* \in X^*$, and $y^* \in Y^*$ (which depend on δz) such that at least one of them is different from zero and the following relations hold:*

$$x^* \in N(G(\widehat{z})|\mathbf{Q}), \tag{1.1}$$

$$\lambda F'(\widehat{z})z + \langle x^*, G'(\widehat{z})z\rangle + \langle y^*, H'(\widehat{z})z\rangle = 0 \qquad \textit{for} \quad z \in Z, \tag{1.2}$$

and

$$\lambda F''(\widehat{z}, \delta z) + \langle x^*, G''(\widehat{z}, \delta z)\rangle + \langle y^*, H''(\widehat{z}, \delta z)\rangle \geq 2\delta^*\big(x^*\big|V(G(\widehat{z}), G'(\widehat{z})\delta z|\mathbf{Q})\big). \tag{1.3}$$

(Here δ^* denotes the support functional defined by $\delta^*(x^*|V) := \sup_{v\in V}\langle x^*, v\rangle$ for $(x^* \in X^*)$, and $N(x|\mathbf{Q})$ denotes the adjoint cone of $T(x|\mathbf{Q})$, that is, the cone of outward normals to the set $\mathbf{Q}$ at the point x.)

As we have seen, the criticality of δz requires that $d \in C(x|\mathbf{Q})$, where $x := G(\widehat{z})$ and $d := G'(\widehat{z})\delta z$. However, in order that d be in $C(x|\mathbf{Q})$, it is only necessary that $\mathbf{Q}$ have a nonempty interior and that d belong to $T(x|\mathbf{Q})$. If $d \in S(x|\mathbf{Q})$, then $V(x, d|\mathbf{Q})$ is nonempty and $V(x, d|\mathbf{Q}) = \operatorname{cone}(\operatorname{cone}(\operatorname{intr}\mathbf{Q} - x) - d)$ (cf. [PZ94a, Theorem 4]). In this case the right-hand side in the second-order condition (1.3) vanishes. However, examples are provided by Kawasaki [Kaw88a] which show that the necessary conditions with *extra term*, that is, when $d \in T(x|\mathbf{Q})$, handle situations that cannot be handled with previous results where d is taken from $S(x|\mathbf{Q})$. Thus, one has to also consider directions $d \in T(x|\mathbf{Q}) \setminus \operatorname{cone}(\mathbf{Q} - x)$. In this important case two questions naturally arise from Theorem 1.1:

(i) How can we check the nonemptiness of $V(x, d|\mathbf{Q})$; that is, how can the critical cone $C(x|\mathbf{Q})$ be characterized in terms of $\mathbf{Q}$?

(ii) How can we evaluate the support function of $V(x, d|\mathbf{Q})$?

A significant setting is the case when $\mathbf{Q}$ is a subset of $\mathcal{C}(T, \mathbb{R}^\kappa)$ defined by

$$\mathbf{Q} = \operatorname{sel}_C(Q) := \{\ x \in \mathcal{C}(T, \mathbb{R}^\kappa) \mid x(t) \in Q(t) \ \text{ for all } t \in T\ \}, \tag{1.4}$$

where Q is a lower semicontinuous set-valued map whose images are closed, convex sets with nonempty interior, and T is a compact Hausdorff space. The importance of this type of constraint stems from control problems with state constraints.

Another case of interest is when $\mathbf{Q}$ is a subset of $\mathcal{L}^\infty(\Omega, \mathbb{R}^\gamma)$ defined by

$$\mathbf{Q} = \operatorname{sel}_\infty(Q) := \{\ x \in \mathcal{L}^\infty(\Omega, \mathbb{R}^\gamma) \mid x(t) \in Q(t) \ \text{ for a.e. } t \in \Omega\ \}, \tag{1.5}$$

where Q is a measurable set-valued map whose images are closed and have nonempty interior, and $(\Omega, \mathcal{A}, \nu)$ is a complete finite measure space. This type of constraint is typical for control constraints in control problems.

The main goal of this paper is to investigate these two types of constraints so that the application of Theorem 1.1 to optimal control problems leads to weak-local optimality necessary conditions that are phrased in terms of the original data; part of these results is announced in [PZ01]. However, given the fact that the state variable x and the control variable u belong to different spaces, it has been known for a long time (see, e.g., [PZ94b, Theorem 3]) that to obtain a result for an optimal control problem by applying an abstract result like Theorem 1.1, one should first derive a specialized version of that abstract result that takes into account the distinct features of each of these variables. Such a result has been developed in [PZ94b] and will be recalled in the next section.

Since, for control problems, the constraint set $\mathbf{Q}$ could be a product of different types of constraints, that is, endpoint set-inclusion, control and state set-inclusion, therefore, we shall need the following *sum rule* for the extra term in (1.3):

$$\delta^*(x^*|V(x,d|\mathbf{Q})) = \sum_{i=1}^{k} \delta^*(x_i^*|V(x_i,d_i|\mathbf{Q}_i)),$$

where $\mathbf{Q}_1,\dots,\mathbf{Q}_k$ are subsets of vector spaces, $\mathbf{Q} := \mathbf{Q}_1 \times \cdots \times \mathbf{Q}_k$, $x = (x_1,\dots,x_k) \in \mathbf{Q}$, $d = (d_1,\dots,d_k) \in C(x|\mathbf{Q})$, and $x^* = (x_1^*,\dots,x_k^*)$.

The paper is divided as follows. In section 2, auxiliary results needed for the main result are presented. In particular, when $\mathbf{Q}$ is given by (1.4) or (1.5), we recall the characterizations of both normal and critical cones ($N(x|\mathbf{Q})$ and $C(x|\mathbf{Q})$), and the evaluation of the support function of $V(x,d|\mathbf{Q})$ in terms of the images of the set-valued map Q. Also, we state a special version of Theorem 1.1 which is tailored for the abstract control setting and which will be used later in proving the main result. However, when $\mathbf{Q}$ is given by (1.5), the multiplier x^* in Theorem 1.1 corresponding to the set inclusion constraint is in general in $(\mathcal{L}^\infty(\Omega,\mathbb{R}^\gamma))^*$. Therefore, it is important in this case to obtain a reasonable sufficiency criterion for x^* to be represented by an integrable function. This is accomplished in section 3 by using a uniform solvability criterion. In section 4, the results of the preceding sections are used to obtain second-order necessary conditions for optimality in a general optimal control problem with control and state set-valued constraints. These conditions are phrased in terms of the critical tangent cone. A specialization of Theorem 4.1 to the case of *inequality* constraints is presented in Corollary 4.1. Therein, only the extra term corresponding to the pure-state constraints remains present. This term is phrased in terms of the function σ defined in (2.14). Finally, a numerical example is provided at the end of section 4 in order to illustrate the utility of these results.

2. Auxiliary results. When dealing with control problems, there are two special cases for X and $\mathbf{Q}$ where the characterization of the critical cone $C(x|\mathbf{Q})$ and the evaluation of the support function of $V(x,d|\mathbf{Q})$ are imperative.

The first setting considers $X = \mathcal{C}(T,\mathbb{R}^\kappa)$, where $T = (T,\rho)$ is a compact metric space, and $Q : T \to 2^{\mathbb{R}^\kappa}$ is a lower semicontinuous set-valued function whose images are closed and convex with nonempty interior. Define the $\mathbf{Q} \subset \mathcal{C}(T,\mathbb{R}^\kappa)$ as the set of continuous selections of Q by

$$\mathbf{Q} = \mathrm{sel}_C(Q) := \{x \in \mathcal{C}(T,\mathbb{R}^\kappa) \mid x(t) \in Q(t) \text{ for } t \in T\}. \tag{2.1}$$

Then $\mathrm{sel}_C(Q)$ is a closed convex set of $\mathcal{C}(T,\mathbb{R}^\kappa)$.

Regarding $\mathbf{Q} = \mathrm{sel}_C(Q)$, a thorough study of convex analysis concepts (normal and tangent cones, support function, etc.) was developed in [PZ99a]. For instance, if

we denote by $\frac{d\mu}{d|\mu|}$ the Radon–Nikodým derivative of μ with respect to $|\mu|$, it is shown that

(2.2)
$$\mu \in N(x|\operatorname{sel}_C(Q)) \qquad \text{if and only if} \qquad \frac{d\mu}{d|\mu|}(t) \in N(x(t)|Q(t)) \quad \text{for } \mu\text{-a.e. } t \in T.$$

Results concerning the second-order admissible variations, critical cone, and application to abstract optimization were derived in [PZ98]. The nonemptiness of the interior of the images of the set-valued function Q implies, by [PZ99a, Theorem 4.2], that $\operatorname{sel}_C(Q)$ has a nonempty interior, too. A characterization of the set of critical directions is offered by the following results from [PZ98, Theorem 3.5, Lemmas 3.6 and 3.8]. Note that condition (2.3) below needs to be verified for $\xi \in \mathbb{R}^\kappa$, i.e., over a finite-dimensional space.

THEOREM 2.1. *Let $x \in \operatorname{sel}_C(Q)$. Then $d \in \mathcal{C}(T, \mathbb{R}^\kappa)$ is in the critical cone $C(x|\operatorname{sel}_C(Q))$ if and only if there exists a constant $M > 0$ such that, for all $t \in T$,*

(2.3)
$$\langle \xi, d(t)\rangle^2 \le M|\xi|(\delta^*(\xi|Q(t)) - \langle \xi, x(t)\rangle) \qquad \textit{whenever} \quad \xi \in \mathbb{R}^\kappa \ \textit{ and } \ \langle \xi, d(t)\rangle > 0.$$

A consequence of Theorem 2.1 concerns the connection between $C(x|\operatorname{sel}_C(Q))$ and the set-valued mapping $t \mapsto C(x(t)|Q(t))$. From Theorem 2.1 applied to $Q := Q(t)$, $T = \{t\}$, and $d = d(t)$ (where t is kept fixed), it results that $d(t) \in C(x(t)|Q(t))$ is equivalent to the fact that (2.3) holds for some constant $M_t > 0$. Therefore, Theorem 2.1 can be reformulated as follows:

A continuous function d belongs to $C(x|\operatorname{sel}_C(Q))$ if and only if

$$d(t) \in C(x(t)|Q(t)) \qquad (t \in T), \tag{2.4}$$

and the corresponding constants M_t from (2.3) *can be chosen to be uniformly bounded.*

When (2.3) is valid for some constant M and for all $t \in T$, then we say that (2.4) *holds uniformly in $t \in T$.*

The second special setting is when $X = \mathcal{L}^\infty(\Omega, \mathbb{R}^\gamma)$, where $(\Omega, \mathcal{A}, \nu)$ is a complete finite measure space, and $Q : \Omega \to 2^{\mathbb{R}^\gamma}$ is a measurable set-valued function whose images are closed sets with nonempty interior and $\operatorname{sel}_\infty(Q)$ is defined by

$$\mathbf{Q} = \operatorname{sel}_\infty(Q) := \{x \in \mathcal{L}^\infty(\Omega, \mathbb{R}^\gamma) \mid x(t) \in Q(t) \text{ for a.e. } t \in \Omega\}.$$

For this case, the concept of convex analysis was studied in [PZ99b], [PZ99c]. In particular, for $x \in \operatorname{sel}_\infty(Q)$ and for $\varphi \in L^1(\Omega, \mathbb{R}^\gamma)$,

(2.5)
$$\varphi \in N(x|\operatorname{sel}_\infty(Q)) \qquad \text{if and only if} \qquad \varphi(t) \in N(x(t)|Q(t)) \qquad \text{for a.e. } t \in \Omega.$$

For the second-order admissible variations, critical cone, and the application to second-order optimality conditions in an abstract setting, results were obtained in [PZ00].

In order that the interior of $\operatorname{sel}_\infty(Q)$ be nonempty it is necessary and sufficient (by [PZ99c, Theorem 3]; see also [PZ99b]) to assume that Q satisfies

$$\exists r \ge \rho > 0 \text{ and, for a.e. } t \in \Omega, \exists x_t \in \mathbb{R}^\gamma \text{ such that } B_\rho(x_t) \subset Q(t) \cap B_r, \tag{2.6}$$

where $B_\rho(x)$ stands for the ball centered at x of radius ρ, and B_r stands for the ball centered at 0 of radius r.

The following consists of a characterization of $C(x|\operatorname{sel}_\infty(Q))$. It provides a verifiable condition over a finite-dimensional space. As was the case in Theorem 2.1 for the space of continuous functions, (2.7) below is to be checked for elements $\xi \in \mathbb{R}^\gamma$ even though the underlying space is $\mathcal{L}^\infty(\Omega, \mathbb{R}^\gamma)$.

THEOREM 2.2. *Let $Q : \Omega \to 2^{\mathbb{R}^\gamma}$ be a measurable set-valued map whose images are closed convex sets and satisfy* (2.6). *Let $x \in \operatorname{sel}_\infty(Q)$ and $d \in \mathcal{L}^\infty(\Omega, \mathbb{R}^\gamma)$. Then $d \in C(x|\operatorname{sel}_\infty(Q))$ if and only if there exists a constant $M > 0$ such that, for a.e. $t \in \Omega$, the following condition is valid:*

(2.7)
$$\langle \xi, d(t)\rangle^2 \le M|\xi|(\delta^*(\xi|Q(t)) - \langle \xi, x(t)\rangle) \qquad \text{whenever} \quad \xi \in \mathbb{R}^\gamma \text{ and } \langle \xi, d(t)\rangle > 0.$$

From Theorem 2.2 it readily follows that, for a.e. $t \in \Omega$, $d(t) \in C(x(t)|Q(t))$ if and only if (2.7) holds for some $M_t > 0$ on the domain indicated. Therefore, Theorem 2.2 can be rephrased as

A bounded measurable function d belongs to $C(x|\operatorname{sel}_\infty(Q))$ if and only if

$$d(t) \in C(x(t)|Q(t)) \qquad \text{for a.e. } (t \in \Omega), \tag{2.8}$$

and the corresponding constants M_t from (2.7) *can be chosen to be uniformly bounded on a set of full measure.*

When (2.7) is valid for some constant M and for a.e. $t \in \Omega$, then we say that (2.8) *holds almost uniformly on* Ω.

The rest of this section is devoted to recalling the results on the calculation of the support functional to the second-order admissible variation set of $\operatorname{sel}_C(Q)$ and $\operatorname{sel}_\infty(Q)$, respectively.

We introduce the following notation. Let Q be a subset of $\mathbb{R}^\gamma$, $x \in Q$, and $d \in \mathbb{R}^\gamma$. Denote

$$E(x, d|Q)(\xi) := \frac{\langle \xi, d\rangle^2}{4[\langle \xi, x\rangle - \delta^*(\xi|Q)]} \qquad \text{for} \quad \xi \in \mathbb{R}^\gamma \text{ such that } \xi \notin N(x|Q).$$

Note that $E(x, d|Q)(\xi)$ is well defined, because $\langle \xi, x\rangle - \delta^*(\xi|Q) \ne 0$ if and only if $\xi \notin N(x|Q)$. If $d \in T(x|Q)$ and $\langle \xi, d\rangle > 0$, then $\xi \notin N(x|Q)$; hence, in this case, $E(x, d|Q)(\xi)$ is defined for $\langle \xi, d\rangle > 0$.

Set

$$d^\perp := \{\xi \in \mathbb{R}^\gamma \mid \langle \xi, d\rangle = 0\}, \qquad d^> := \{\xi \in \mathbb{R}^\gamma \mid \langle \xi, d\rangle > 0\},$$

and define from $\mathbb{R}^\gamma$ to the extended reals the function

$$\boldsymbol{\mathcal{E}}(x, d|Q)(\xi) := \begin{cases} \liminf\limits_{\substack{\zeta \to \xi \\ \zeta \in d^>}} E(x, d|Q)(\zeta) & \text{if } \xi \in N(x|Q) \cap d^\perp, \\ +\infty, & \text{otherwise.} \end{cases} \tag{2.9}$$

One can see that $\boldsymbol{\mathcal{E}}(x, d|Q)(\cdot)$ is a positively homogeneous and also lower semicontinuous function on $\mathbb{R}^\gamma \setminus \{0\}$.

Define the convex regularization $\overline{\operatorname{co}}\,\boldsymbol{\mathcal{E}}(x, d|Q)(\cdot)$ to be the largest lower semicontinuous convex function below $\boldsymbol{\mathcal{E}}(x, d|Q)(\cdot)$; that is,

$$\overline{\operatorname{co}}\,\boldsymbol{\mathcal{E}}(x, d|Q)(\xi) = \sup\{\, \varphi(\xi) \mid \varphi : \mathbb{R}^\gamma \to [-\infty, \infty] \text{ is convex and lsc,} \\ \varphi(\zeta) \le \boldsymbol{\mathcal{E}}(x, d|Q)(\zeta)\, \forall \zeta \in \mathbb{R}^\gamma \setminus \{0\} \,\}.$$

It results that $\overline{\operatorname{co}}\,\boldsymbol{\mathcal{E}}(x, d|Q)(\cdot)$ is also sublinear.

The following result offers an evaluation of the support function of the set $V(x,d|\operatorname{sel}_\infty(Q))$ at linear functionals that can be represented in terms of integrable functions (cf. [PZ00, Corollary 2.7]).

THEOREM 2.3. *Let Q be a closed convex set–valued measurable set-valued map on Ω, $x\in\operatorname{sel}_\infty(Q)$, and $d\in C(x|\operatorname{sel}_\infty(Q))$, and let $\varphi\in\mathcal{L}^1(\Omega,\mathbb{R}^\gamma)$. Then*

$$\delta^*\big(\varphi\big|V(x,d|\operatorname{sel}_\infty(Q))\big)=\int_\Omega \overline{\operatorname{co}}\,\mathcal{E}\Big(x(t),d(t)|Q(t)\Big)(\varphi(t))\,d\nu(t). \tag{2.10}$$

A common type of constraint is when Q comes from inequality constraints, that is, when $Q(t)=\mathbb{R}^\gamma_-$ for all $t\in\Omega$. In this case the description of the critical cone and the evaluation of the support function simplify drastically.

COROLLARY 2.1. *Let $x\in\operatorname{sel}_\infty(\mathbb{R}^\gamma_-)$. Then a bounded measurable function $d=(d_1,\dots,d_\gamma):\Omega\to\mathbb{R}^\gamma$ is in $C(x|\operatorname{sel}_\infty(\mathbb{R}^\gamma_-))$ if and only if*

(i) *there exists a constant $M\geq 0$ such that, for all $i=1,\dots,\gamma$ and for a.e. $t\in\Omega$, $d_i^2(t)\leq -Mx_i(t)$ whenever $x_i(t)\leq 0$ and $d_i(t)>0$ hold;*

(ii) *for a.e. $t\in\Omega$ with $x_i(t)=0$, we have $d_i(t)\leq 0$.*

Furthermore, let $x\in\operatorname{sel}_\infty(\mathbb{R}^\gamma_-)$, $d\in C(x|\operatorname{sel}_\infty(\mathbb{R}^\gamma_-))$, and let $\varphi:\Omega\to\mathbb{R}^\gamma_+$ be an integrable function such that $\varphi^T(t)x(t)=0$ and $\varphi^T(t)d(t)=0$ for a.e. $t\in\Omega$. Then

$$\delta^*\big(\varphi|V(x,d|\operatorname{sel}_\infty(\mathbb{R}^\gamma_-))\big)=0. \tag{2.11}$$

Proof. Using the product rule for the critical cone and the first part of [PZ94b, Lemma 7], we get that the inclusion $d\in C(x|\operatorname{sel}_\infty(\mathbb{R}^\gamma_-))$ is characterized by conditions (i) and (ii).

Observe that the nonnegativity of φ and the conditions $\varphi^Tx=0$ and $\varphi^Td=0$ yield that $\varphi_i^Tx_i=0$ and $\varphi_i^Td_i=0$ for all $i=1,\dots,\gamma$ almost everywhere in Ω. Thus, applying the sum rule for the evaluation of the support function of second-order variation sets and the second part of [PZ94b, Lemma 7], the second statement of the corollary will follow. □

The analogous result for the case of $\mathbf{Q}=\operatorname{sel}_C(Q)$ requires more involved notions (see [PZ98]). Let T be a compact metric space, and let $Q:T\to 2^{\mathbb{R}^\kappa}$ be a set-valued function whose images are closed and convex sets with nonempty interior. Let $x\in\operatorname{sel}_C(Q)$ and $d\in C(x|\operatorname{sel}_C(Q))$. Denote by $d^\#:T\to 2^{\mathbb{R}^\kappa}$ the following set-valued function:

$$d^\#(t)=\{\ \xi\in\mathbb{R}^\kappa\mid \exists t_n\to t,\ \exists\xi_n\to\xi \text{ with } \xi_n\in d(t_n)^>\ \forall n\ \}.$$

Define

$$\mathbb{E}(x,d|Q)(t,\xi):=\begin{cases}\liminf\limits_{\substack{(s,\zeta)\to(t,\xi)\\ \zeta\in d(s)^>}} E(x(s),d(s)|Q(s))(\zeta) & \text{if } \xi\in N(x(t)|Q(t))\cap d(t)^\perp\cap d^\#(t),\\ 0 & \text{if } \xi\in N(x(t)|Q(t))\cap d(t)^\perp\setminus d^\#(t),\\ +\infty, & \text{otherwise.}\end{cases} \tag{2.12}$$

Define the convex regularization $\overline{\operatorname{co}}\,\mathbb{E}(x,d|Q)(\cdot,\cdot)$ to be the largest lower semicontinuous function $\varphi:T\times\mathbb{R}^\kappa\to[-\infty,\infty]$ below $\mathbb{E}(x,d|Q)(\cdot,\cdot)$ such that, for each $t\in T$, the function $\xi\mapsto\varphi(t,\xi)$ is convex on $\mathbb{R}^\kappa$.

In the following result (cf. [PZ98, Theorem 3.10]), we describe how the support functional of $V(x,d|\operatorname{sel}_C(Q))$ can be evaluated in terms of $\overline{\operatorname{co}}\,\mathbb{E}$.

THEOREM 2.4. *Let T be a compact metric space, and let $Q : T \to 2^{R^\kappa}$ be a lower semicontinuous set-valued function whose images are closed and convex with nonempty interior. Let $x \in \operatorname{sel}_C(Q)$, $d \in C(x|\operatorname{sel}_C(Q))$, and let μ be a bounded vector-valued Borel measure on T. Then*

$$\delta^*\big(\mu\big|V(x,d|\operatorname{sel}_C(Q))\big) = \int_T \overline{\operatorname{co}}\,\mathbb{E}(x,d|Q)\left(t, \frac{d\mu}{d|\mu|}(t)\right) d|\mu|(t), \tag{2.13}$$

where $\frac{d\mu}{d|\mu|}(\cdot)$ is the Radon–Nikodým derivative of μ with respect to its total variation $|\mu|$.

For given continuous functions $a, b : T \to \mathbb{R}$, define $\sigma_{a,b} : T \to [-\infty, \infty]$ by

$$\sigma_{a,b}(t) := \begin{cases} \liminf\limits_{a(\tau) \overset{\tau}{<} 0,\ b(\tau) > 0} \dfrac{b^2(\tau)}{4a(\tau)} & \text{if} \quad t \in T_{a=0,\,b=0} \cap \partial(T_{a<0,\,b>0}), \\ 0 & \text{if} \quad t \in T_{a=0,\,b=0} \setminus \partial(T_{a<0,\,b>0}), \\ +\infty, & \text{otherwise,} \end{cases} \tag{2.14}$$

where

$$T_{a=0,\,b=0} := \{t \in T \,|\, a(t) = 0,\ b(t) = 0\}, \qquad T_{a<0,\,b>0} := \{t \in T \,|\, a(t) < 0,\ b(t) > 0\}.$$

COROLLARY 2.2. *Let $x = (x_1, \dots, x_\kappa) \in \operatorname{sel}_C(\mathbb{R}^\kappa_-)$. Then a continuous function $d = (d_1, \dots, d_\kappa) : T \to \mathbb{R}^\kappa$ is in $C(x|\operatorname{sel}_C(\mathbb{R}^\kappa_-))$ if and only if*

(i) *there exists a constant $M \geq 0$ such that, for all $i = 1, \dots, \kappa$ and for all $t \in T$, $d_i^2(t) \leq -M x_i(t)$ whenever $x_i(t) \leq 0$ and $d_i(t) > 0$ hold;*

(ii) *for all $t \in T$ with $x_i(t) = 0$, we have $d_i(t) \leq 0$.*

Furthermore, let $x \in \operatorname{sel}_C(\mathbb{R}^\kappa_-)$, $d \in C(x|\operatorname{sel}_C(\mathbb{R}^\kappa_-))$, and let $\mu = (\mu_1, \dots, \mu_\kappa)$ be a bounded vector-valued Borel measure on T with nonnegative components such that $\operatorname{supp}\mu_i \subset \{t \in T \,|\, x_i(t) = 0,\ d_i(t) = 0\}$ for all $i = 1, \dots, \kappa$. Then

$$\delta^*\big(\mu\big|V(x,d|\operatorname{sel}_C(\mathbb{R}^\kappa_-))\big) = \sum_{i=1}^{\kappa} \int_T \sigma_{x_i,d_i}(t)\, d\mu_i(t). \tag{2.15}$$

Proof. Using the product rule for the critical cone and [PZ98, Corollary 4.2(i)], we get that the inclusion $d \in C(x|\operatorname{sel}_C(\mathbb{R}^\kappa_-))$ is characterized by conditions (i) and (ii).

Applying the sum rule for the evaluation of the support function of second-order variation sets and [PZ98, Corollary 4.2(iv)], the second statement of the corollary will also follow. □

In the rest of this section we present second-order optimality conditions for the following abstract control problems, which are a special form of the problem $(\mathcal{P})$. This problem allows the distinction between the control and the state variables:

Assume that X, U, Y, V, and W are Banach spaces (over $\mathbb{R}$), and $D \subset X \times U$ is nonempty and open. Let $\mathbf{F} : D \to \mathbb{R}$, $\mathbf{G} : D \to V$, $\mathbf{H} : D \to W$, $\mathbf{K} : D \to Y$, and, further, that $\mathbf{Q} \subset V$ is a closed convex set with nonempty interior. The problem **$(\mathcal{P})$** is to minimize $F(x,u)$ in $(x,u) \in D$ subject to

(i) $\mathbf{G}(x,u) \in \mathbf{Q}$ (Banach space–valued mixed state-control inequality and control set constraint),

(ii) $\mathbf{H}(x,u) = 0$ (Banach space–valued mixed state-control equality),

(iii) $\mathbf{K}(x,u) = 0$ (control system).

The admissibility and optimality of a pair $(x,u)\in D$ is defined similarly to that of problem $(\mathcal{P})$.

The second constraint $\mathbf{G}(x,u)\in\mathbf{Q}$ is able to handle Banach space–valued inequalities and control set constraint as well. For instance, if $\mathbf{Q}$ is a closed convex cone with nonempty interior, then introducing the ordering $\leq_{\mathbf{Q}}$ in V by $x\leq_{\mathbf{Q}} y\iff x-y\in\mathbf{Q}$, one can see that our first constraint can be rewritten as $\mathbf{G}(x,u)\leq_{\mathbf{Q}}0$. On the other hand the constraints $u\in Q$ or $x\in Q$, where Q is a convex set with nonempty interior, are obviously a particular case of (i). In this case problem **(𝒫)** specializes to the mixed problem dealt with in [IT79, section 1.1.3, p. 70]. However, both the regularity assumptions and the results are of a different nature from those in our case.

At this stage one cannot make any difference between the mixed state-control equality and control system constraints. However, the difference becomes clear when evoking the regularity conditions stated below.

A pair $(\widehat{x},\widehat{u})\in D$ is called *regular* for problem **(𝒫)** if the following conditions are satisfied:

- $(\mathbf{R}_1)$ $\mathbf{G}$ is strictly Fréchet differentiable at $(\widehat{x},\widehat{u})$;
- $(\mathbf{R}_2)$ $\mathbf{H}$ is strictly Fréchet differentiable at $(\widehat{x},\widehat{u})$ and the partial Fréchet derivative $\mathbf{H}_u(\widehat{x},\widehat{u}):U\to W$ has the *full rank property*; that is, it has a bounded right inverse;
- $(\mathbf{R}_3)$ $\mathbf{K}$ is strictly Fréchet differentiable at $(\widehat{x},\widehat{u})$ and the equation is an *abstract control system* at $(\widehat{x},\widehat{u})$; i.e., the partial derivative $\mathbf{K}_x(\widehat{x},\widehat{u})$ is a Fredholm operator and $\mathbf{K}_u(\widehat{x},\widehat{u})$ is compact.

We note that when $\mathbf{K}$ fulfills the above assumption at each point of D, then the equation $\mathbf{K}=0$ will be called a *(global) control system*. It is worth noting that if $\mathbf{K}$ is continuously Fréchet differentiable on D, $\mathbf{K}_x$ is a Fredholm operator, and D is a connected set, then $\operatorname{ind}\mathbf{K}_x$ is constant on D and hence the index of a control system could be defined.

We indicate by $\widehat{\Phi}$ the evaluation of the function Φ at $(\widehat{x},\widehat{u})$.

Let $(\widehat{x},\widehat{u})$ be a regular admissible pair for problem **(𝒫)**. A direction $(\delta x,\delta u)\in X\times U$ is called *regular* for our problem **(𝒫)** at $(\widehat{x},\widehat{u})$ if

- $(\mathbf{R}_4)$ the second-order directional derivatives $\widehat{\mathbf{G}}''(\delta x,\delta u)$, $\widehat{\mathbf{H}}''(\delta x,\delta u)$, and $\widehat{\mathbf{K}}''(\delta x,\delta u)$ of $\mathbf{G},\mathbf{H}$, and $\mathbf{K}$, respectively, exist at $(\widehat{x},\widehat{u})$ in the direction $(\delta x,\delta u)$.

A direction $(\delta x,\delta u)\in X\times U$ is called *critical* for **(𝒫)** at $(\widehat{x},\widehat{u})$ if

- $(\mathbf{C}_1)$ $\widehat{\mathbf{F}}_x\delta x+\widehat{\mathbf{F}}_u\delta u\leq 0$;
- $(\mathbf{C}_2)$ $\widehat{\mathbf{G}}_x\delta x+\widehat{\mathbf{G}}_u\delta u\in C(\widehat{\mathbf{G}}|\mathbf{Q})$, $\widehat{\mathbf{H}}_x\delta x+\widehat{\mathbf{H}}_u\delta u=0$, $\widehat{\mathbf{K}}_x\delta x+\widehat{\mathbf{K}}_u\delta u=0$.

One can check that $(\delta x,\delta u)=(0,0)$ is always a regular and critical direction at $(\widehat{x},\widehat{u})$ for **(𝒫)**.

The next result is the multiplier rule for problem **(𝒫)** obtained in [PZ94b, Theorem 3].

THEOREM 2.5. *Let $(\widehat{x},\widehat{u})$ be a regular solution for problem **(𝒫)**. Then, for every regular critical direction $(\delta x,\delta u)\in X\times U$, there exist Lagrange multipliers $v^*\in V^*$, $w^*\in W^*$, and $y^*\in Y^*$ such that at least one of them is different from zero and the following relations hold:*

$$\langle v^*,v\rangle\leq 0\quad \text{for}\quad v\in\mathbf{Q}-\widehat{\mathbf{G}},\qquad \langle v^*,\widehat{\mathbf{G}}_x\delta x+\widehat{\mathbf{G}}_u\delta u\rangle=0, \tag{2.16}$$

$$\lambda\widehat{\mathbf{F}}_x+v^*\circ\widehat{\mathbf{G}}_x+w^*\circ\widehat{\mathbf{H}}_x+y^*\circ\widehat{\mathbf{K}}_x=0, \tag{2.17}$$

$$\lambda\widehat{\mathbf{F}}_u+v^*\circ\widehat{\mathbf{G}}_u+w^*\circ\widehat{\mathbf{H}}_u+y^*\circ\widehat{\mathbf{K}}_u=0, \tag{2.18}$$

and

$$\begin{aligned}\lambda\widehat{\mathbf{F}}''(\delta x,\delta u)+\big\langle v^*,\widehat{\mathbf{G}}''(\delta x,\delta u)\big\rangle+\big\langle w^*,\widehat{\mathbf{H}}''(\delta x,\delta u)\big\rangle+\big\langle y^*,\widehat{\mathbf{K}}''(\delta x,\delta u)\big\rangle\\ \geq 2\delta^*\big(v^*\big|V(\widehat{\mathbf{G}},\widehat{\mathbf{G}}_x\delta x+\widehat{\mathbf{G}}_u\delta u|\mathbf{Q})\big).\end{aligned}\tag{2.19}$$

3. Uniform solvability criteria. In the next result, we characterize the solvability of a system of linear equations over cones in different ways.

THEOREM 3.1. *Let $G\in\mathbb{R}^{\gamma\times m}$, $H\in\mathbb{R}^{\delta\times m}$, and $D\in\mathbb{R}^{\gamma\times q}$ be matrices and let $C\subset\mathbb{R}^q$ be a closed convex cone. Then the following four statements are equivalent to each other:*

(i) *For all vectors $v\in\mathbb{R}^\gamma$ and $w\in\mathbb{R}^\delta$, there exist $a\in\mathbb{R}^m$ and $c\in C$ such that*

$$v=Ga-Dc\qquad\text{and}\qquad w=Ha.\tag{3.1}$$

(ii) *If $\xi\in\mathbb{R}^\gamma$, $\eta\in\mathbb{R}^\delta$, then*

$$\xi^TG+\eta^TH=0\qquad\text{and}\qquad\xi^TD\in C^\circ\tag{3.2}$$

are valid if and only if $(\xi,\eta)=(0,0)$. (Here C° denotes the polar cone of C.)

(iii) *There exists a constant $\tau>0$ such that*

$$|\xi^TG+\eta^TH|^2+[\operatorname{dist}(\xi^TD,C^\circ)]^2\geq\tau|(\xi,\eta)|^2\qquad((\xi,\eta)\in\mathbb{R}^\gamma\times\mathbb{R}^\delta).\tag{3.3}$$

(iv) *The matrix H is of full rank and there exist two maps $a:\mathbb{R}^\gamma\times\mathbb{R}^\delta\to\mathbb{R}^m$, $c:\mathbb{R}^\gamma\times\mathbb{R}^\delta\to\mathbb{R}^q$ and a constant $\rho>0$ such that*

$$Ga(v,w)-Dc(v,w)=v,\quad Ha(v,w)=w,\quad c(v,w)\in C\qquad((v,w)\in\mathbb{R}^\gamma\times\mathbb{R}^\delta),\tag{3.4}$$

and $\|(HH^T)^{-1}\|\leq\rho$,

$$\begin{aligned}&|a(v,w)|\leq\rho\big[\|G\|+\|H\|\big]|(v,w)|,\\ &|c(v,w)|\leq\rho\|D\||(v,w)|\qquad((v,w)\in\mathbb{R}^\gamma\times\mathbb{R}^\delta).\end{aligned}\tag{3.5}$$

Moreover, if (iii) *holds, then ρ can be chosen such that $\rho\leq1/\tau$.*

Remark 3.1. As we shall soon see in the proof below, the equivalence (i) $\Longleftrightarrow$ (ii) $\Longleftrightarrow$ (iii) and the implication (iv) $\Longrightarrow$ (i) are straightforward. Note that the equivalence between (i) and (iv) could be obtained via an open-mapping theorem for convex processes (i.e., the Robinson–Ursescu theorem). However, the main contribution of Theorem 3.1 lies in the implication (iii) $\Longrightarrow$ (iv), and more specifically in the fact that the constant ρ turns out to be less than or equal to $1/\tau$, where τ is the constant in (3.3). This fact becomes crucial when applying the result of Theorem 3.1 to data that consist of essentially bounded matrix-valued functions.

Proof. (i) $\Longleftrightarrow$ (ii). Assume that (i) is true and let $\xi\in\mathbb{R}^\gamma$ and $\eta\in\mathbb{R}^\delta$ such that (3.2) holds. Let $v\in\mathbb{R}^\gamma$ and $w\in\mathbb{R}^\delta$ be arbitrary. By (i), there exist $a\in\mathbb{R}^m$ and $c\in C$ such that (3.1) holds. Multiplying these equations by ξ and η, respectively, we get

$$\xi^Tv+\eta^Tw=\xi^TGa-\xi^TDc+\eta^THa=-\xi^TDc\geq0.$$

Hence $\xi^Tv+\eta^Tw\geq0$ for all v and w. This implies that $\xi=0$ and $\eta=0$.

Conversely, assume that (ii) holds but (i) is not true. Then the set

$$K := \{(Ga - Dc, Ha) \mid a \in \mathbb{R}^m,\, c \in C\}$$

is a proper subcone of $\mathbb{R}^\gamma \times \mathbb{R}^\delta$. Thus, there exists $(\xi, \eta) \neq (0,0)$ such that $(\xi, \eta) \in -K^\circ$, that is,

$$\xi^T(Ga - Dc) + \eta^T Ha \geq 0$$

for all $a \in \mathbb{R}^m$, $c \in C$. This yields the fact that (3.2) is valid. Hence, by (ii), $\xi = 0$ and $\eta = 0$. The contradiction shows that (ii) implies (i).

(ii) $\iff$ (iii). If (ii) holds, then

$$\varphi(\xi, \eta) := |\xi^T G + \eta^T H|^2 + [\operatorname{dist}(\xi^T D, C^\circ)]^2 > 0$$

for all $(\xi, \eta) \neq (0,0)$. Hence, the infimum of φ on the unit sphere of $\mathbb{R}^\gamma \times \mathbb{R}^\delta$, which we denote by τ, is positive. Using quadratic homogeneity, the statement of (iii) follows. The reverse implication holds trivially.

Thus, we have obtained that conditions (i), (ii), and (iii) are equivalent.

(iii) $\Longrightarrow$ (iv). Putting $\xi = 0$ into (3.3), we get

$$|\eta^T H|^2 \geq \tau|\eta|^2, \qquad \text{i.e.,} \qquad \eta^T H H^T \eta \geq \tau|\eta|^2 \qquad (\eta \in \mathbb{R}^\delta).$$

Hence, HH^T is positive definite, invertible, and $\|(HH^T)^{-1}\| \leq \rho$, where $\rho := 1/\tau$.

Let $v \in \mathbb{R}^\gamma$ and $w \in \mathbb{R}^\delta$ be fixed arbitrarily. Using the equivalence of (i) and (iii), we can see that there exist $x \in \mathbb{R}^m$ and $y \in C$ such that $Gx - Dy = v$ and $Hx = w$. Thus, the following optimization problem has a unique solution (x, y):

$$\text{(3.6)} \qquad \frac{1}{2}\big(\|x\|^2 + \|y\|^2\big) \longrightarrow \min \qquad \text{w.r.t.} \qquad Gx - Dy = v, \quad Hx = w, \quad y \in C.$$

(The uniqueness follows from the strict convexity of the objective function.) Hence, there are multipliers $\lambda \geq 0$, $\xi \in \mathbb{R}^\gamma$, $\eta \in \mathbb{R}^\delta$, and $\zeta \in C^\circ$, not all zero, such that

$$\text{(3.7)} \qquad \lambda x^T + \xi^T G + \eta^T H = 0, \qquad \lambda y^T - \xi^T D + \zeta^T = 0, \qquad \zeta^T y = 0.$$

If λ were zero, then $\xi^T G + \eta^T H = 0$ and $\xi^T D \in C^\circ$, which, due to (3.3), yields $\xi = 0$, $\eta = 0$. Thus also $\zeta = 0$, which is a contradiction. Thus, we may assume that $\lambda = 1$. Then

$$\begin{aligned} -(\xi, \eta)^T(v, w) &= -\xi^T(Gx - Dy) - \eta^T Hx = -(\xi^T G + \eta^T H)x + (y^T + \zeta^T)y \\ &= |\xi^T G + \eta^T H|^2 + |y|^2 = |\xi^T G + \eta^T H|^2 + |D^T\xi - \zeta|^2 \\ &\geq |\xi^T G + \eta^T H|^2 + [\operatorname{dist}(\xi^T D, C^\circ)]^2 \geq \tau|(\xi, \eta)|^2. \end{aligned}$$

Using the Cauchy–Schwarz inequality, this yields

$$|(\xi, \eta)||(v, w)| \geq \tau|(\xi, \eta)|^2, \qquad \text{i.e.,} \qquad |(\xi, \eta)| \leq \rho|(v, w)|.$$

Hence

$$\text{(3.8)} \qquad |x| = |G^T\xi + H^T\eta| \leq \rho\big[\|G\| + \|H\|\big]|(v, w)|$$

and

$$\text{(3.9)} \qquad |y|^2 = y^T(D^T\xi - \zeta) = |y^T D^T \xi| \leq |y|\|D\||\xi| \qquad \Longrightarrow \qquad |y| \leq \|D\||\xi| \leq \rho\|D\||(v, w)|.$$

Define $a(v,w)$ and $c(v,w)$ (for fixed $v \in \mathbb{R}^\gamma$ and $w \in \mathbb{R}^\delta$) to be, respectively, the solutions x and y of the optimization problem (3.6). Then the feasibility of (x,y) yields (3.4); furthermore, the estimates (3.8) and (3.9) imply (3.5).

Finally, we note that the implication (iv) $\Longrightarrow$ (i) is obvious. Thus the proof of the theorem is complete. □

Now we apply the implication (iii) $\Longrightarrow$ (iv) of the above result to essentially bounded matrix functions G, H, and D, where C is the nonnegative orthant in $\mathbb{R}^q$. Below, $\mathcal{B}$ denotes the σ-algebra of Borel sets. The notation x^+ stands for the nonnegative part of a real number x, that is, $x^+ := \max(0,x)$.

THEOREM 3.2. *Let $(\Omega, \mathcal{A}, \nu)$ be a finite measure space, $G : \Omega \to \mathbb{R}^{\gamma\times m}$, $H : \Omega \to \mathbb{R}^{\delta\times m}$, and let $d_1, \dots, d_q : \Omega \to \mathbb{R}^\gamma$ be bounded measurable functions. Assume that there exists a constant $\tau > 0$ such that, for a.e. $t \in \Omega$,*

$$|\xi^T G(t) + \eta^T H(t)|^2 + \sum_{i=1}^{q} \Big[(\xi^T d_i(t))^+\Big]^2 \geq \tau |(\xi,\eta)|^2 \qquad ((\xi,\eta) \in \mathbb{R}^\gamma \times \mathbb{R}^\delta). \tag{3.10}$$

Then $HH^T : \Omega \to \mathbb{R}^{\delta\times\delta}$ has a bounded measurable inverse, and there exist $\mathcal{A}\times\mathcal{B}\times\mathcal{B}$-measurable maps $a : \Omega \times \mathbb{R}^\gamma \times \mathbb{R}^\delta \to \mathbb{R}^m$ and $c_1, \dots, c_q : \Omega \times \mathbb{R}^\gamma \times \mathbb{R}^\delta \to [0,\infty)$ and a constant $R > 0$ such that, for a.e. $t \in \Omega$,

$$G(t)a(t,v,w) = v + \sum_{i=1}^{q} c_i(t,v,w) d_i(t), \quad H(t)a(t,v,w) = w \qquad ((v,w) \in \mathbb{R}^\gamma \times \mathbb{R}^\delta), \tag{3.11}$$

and

$$|a(t,v,w)| \leq R|(v,w)|, \qquad c_i(t,v,w) \leq R|(v,w)| \qquad ((v,w) \in \mathbb{R}^\gamma \times \mathbb{R}^\delta,\ i = 1,\dots,q). \tag{3.12}$$

Proof. Set $\rho := 1/\tau$,

$$C := \mathbb{R}^q_+ = \{c = (c_1,\dots,c_q) \mid c_1,\dots,c_q \geq 0\},$$
$$\text{and} \qquad D(t) := (d_1(t),\dots,d_q(t)) \qquad (t \in \Omega).$$

Define the set-valued map Φ on $\Omega \times \mathbb{R}^\gamma \times \mathbb{R}^\delta$ by

$$\Phi(t,v,w) := \big\{(a,c) \in \mathbb{R}^m \times C : H(t)a = w,\ G(t)a - D(t)c = v,$$
$$|a| \leq \rho\big[\|G\|_\infty + \|H\|_\infty\big]|(v,w)|, |c| \leq \rho\|D\|_\infty |(v,w)|\big\}.$$

We show that, for a.e. $t \in \Omega$, for all $v \in \mathbb{R}^\gamma$, and for all $w \in \mathbb{R}^\delta$, the set $\Phi(t,v,w)$ is nonempty.

Without loss of generality, we may assume that

$$\|G(t)\| \leq \|G\|_\infty, \qquad \|H(t)\| \leq \|H\|_\infty, \qquad \|D(t)\| \leq \|D\|_\infty,$$

and (3.10) is valid for all $t \in \Omega$. (In fact, (3.10) is valid on a subset of Ω, which is of full measure and which we do not relabel.)

Since $C^\circ = \{(c_1,\dots,c_q) \mid c_1,\dots,c_q \leq 0\}$, then

$$\operatorname{dist}(x, C^\circ) = \sum_{i=1}^{q} (x_i^+)^2.$$

Thus (3.10) yields, for all $t \in \Omega$,

$$|\xi^T G(t) + \eta^T H(t)|^2 + [\operatorname{dist}(\xi^T D(t), C^\circ)]^2 \geq \tau |(\xi, \eta)|^2 \qquad ((\xi, \eta) \in \mathbb{R}^\gamma \times \mathbb{R}^\delta).$$

Whence, by Theorem 3.1 and Remark 3.1, for each $t \in \Omega$, $H(t)$ is of full rank, and there exist two mappings $a_t : \mathbb{R}^\gamma \times \mathbb{R}^\delta \to \mathbb{R}^m$ and $c_t : \mathbb{R}^\gamma \times \mathbb{R}^\delta \to \mathbb{R}^q$ such that

$$G(t)a_t(v,w) - D(t)c_t(v,w) = v, \quad H(t)a_t(v,w) = 0, \quad c_t(v,w) \in C \qquad ((v,w) \in \mathbb{R}^\gamma \times \mathbb{R}^\delta),$$

$$\|(H(t)H(t)^T)^{-1}\| \leq \rho, \tag{3.13}$$

and

$$\begin{aligned} &|a_t(v,w)| \leq \rho\big[\|G\|_\infty + \|H\|_\infty\big]|(v,w)|, \\ &|c_t(v,w)| \leq \rho\|D\|_\infty |(v,w)| \qquad ((v,w) \in \mathbb{R}^\gamma \times \mathbb{R}^\delta). \end{aligned} \tag{3.14}$$

Thus, with $a := a_t(v,w)$ and $c := c_t(v,w)$, we have that $(a,c) \in \Phi(t,v,w)$, whence the nonemptiness of $\Phi(t,v,w)$ follows.

Furthermore, Φ is $\mathcal{A} \times \mathcal{B} \times \mathcal{B}$-measurable with closed images. Hence, by the measurable selection theorem, there exists an $\mathcal{A} \times \mathcal{B} \times \mathcal{B}$-measurable function $(a,c) : \Omega \times \mathbb{R}^\gamma \times \mathbb{R}^\delta \to \mathbb{R}^m \times \mathbb{R}^q$ such that

$$(a(t,v,w), c(t,v,w)) \in \Phi(t,v,w) \qquad ((t,v,w) \in \Omega \times \mathbb{R}^\gamma \times \mathbb{R}^\delta).$$

Therefore, c has nonnegative components and together with a satisfies the relations (3.11) and (3.12), where $R := \rho \max(\|G\|_\infty + \|H\|_\infty, \|D\|_\infty)$.

Using (3.13), it follows that the function matrix–valued B defined by $B(t) = H^T(t)(H(t)H^T(t))^{-1}$ is an essentially bounded right inverse of H. □

4. Main results. We consider the optimal control problem

$$(\mathcal{CP}) \qquad \begin{array}{ll} \text{Minimize} & \ell(x(0), x(1)) \\ \text{subject to} & \left\{ \begin{array}{ll} \text{(i)} & a(x(0),x(1)) \in R, \\ \text{(ii)} & b(x(0),x(1)) = 0, \\ \text{(iii)} & \dot{x}(t) = f(t,x(t),u(t)) \text{ for a.e. } t \in [0,1], \\ \text{(iv)} & g(t,x(t),u(t)) \in Q(t) \text{ for a.e. } t \in [0,1], \\ \text{(v)} & h(t,x(t),u(t)) = 0 \text{ for a.e. } t \in [0,1], \\ \text{(vi)} & k(t,x(t)) \in S(t) \text{ for } t \in [0,1], \end{array} \right. \end{array}$$

where $x : [0,1] \to \mathbb{R}^n$ is absolutely continuous, $u : [0,1] \to \mathbb{R}^m$ is essentially bounded measurable, and the ranges of the functions ℓ, a, b, f, g, h, and k are, respectively, in $\mathbb{R}$, $\mathbb{R}^r$, $\mathbb{R}^s$, $\mathbb{R}^n$, $\mathbb{R}^\gamma$, $\mathbb{R}^\delta$, and $\mathbb{R}^\kappa$. Furthermore, R is a subset of $\mathbb{R}^r$, and Q and S are set-valued maps with images in $\mathbb{R}^\gamma$ and $\mathbb{R}^\kappa$.

The set-valued maps Q and S will be assumed in (R_5) to take *convex* values, while *no* convexity is imposed on the functions g and k. Hence, the forms of the constraints (iv) and (vi) considered here are more general than the traditional forms: $u(t) \in Q(t)$ and $x(t) \in S(t)$. Indeed, the present constraints permit us to consider, for instance, inequality constraints $g(t,x(t),u(t)) \leq 0$ and $k(t,x(t)) \leq 0$ without any convexity assumptions on the functions g and k.

The Hamiltonian function associated to $(\mathcal{CP})$ is

$$\mathcal{H}(t,x,u,p,\varphi,\psi) := p^T f(t,x,u) + \varphi^T g(t,x,u) + \psi^T h(t,x,u).$$

If (x,u) satisfy (i)–(vi), then it is said to be *admissible* for $(\mathcal{CP})$. Given an admissible arc $(\widehat{x},\widehat{u})$, we denote by $\widehat{F}$ the evaluation of a given function F along $(\widehat{x},\widehat{u})$. For instance, $\widehat{a} := a(\widehat{x}(0),\widehat{x}(1))$ and $\widehat{g}$ is defined by $\widehat{g}(t) := g(t,\widehat{x}(t),\widehat{u}(t))$.

To formulate the optimality concept and the regularity assumptions for problem $(\mathcal{CP})$, introduce the following notion: If T is a subset of $[0,1]$ and $\widehat{w} : [0,1] \to \mathbb{R}^\omega$ is an arbitrary function, then the *ε-tube* on T around $\widehat{w}$ is the set

$$\mathcal{T}_\varepsilon(\widehat{w};T) := \{(t,w) \in T \times \mathbb{R}^\omega \mid |w - \widehat{w}(t)| < \varepsilon \text{ for } t \in T\}.$$

When $T = \{t\}$ is a singleton, then $\{w \mid (t,w) \in \mathcal{T}_\varepsilon(\widehat{w};\{t\})\}$ will be denoted by $\mathcal{T}_\varepsilon(\widehat{w}(t))$.

A pair $(\widehat{x},\widehat{u})$ provides a *weak-local minimum* for $(\mathcal{CP})$ if there exists an $\varepsilon > 0$ such that for all admissible pairs $(x,u) \in \mathcal{T}_\varepsilon(\widehat{x},\widehat{u};[0,1])$, we have $\ell(x(0),x(1)) \geq \ell(\widehat{x}(0),\widehat{x}(1))$.

In [OS95] and [MOS98] optimality conditions for the Pontryagin minimum were obtained in the absence of pure-state constraints and when the mixed state-control constraints take the form of equality and *inequality*.

Denote by $\mathcal{L}$ the class of Lebesgue-measurable subsets in $[0,1]$, and by $\mathcal{B}$ the class of Borel-measurable subsets in a metric space.

A pair $(\widehat{x},\widehat{u})$ is called *regular* for $(\mathcal{CP})$ if there exists an $\varepsilon > 0$ such that the following conditions are satisfied:

(R_1) The functions ℓ, a, b are defined on $\mathcal{T}_\varepsilon(\widehat{x};\{0,1\})$ and are strictly Fréchet differentiable at the point $(\widehat{x}(0),\widehat{x}(1))$.

(R_2) The functions f, g, h are defined on $\mathcal{T}_\varepsilon(\widehat{x},\widehat{u};[0,1])$, are $\mathcal{L}\times\mathcal{B}\times\mathcal{B}$-measurable, and the maps

$$\begin{aligned} &(x,u) \mapsto f(t,x,u) \\ \text{and} \quad &(x,u) \mapsto (g(t,x,u),h(t,x,u)) \qquad \big((x,u) \in \mathcal{T}_\varepsilon(\widehat{x}(t),\widehat{u}(t))\big) \end{aligned} \tag{4.1}$$

are strictly Fréchet differentiable at the point $(\widehat{x}(t),\widehat{u}(t))$, $\mathcal{L}^1$-uniformly and $\mathcal{L}^\infty$-uniformly, respectively, for a.e. $t \in [0,1]$. Furthermore, it is also assumed that $\widehat{f}$, $\widehat{f}_x$, and $\widehat{f}_u$ are integrable functions, and $\widehat{g}$, $\widehat{h}$, $\widehat{g}_x$, $\widehat{h}_x$, $\widehat{g}_u$, and $\widehat{h}_u$ are essentially bounded measurable functions.

(R_3) The functions g and h satisfy the following strong normality condition: There exist a constant $\tau > 0$ and bounded measurable functions $d_1,\dots,d_q \in T(\widehat{g}|\operatorname{sel}_\infty(Q))$ such that, for a.e. $t \in [0,1]$,

$$|\xi^T\widehat{g}_u(t) + \eta^T\widehat{h}_u(t)|^2 + \sum_{i=1}^{q}\Big[(\xi^T d_i(t))^+\Big]^2 \geq \tau|(\xi,\eta)|^2 \qquad ((\xi,\eta) \in \mathbb{R}^\gamma \times \mathbb{R}^\delta). \tag{4.2}$$

(R_4) The function k defined on $\mathcal{T}_\varepsilon(\widehat{x};[0,1])$ is Borel-measurable, and the map

$$x \mapsto k(t,x) \qquad \big(x \in \mathcal{T}_\varepsilon(\widehat{x}(t))\big) \tag{4.3}$$

is strictly Fréchet differentiable at the point $\widehat{x}(t)$ uniformly in $t \in [0,1]$. Furthermore, it is also assumed that $\widehat{k}$ and $\widehat{k}_x$ are continuous functions.

(R_5) The set $R \subset \mathbb{R}^r$ is closed convex and has nonempty interior; the set-valued maps $Q : [0,1] \to 2^{\mathbb{R}^\gamma}$ and $S : [0,1] \to 2^{\mathbb{R}^\kappa}$ take closed convex values with nonempty interior and are measurable and lower semicontinuous, respectively. Moreover, Q also satisfies condition (2.6).

We note that, in (R_3), a sufficient condition in order that $d_1,\dots,d_q \in T(\widehat{g}|\operatorname{sel}_\infty(Q))$ be valid is that $d_1,\dots,d_q \in C(\widehat{g}|\operatorname{sel}_\infty(Q))$ be satisfied. This latter condition holds if and only if $d_1(t),\dots,d_q(t) \in C(\widehat{g}(t)|Q(t))$ almost uniformly in t, that is, if there exists a constant $M>0$ such that, for all $i=1,\dots,q$ and for a.e. t,

$$\big[\xi^T d_i(t)\big]^2 \le M|\xi|\big(\delta^*(\xi|Q(t)) - \xi^T\widehat{g}(t)\big) \tag{4.4}$$

whenever $\xi\in\mathbb{R}^\gamma$ satisfies $\xi^T d_i(t)>0$.

A pair $(\delta x,\delta u)$ is said to be *critical* for $(\mathcal{CP})$ at $(\widehat{x},\widehat{u})$ if $\delta x:[0,1]\to\mathbb{R}^n$ is absolutely continuous, $\delta u:[0,1]\to\mathbb{R}^m$ is essentially bounded measurable, and

- (C_1) $\widehat{\ell}_{x_0}\delta x(0)+\widehat{\ell}_{x_1}\delta x(1)\le 0$;
- (C_2) $\widehat{a}_{x_0}\delta x(0)+\widehat{a}_{x_1}\delta x(1)\in C(\widehat{a}|R)$;
- (C_3) $\widehat{b}_{x_0}\delta x(0)+\widehat{b}_{x_1}\delta x(1)=0$;
- (C_4) $\dot{\delta x}(t)=\widehat{f}_x(t)\delta x(t)+\widehat{f}_u(t)\delta u(t)$ holds for a.e. $t\in[0,1]$;
- (C_5) $\widehat{g}_x(t)\delta x(t)+\widehat{g}_u(t)\delta u(t)\in C(\widehat{g}(t)|Q(t))$ almost uniformly in $t\in[0,1]$; that is, there exists a constant $M>0$ such that, for a.e. $t\in[0,1]$,

$$\big[\xi^T\widehat{g}_x(t)\delta x(t)+\xi^T\widehat{g}_u(t)\delta u(t)\big]^2 \le M|\xi|\big(\delta^*(\xi|Q(t)) - \xi^T\widehat{g}(t)\big) \tag{4.5}$$

  whenever $\xi\in\mathbb{R}^\gamma$ satisfies $\xi^T\widehat{g}_x(t)\delta x(t)+\xi^T\widehat{g}_u(t)\delta u(t)>0$;
- (C_6) $\widehat{h}_x(t)\delta x(t)+\widehat{h}_u(t)\delta u(t)=0$ holds for a.e. $t\in[0,1]$;
- (C_7) $\widehat{k}_x(t)\delta x(t)\in C(\widehat{k}(t)|S(t))$ uniformly in $t\in[0,1]$; that is, there exists a constant $M>0$ such that, for all $t\in[0,1]$,

$$\big[\zeta^T\widehat{k}_x(t)\delta x(t)\big]^2 \le M|\zeta|\big(\delta^*(\zeta|S(t)) - \zeta^T\widehat{k}(t)\big) \tag{4.6}$$

  whenever $\zeta\in\mathbb{R}^\kappa$ satisfies $\zeta^T\widehat{k}_x(t)\delta x(t)>0$.

A critical arc $(\delta x,\delta u)$ is called *regular* for $(\mathcal{CP})$ at $(\widehat{x},\widehat{u})$ if

- (R_6) $\widehat{\ell}$, $\widehat{a}$, and $\widehat{b}$ are twice directionally differentiable at $(\widehat{x}(0),\widehat{x}(1))$ in direction $(\delta x(0),\delta x(1))$;
- (R_7) for a.e. $t\in[0,1]$, the maps in (4.1) are twice directionally differentiable at $(\widehat{x}(t),\widehat{u}(t))$ in direction $(\delta x(t),\delta u(t))$ $\mathcal{L}^1$- and $\mathcal{L}^\infty$-uniformly in t, respectively;
- (R_8) for all $t\in[0,1]$, the map (4.3) is twice directionally differentiable at $\widehat{x}(t)$ in direction $\delta x(t)$ uniformly in t.

The following result consists of necessary conditions for optimality in $(\mathcal{CP})$. Its proof makes use of all the results of sections 2 and 3 and applies the argument followed in [PZ94b].

THEOREM 4.1. *Let $(\widehat{x},\widehat{u})$ be a regular weak local minimum for the problem $(\mathcal{CP})$. Then, for every regular critical arc $(\delta x,\delta u)$, there exist constants $\lambda\in\mathbb{R}$, $\alpha=(\alpha_1,\dots,\alpha_r)\in\mathbb{R}^r$, $\beta=(\beta_1,\dots,\beta_s)\in\mathbb{R}^s$, an absolutely continuous function $p:[0,1]\to\mathbb{R}^n$, two integrable functions $\varphi:[0,1]\to\mathbb{R}^\gamma$ and $\psi:[0,1]\to\mathbb{R}^\delta$, and a Borel regular vector-valued measure $\mu=(\mu_1,\dots,\mu_\kappa)$, not all zero, such that $\lambda\ge 0$,*

$$\alpha\in N(\widehat{a}|R),\qquad \alpha^T(\widehat{a}_{x_0}\delta x(0)+\widehat{a}_{x_1}\delta x(1))=0, \tag{4.7}$$

(4.8)
$$\varphi(t)\in N(\widehat{g}(t)|Q(t)),\qquad \varphi^T(t)(\widehat{g}_x(t)\delta x(t)+\widehat{g}_u\delta u(t))=0 \quad \text{for a.e. } t\in[0,1],$$

$$\frac{d\mu}{d|\mu|}(t)\in N(\widehat{k}(t)|S(t)),\qquad \Big(\frac{d\mu}{d|\mu|}\Big)^T(t)\,\widehat{k}_x(t)\delta x(t)=0 \quad \text{for } \mu\text{-a.e. } t\in[0,1], \tag{4.9}$$

$$\dot{p}^T(t) = -\widehat{\mathcal{H}}_x\Big(t, p(t) + \int_{]t,1]} \widehat{k}_x^T(s)d\mu(s), \varphi(t), \psi(t)\Big) \qquad \textit{for a.e. } t \in [0,1], \tag{4.10}$$

$$-p^T(0) = \lambda\widehat{\ell}_{x_0} + \alpha^T\widehat{a}_{x_0} + \beta^T\widehat{b}_{x_0} + \Big(\int_{[0,1]} \widehat{k}_x^T(t)d\mu(t)\Big)^T, \tag{4.11}$$

$$p^T(1) = \lambda\widehat{\ell}_{x_1} + \alpha^T\widehat{a}_{x_1} + \beta^T\widehat{b}_{x_1}, \tag{4.12}$$

$$\widehat{\mathcal{H}}_u\Big(t, p(t) + \int_{]t,1]} \widehat{k}_x^T(s)d\mu(s), \varphi(t), \psi(t)\Big) = 0 \qquad \textit{for a.e. } t \in [0,1], \tag{4.13}$$

and

(4.14)

$$\begin{aligned}
&\big(\lambda\widehat{\ell}'' + \alpha^T\widehat{a}'' + \beta^T\widehat{b}''\big)(\delta x(0), \delta x(1)) + \int_0^1 \widehat{k}''(t; \delta x(t))d\mu(t) \\
&\qquad + \int_0^1 \widehat{\mathcal{H}}''\Big(t, p(t) + \int_{]t,1]} \widehat{k}_x^T(s)d\mu(s), \varphi(t), \psi(t); \delta x(t), \delta u(t)\Big)dt \\
&\geq 2\,\overline{\mathrm{co}}\,\boldsymbol{\mathcal{E}}\big(\widehat{a}, \widehat{a}'(\delta x(0), \delta x(1))|R\big)(\alpha) + 2\int_0^1 \overline{\mathrm{co}}\,\mathbb{E}(\widehat{k}, \widehat{k}_x\delta x|S)\Big(t, \frac{d\mu}{d|\mu|}(t)\Big)d|\mu|(t) \\
&\qquad + 2\int_0^1 \overline{\mathrm{co}}\,\boldsymbol{\mathcal{E}}\big(\widehat{g}(t), \widehat{g}_x(t)\delta x(t) + \widehat{g}_u(t)\delta u(t)\big|Q(t)\big)(\varphi(t))dt,
\end{aligned}$$

where $\mathcal{H}''$ denotes the second-order strong directional derivative of $\mathcal{H}$ with respect to the variable (x, u).

Proof. First we are going to apply the result of Theorem 2.5, which is a special case of [PZ94b, Theorem 3]. Introduce the following spaces

$$X := \mathcal{C}(\mathbb{R}^n), \qquad U := \mathcal{L}_m^\infty := \mathcal{L}^\infty(\mathbb{R}^m), \qquad Y := \mathbb{R}^s \times \{y \in \mathcal{C}(\mathbb{R}^n) \mid y(0) = 0\},$$

$$V := \mathbb{R}^r \times \mathcal{C}(\mathbb{R}^\kappa) \times \mathcal{L}_\gamma^\infty, \qquad W := \mathcal{L}_\delta^\infty$$

(where we suppress $[0,1]$ in this notation) and denote, for $(x, u) \in X \times U$,

$$\begin{aligned}
\mathbf{F}(x,u) &:= \ell(x(0), x(1)), \\
\mathbf{G}(x,u)(t) &:= \begin{pmatrix} a(x(0), x(1)) \\ k(t, x(t)) \\ g(t, x(t), u(t)) \end{pmatrix}, \\
\mathbf{H}(x,u)(t) &:= h(t, x(t), u(t)), \\
\mathbf{K}(x,u)(t) &:= \begin{pmatrix} b(x(0), x(1)) \\ \int_0^t (f(\tau, x(\tau), u(\tau))d\tau - x(t) + x(0), \end{pmatrix}, \\
\mathbf{Q} &:= R \times \mathrm{sel}_C(S) \times \mathrm{sel}_\infty(Q).
\end{aligned}$$

Then, with this notation, our control problem $(\mathcal{CP})$ is equivalent to the abstract control problem **(P)** in section 2.

The regularity condition (R_5) yields that the set $\mathbf{Q}$ defined above is closed, convex, with nonempty interior.

Let $\varepsilon > 0$ be the constant for which the regularity assumptions of the arc $(\widehat{x}, \widehat{u})$ are satisfied and define

$$D := \{(x,u) \in X \times U \mid \|x - \widehat{x}\| < \varepsilon,\ \|u - \widehat{u}\|_\infty < \varepsilon\}.$$

Then D is an open subset of $X \times U$, and the functions $\mathbf{F}$, $\mathbf{G}$, $\mathbf{H}$, $\mathbf{K}$ are defined on D. Since the arc $(\widehat{x}, \widehat{u})$ satisfies the regularity conditions (R_1), (R_2), and (R_4) for $(\mathcal{CP})$, the functions $\mathbf{F}$, $\mathbf{G}$, $\mathbf{H}$, and $\mathbf{K}$ are strictly Fréchet differentiable at $(\widehat{x}, \widehat{u})$, whence we have the following relations:

$$\widehat{\mathbf{H}}'(x,u)(t) = \widehat{H}_x(t)x(t) + \widehat{H}_u(t)u(t)$$

and

$$\widehat{\mathbf{K}}_u(x,u)(t) := \begin{pmatrix} 0 \\ \int_0^t (\widehat{f}_u(\tau)u(\tau)d\tau \end{pmatrix},$$

$$\widehat{\mathbf{K}}_x(x,u)(t) := \begin{pmatrix} \widehat{b}(x(0),x(1)) \\ \int_0^t (\widehat{f}_x(\tau)x(\tau)d\tau - x(t) + x(0) \end{pmatrix}.$$

We need to show that the partial Fréchet derivatives $\widehat{\mathbf{K}}_x$ and $\widehat{\mathbf{K}}_u$ of the mapping $\mathbf{K}$ are Fredholm and compact operators, respectively. Since $\mathbf{K}_u$ is a Volterra integral operator, it is compact. On the other hand, the operator $\mathbf{K}_x$ is the sum of a compact (Volterra integral) operator and the operator $F : X \to Y$ defined by $Fx(t) := -x(t) + x(a)$, which is clearly a Fredholm operator. Therefore, by [PZ94b, Lemmas 3 and 5], $\mathbf{K}_x$ is also Fredholm. Thus, for $(\mathbf{R}_1)$–$(\mathbf{R}_3)$ to hold, it remains to show that $\widehat{\mathbf{H}}_u$ has a bounded right inverse. The strong normality condition, i.e., (R_3), and Theorem 3.2 yield that the function $\widehat{h}_u\widehat{h}_u^T : [0,1] \to \mathbb{R}^\delta$ has a bounded measurable inverse, and hence the linear operator $\mathbf{B} : W \to U$ defined by

$$(\mathbf{B}w)(t) := \widehat{h}_u^T(t)\big(\widehat{h}_u(t)\widehat{h}_u^T(t)\big)^{-1}w(t) \qquad (t \in [0,1])$$

is a bounded linear right inverse for $\mathbf{H}_u(\widehat{x}, \widehat{u})$.

Hence, the arc $(\widehat{x}, \widehat{u})$ is a regular arc with respect to the problem **(P)**.

Now we prove that the pair $(\delta x, \delta u)$ is a regular and critical arc for **(P)** at the point $(\widehat{x}, \widehat{u})$ where $\mathbf{F}$, $\mathbf{G}$, $\mathbf{H}$, and $\mathbf{K}$ are defined above. The regularity assumptions (R_6)–(R_8) imposed on $(\delta x, \delta u)$ yield that the functions $\mathbf{F}$, $\mathbf{G}$, $\mathbf{H}$, and $\mathbf{K}$ are twice directionally differentiable at $(\widehat{x}, \widehat{u})$ in the direction $(\delta x, \delta u)$; that is, $(\mathbf{R}_4)$ holds. Note that (C_1) implies $(\mathbf{C}_1)$. Using (C_7) together with Theorem 2.1, (C_5) together with Theorem 2.2, and (C_2), then applying the product rule, we can see that

$$\widehat{\mathbf{G}}_x\delta x + \widehat{\mathbf{G}}_u\delta u \in C(\widehat{\mathbf{G}}|\mathbf{Q});$$

that is, the second-order variation set $V(\widehat{\mathbf{G}}, \widehat{\mathbf{G}}_x\delta x + \widehat{\mathbf{G}}_u\delta u|\mathbf{Q})$ is nonempty. Furthermore, (C_3), (C_4), and (C_6) yield that

$$\widehat{\mathbf{H}}_x\delta x + \widehat{\mathbf{H}}_u\delta u = 0, \qquad \widehat{\mathbf{K}}_x\delta x + \widehat{\mathbf{K}}_u\delta u = 0.$$

Then $(\mathbf{C}_2)$ is satisfied, and hence $(\delta x, \delta u)$ is a regular and critical arc for **(P)** at the point $(\widehat{x}, \widehat{u})$.

Therefore, the statement of Theorem 2.5 can also be applied to produce multipliers $\lambda \geq 0$, $v^* = (v_1^*, v_2^*, v_3^*) \in V^*$, $w^* \in W^*$, and $y^* = (y_1^*, y_2^*) \in Y^*$, not all zero, satisfying (2.16)–(2.19). The first and second components v_1^* and v_2^* of v^* can be identified, respectively, by a vector $\alpha \in \mathbb{R}^r$ and (due to the Riesz representation theorem) by a bounded signed $\mathbb{R}^\kappa$-valued Borel measure μ. For the third component, we have $v_3^* \in \left(\mathcal{L}_\gamma^\infty\right)^*$. Then (2.16) yields that

$$(\alpha, \mu, v_3^*) \in N(\widehat{\mathbf{G}}|\mathbf{Q}) = N(\widehat{a}|R) \times N(\widehat{k}|\operatorname{sel}_C(S)) \times N(\widehat{g}|\operatorname{sel}_\infty(Q))$$

and

$$\alpha^T(\widehat{a}_{x_0}\delta x(0) + \widehat{a}_{x_1}\delta x(1)) + \int_{[0,1]} \left(\widehat{k}_x(t)\delta x(t)\right)^T d\mu(t) + \left\langle v_3^*, \widehat{g}_x\delta x + \widehat{g}_u\delta u\right\rangle = 0. \tag{4.15}$$

Therefore, we get that the first equation of (4.7) is valid and that $\mu \in N(\widehat{k}|\operatorname{sel}_C(S))$, which, via (2.2), yields the first equation of (4.9); furthermore,

$$v_3^* \in N(\widehat{g}|\operatorname{sel}_\infty(Q)). \tag{4.16}$$

The first component y_1^* of y^* can be identified by an element $\beta \in \mathbb{R}^s$, and, by the Riesz representation theorem, there exists a bounded signed $\mathbb{R}^n$–valued measure ν with $\nu(\{0\}) = 0$ such that y_2^* is represented via ν; that is, for $y \in \mathcal{C}(\mathbb{R}^n)$ with $y(0) = 0$, we have

$$\langle y_2^*, y\rangle = \int_{[0,1]} y^T(t)d\nu(t).$$

Define $\bar{p} : [0,1] \to \mathbb{R}^n$ by

$$\bar{p}(t) = \nu(]t, 1]).$$

Clearly, $\bar{p}(1) = 0$, and $\bar{p}$ is of bounded variation (and hence it is also bounded). Then, by standard argument (see, e.g., [PZ94b, p. 441]), we get that, for $x \in \mathcal{C}(\mathbb{R}^n)$ and $u \in \mathcal{L}_m^\infty$,

$$\langle y^*, \widehat{\mathbf{K}}_x x\rangle = \beta^T\left(\widehat{b}_{x_0}x(0) + \widehat{b}_{x_1}x(1)\right) + \int_0^1 \bar{p}^T(t)\widehat{f}_x(t)x(t)dt - \int_0^1 x^T(t)d\nu(t) + \bar{p}^T(0)x(0), \tag{4.17}$$

$$\langle y^*, \widehat{\mathbf{K}}_u u\rangle = \int_0^1 \bar{p}^T(t)\widehat{f}_u(t)u(t)dt, \tag{4.18}$$

and

$$\langle y^*, \widehat{\mathbf{K}}''(\delta x, \delta u)\rangle = \beta^T\widehat{b}''\left(\delta x(0) + \delta x(1)\right) + \int_0^1 \bar{p}^T(t)\widehat{f}''(t; \delta x(t), \delta u(t))dt. \tag{4.19}$$

Using (4.18), equation (2.18) reduces to

$$\langle v_3^*, \widehat{g}_u u\rangle + \langle w^*, \widehat{h}_u u\rangle + \int_0^1 \bar{p}^T(t)\widehat{f}_u(t)u(t)dt = 0 \qquad (u \in \mathcal{L}_m^\infty). \tag{4.20}$$

We are going to show that v_3^* and w are represented via integrable functions. To achieve this goal, we shall apply Theorem 3.2. Observe that condition (4.2) is equivalent to (3.10), where G and H are replaced by $\widehat{g}_u$ and $\widehat{h}_u$, respectively, and $d_1, \dots, d_q$ are the functions of hypothesis (R_3) of the theorem. Thus, by Theorem 3.2, $\widehat{h}_u \widehat{h}_u^T$ has a bounded measurable inverse and there exist $\mathcal{L}\times\mathcal{B}\times\mathcal{B}$-measurable functions $a : [0,1] \times \mathbb{R}^\gamma \times \mathbb{R}^\delta \to \mathbb{R}^m$, $c_1, \dots, c_q : [0,1] \times \mathbb{R}^\gamma \times \mathbb{R}^\delta \to [0, \infty[$ and a constant $R > 0$ such that, for a.e. $t \in [0,1]$,

(4.21)
$$\widehat{g}_u(t)a(t,v,w) = v + \sum_{i=1}^{q} c_i(t,v,w)d_i(t), \quad \widehat{h}_u(t)a(t,v,w) = w \qquad ((v,w) \in \mathbb{R}^\gamma \times \mathbb{R}^\delta),$$

and

(4.22)
$$|a(t,v,w)| \le R|(v,w)|, \qquad c_i(t,v,w) \le R|(v,w)| \qquad ((v,w) \in \mathbb{R}^\gamma \times \mathbb{R}^\delta,\ i = 1, \dots, q).$$

Let $(v,w) \in \mathcal{L}_\gamma^\infty \times \mathcal{L}_\delta^\infty$ be fixed. Set

$$\begin{aligned} \mathbb{A}(v,w)(t) &:= \boldsymbol{a}(t) := a(t, v(t), w(t)), \\ \boldsymbol{c}_i(t) &:= c_i(t, v(t), w(t)) \qquad (t \in [0,1],\ i = 1, \dots, q). \end{aligned}$$

Then, due to the second inequality in (4.22), $\boldsymbol{a}$ and $\boldsymbol{c}_i$ are bounded measurable functions. Thus $\mathbb{A}$ is defined on $\mathcal{L}_\gamma^\infty \times \mathcal{L}_\delta^\infty$ with a range in $\mathcal{L}_m^\infty$. Since $\operatorname{sel}_\infty(Q)$ is decomposable, $T(\widehat{g}|\operatorname{sel}_\infty(Q))$ is an $\mathcal{L}$-cone, and we have that

$$\sum_{i=1}^{q} \boldsymbol{c}_i d_i \in T(\widehat{g}|\operatorname{sel}_\infty(Q)).$$

Hence, by (4.21),

$$\widehat{g}_u\mathbb{A}(v,w) - v \in T(\widehat{g}|\operatorname{sel}_\infty(Q)), \qquad \widehat{h}_u\mathbb{A}(v,w) = w \qquad ((v,w) \in \mathcal{L}_\gamma^\infty \times \mathcal{L}_\delta^\infty). \tag{4.23}$$

Using (4.16), the first inclusion in (4.23) yields that

$$\langle v_3^*, \widehat{g}_u\mathbb{A}(v,w)\rangle \le \langle v_3^*, v\rangle \qquad ((v,w) \in \mathcal{L}_\gamma^\infty \times \mathcal{L}_\delta^\infty).$$

Now, substituting $u = \mathbb{A}(v,w)$ into (4.20) and using also the second relation in (4.23), we get that

(4.24)
$$\langle v_3^*, v\rangle + \langle w^*, w\rangle + \int_0^1 \bar{p}^T(t)\widehat{f}_u(t)a(t, v(t), w(t))dt \ge 0 \qquad ((v,w) \in \mathcal{L}_\gamma^\infty \times \mathcal{L}_\delta^\infty).$$

Putting $w = 0$, we deduce that

$$\big|\langle v_3^*, v\rangle\big| \le \Big|\int_0^1 \bar{p}(t)\widehat{f}_u(t)a(t, v(t), 0)dt\Big| \le \int_0^1 |\bar{p}(t)||\widehat{f}_u(t)|R|v(t)|dt \qquad (v \in \mathcal{L}_\gamma^\infty).$$

Hence, v_3^* is $\mathcal{L}^1$-continuous; i.e., for any bounded (in $\mathcal{L}_\gamma^\infty$) sequence (v_i) that converges almost everywhere to zero, we have that $\langle v_3^*, v_i\rangle$ tends to zero. Therefore, by the

Yosida–Hewitt representation theorem, there exists an integrable function $\varphi : [0,1] \to \mathbb{R}^\gamma$ such that

$$\langle v_3^*, v\rangle = \int_0^1 \varphi^T(t)v(t)dt \qquad (v \in \mathcal{L}_\gamma^\infty). \tag{4.25}$$

Arguing analogously for w^*, (4.24) also yields the existence of an integrable function $\psi : [0,1] \to \mathbb{R}^\delta$ such that

$$\langle w^*, w\rangle = \int_0^1 \psi^T(t)w(t)dt \qquad (w \in \mathcal{L}_\delta^\infty). \tag{4.26}$$

By (2.5), we have that (4.16) is equivalent to the first equation of (4.8). Furthermore, (4.15) and conditions (C_2), (C_5), and (C_7) combined with the first equations of (4.7)–(4.9) yield the second equations of (4.7)–(4.9).

Using the representations of v_3^* and w^* and (4.17), equation (2.17) can be rewritten in the following way:

For all $x \in \mathcal{C}(\mathbb{R}^n)$,

(4.27)
$$\begin{aligned}&\big(\lambda\widehat{l}_{x_0} + \alpha^T\widehat{a}_{x_0} + \beta^T\widehat{b}_{x_0} + \bar{p}^T(0)\big)x(0) + \big(\lambda\widehat{l}_{x_1} + \alpha^T\widehat{a}_{x_1} + \beta^T\widehat{b}_{x_1}\big)x(1) - \int_0^1 x^T(t)d\nu(t)\\ &\quad + \int_0^1 x^T(t)\widehat{k}_x^T(t)d\mu(t) + \int_0^1 \big[\varphi^T(t)\widehat{g}_x(t) + \psi^T(t)\widehat{h}_x(t) + \bar{p}^T(t)\widehat{f}_x(t)\big]x(t)dt = 0.\end{aligned}$$

Set

$$p(t) := \begin{cases} \bar{p}(t) - \displaystyle\int_{]t,1]} \widehat{k}_x^T(t)d\mu(t) & \text{for } t \in [0,1[,\\ \displaystyle\lim_{t\to 1-} p(t) = \nu(\{1\}) - \widehat{k}_x^T(1)\mu(\{1\}) & \text{for } t = 1.\end{cases}$$

Observe that (4.27) is also true for all functions x of the form $x(t) = \bar{x}\chi_\Omega(t)$, where Ω is a subinterval of $[0,1]$ and $\bar{x} \in \mathbb{R}^n$ is arbitrary.

First taking $\Omega := \{1\}$, it follows from (4.27) that

$$\Big(\lambda\widehat{l}_{x_1} + \alpha^T\widehat{a}_{x_1} + \beta^T\widehat{b}_{x_1}\Big)^T - \nu(\{1\}) + \widehat{k}_x^T(1)\mu(\{1\}) = 0;$$

hence

$$p(1) = \nu(\{1\}) - \widehat{k}_x^T(1)\mu(\{1\}) = \Big(\lambda\widehat{l}_{x_1} + \alpha^T\widehat{a}_{x_1} + \beta^T\widehat{b}_{x_1}\Big)^T,$$

which is exactly (4.12).

With the substitution $x(t) := \bar{x}\chi_{\{0\}}(t)$ $(\bar{x} \in \mathbb{R}^n)$, we deduce from (4.27) that

$$\lambda\widehat{l}_{x_0} + \alpha^T\widehat{a}_{x_0} + \beta^T\widehat{b}_{x_0} + p^T(0) + \Big(\int_{]0,1]} \widehat{k}_x^T(t)d\mu(t) + \widehat{k}_x^T(0)\mu(\{0\})\Big)^T = 0$$

since $\nu(\{0\}) = 0$. This yields (4.11).

Finally, we put $x(t) := \bar{x}\chi_{]\tau,1]}(t)$ into (4.27), where $\bar{x} \in \mathbb{R}^n$ and $\tau \in [0,1[$ are fixed arbitrarily. Then we obtain, for all $\tau \in [0,1[$, that

$$\begin{aligned}&\lambda\widehat{l}_{x_1} + \alpha^T\widehat{a}_{x_1} + \beta^T\widehat{b}_{x_1} + \Big(\int_{]\tau,1]} \widehat{k}_x^T(t)d\mu(t) - \nu(]\tau,1])\Big)^T\\ &\qquad + \int_\tau^1 \big[\varphi^T(t)\widehat{g}_x(t) + \psi^T(t)\widehat{h}_x(t) + \bar{p}^T(t)\widehat{f}_x(t)\big]dt = 0.\end{aligned}$$

Using (4.12) and the definitions of p, $\bar{p}$, and the Hamiltonian $\mathcal{H}$, the above equation can be rewritten as

$$p^T(\tau) = p^T(1) + \int_\tau^1 \mathcal{H}_x\Big(t, p(t) + \int_{]t,1]} \widehat{k}_x^T(s)d\mu(s), \varphi(t), \psi(t)\Big)dt \qquad (\tau \in [0,1[).$$

It follows from the above equation that p is absolutely continuous, and after differentiation, we obtain (4.10).

Now we consider (2.18). Using (4.18), (4.25), and (4.26), equation (2.18) can be rewritten as

$$\int_0^1 \big[\varphi^T(t)\widehat{g}_u(t) + \psi^T(t)\widehat{h}_u(t) + \bar{p}^T(t)\widehat{f}_u(t)\big]u(t)dt = 0 \qquad (u \in \mathcal{L}_m^\infty).$$

This is equivalent to

$$\int_0^1 \mathcal{H}_u\Big(t, p(t) + \int_{]t,1]} \widehat{k}_x^T(s)d\mu(s), \varphi(t), \psi(t)\Big)u(t)dt = 0 \qquad (u \in \mathcal{L}_m^\infty).$$

By a standard argument, the above equation yields (4.13).

Now, (2.19) becomes

$$\begin{aligned}
&\big(\lambda\widehat{l}'' + \lambda^T\widehat{a}'' + \beta^T\widehat{b}''\big)(\delta x(0), \delta x(1)) + \int_0^1 \widehat{k}''(t;\delta x(t))d\mu(t) \\
&\qquad + \int_0^1 \mathcal{H}''\Big(t, p(t) + \int_{]t,1]} \widehat{k}_x^T(s)d\mu(s), \varphi(t), \psi(t); \delta x(t), \delta u(t)\Big)dt \\
&\geq 2\delta^*\big(v^*\big|V(\widehat{\mathbf{G}}, \widehat{\mathbf{G}}_x\delta x + \widehat{\mathbf{G}}_u\delta u|\mathbf{Q})\big) \\
&= 2\delta^*\big(\alpha\big|V(\widehat{a}, \widehat{a}_{x_0}\delta x(0) + \widehat{a}_{x_1}\delta x(1)|R)\big) + 2\delta^*\big(\mu\big|V(\widehat{k}, \widehat{k}_x\delta x|\operatorname{sel}_C(S))\big) \\
&\qquad + 2\delta^*\big(v_3^*\big|V(\widehat{g}, \widehat{g}_x\delta x + \widehat{g}_u\delta u|\operatorname{sel}_\infty(Q))\big)
\end{aligned}$$

since $\mathbf{Q}$ is the Cartesian product of three sets and therefore the sum rule applies. Now, the second and third terms on the right-hand side can be computed via Theorems 2.4 and 2.3, respectively. The first term can also be calculated via Theorem 2.5, where the measure space Ω is chosen to be the singleton $\{0\}$ with $\mathcal{A} = \{\{0\}\}$, $\nu(\{0\}) = 1$. Thus the above inequality yields (4.14). □

Now we consider a special case ($\widetilde{\mathcal{CP}}$) of problem ($\mathcal{CP}$), where

$$R = \mathbb{R}_-^r, \qquad \text{and} \qquad Q(t) = \mathbb{R}_-^\gamma, \qquad S(t) = \mathbb{R}_-^\kappa \qquad \text{for all} \quad t \in [0,1]. \tag{4.28}$$

In this case, we intend to simplify the results given by Theorem 4.1. Then (R_5) is automatically satisfied. The focus is on reformulating conditions (R_3), (C_2), (C_5), and (C_7), and, in Theorem 4.1, conditions (4.7), (4.8), (4.9), and (4.14).

Condition (R_3) is replaced by the following:

($\widetilde{R}_3$) There exist bounded measurable functions $d_1, \ldots, d_q \in \mathcal{L}_\gamma^\infty$ and a constant M such that, for almost all $t \in [0,1]$ and for all $i = 1, \ldots, q$, $j = 1, \ldots, \gamma$,

$$\frac{d_{ij}^2(t)}{\widehat{g}_j(t)} > -M \qquad \text{whenever} \qquad \widehat{g}_j(t) < 0 \text{ and } d_{ij}(t) > 0, \tag{4.29}$$

$$d_{ij}(t) \leq 0 \qquad \text{whenever} \qquad \widehat{g}_j(t) = 0. \tag{4.30}$$

Furthermore, there exists a constant $\tau > 0$ such that, for a.e. $t \in [0,1]$, (4.2) is satisfied.

Remark 4.1. It follows from Corollary 2.1 that if $d_1, \dots, d_q$ satisfy (4.29) and (4.30), then $d_1, \dots, d_q \in C(\widehat{g}|\operatorname{sel}_\infty(Q))$; therefore, they also belong to $T(\widehat{g}|\operatorname{sel}_\infty(Q))$.

Taking the choice $q = 2\gamma$, $d_i = (d_{i,1}, \dots, d_{i,\gamma})$, where $d_{i,j}$ is defined by

$$d_{i,j} = \begin{cases} \sqrt{-\widehat{g}_i}, & i = j, \\ 0, & i \neq j, \end{cases} \qquad d_{\gamma+i,j} = \begin{cases} -\sqrt{-\widehat{g}_i}, & i = j, \\ 0, & i \neq j \end{cases} \qquad (i, j = 1, \dots, \gamma),$$

we can see that $d_1, \dots, d_{2\gamma}$ satisfy (4.29) and (4.30). In this case, (4.2) is equivalent to the following quadratic inequality: for a.e. $t \in [0, 1]$,

$$|\xi^T \widehat{g}_u(t) + \eta^T \widehat{h}_u(t)|^2 - \sum_{i=1}^{\gamma} \xi_i^2 \widehat{g}_i(t) \geq \tau |(\xi, \eta)|^2 \qquad ((\xi, \eta) \in \mathbb{R}^\gamma \times \mathbb{R}^\delta). \tag{4.31}$$

Introducing the notation

$$J(t) := \begin{pmatrix} \widehat{g}_{1u}(t) & \sqrt{-\widehat{g}_1(t)} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \widehat{g}_{\gamma u}(t) & 0 & \dots & \sqrt{-\widehat{g}_\gamma(t)} \\ \widehat{h}_{1u}(t) & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \widehat{h}_{\delta u}(t) & 0 & \dots & 0 \end{pmatrix},$$

we can rewrite (4.31) as

$$(\xi^T, \eta^T) J(t) J^T(t) \begin{pmatrix} \xi \\ \eta \end{pmatrix} \geq \tau |(\xi, \eta)|^2 \qquad ((\xi, \eta) \in \mathbb{R}^\gamma \times \mathbb{R}^\delta).$$

Therefore, it is necessary and sufficient that

$$\det(J(t) J^T(t)) \geq \tau \qquad \text{for a.e. } t \in [0, 1].$$

This latter condition appeared among the assumptions of Theorem 5 in [PZ94b].

The conditions (C_2), (C_5), and (C_7) are replaced by the following:

$(\widetilde{C}_2)$ $\widehat{a}_{i,x_0} \delta x(0) + \widehat{a}_{i,x_1} \delta x(1) \leq 0$ whenever $\widehat{a}_i = 0$ $(i = 1, \dots, r)$.

$(\widetilde{C}_5)$ For all $i = 1, \dots, \gamma$,

$$\widehat{g}_{i,x}(t) \delta x(t) + \widehat{g}_{i,u}(t) \delta u(t) \leq 0 \qquad \text{whenever} \quad \widehat{g}_i(t) = 0, \tag{4.32}$$

and there exists a constant $M > 0$ such that, for a.e. $t \in [0, 1]$,

$$\frac{[\widehat{g}_{i,x}(t) \delta x(t) + \widehat{g}_{i,u}(t) \delta u(t)]^2}{-\widehat{g}_i(t)} \leq M \quad \text{whenever} \quad \widehat{g}_i(t) < 0, \quad \widehat{g}_{i,x}(t) \delta x(t) + \widehat{g}_{i,u}(t) \delta u(t) > 0. \tag{4.33}$$

$(\widetilde{C}_7)$ For all $i = 1, \dots, \kappa$,

$$\widehat{k}_{i,x}(t) \delta x(t) \leq 0 \qquad \text{whenever} \quad \widehat{k}_i(t) = 0, \tag{4.34}$$

and there exists a constant $M > 0$ such that, for all $t \in [0, 1]$,

$$\frac{[\widehat{k}_{i,x}(t) \delta x(t)]^2}{-\widehat{k}_i(t)} \leq M \qquad \text{whenever} \quad \widehat{k}_i(t) < 0, \quad \widehat{k}_{i,x}(t) \delta x(t) > 0. \tag{4.35}$$

Using these new conditions, the statement of Theorem 4.1 simplifies to the following result (cf. [PZ94b, Theorem 5]).

COROLLARY 4.1. *Let $(\widehat{x},\widehat{u})$ be a regular weak local minimum for problem $(\widetilde{\mathcal{CP}})$. Then, for every regular critical arc $(\delta x,\delta u)$, there exist constants $\lambda\in\mathbb{R}$, $\alpha=(\alpha_1,\dots,\alpha_r)\in\mathbb{R}^r$, and $\beta=(\beta_1,\dots,\beta_s)\in\mathbb{R}^s$, an absolutely continuous function $p:[0,1]\to\mathbb{R}^n$, two integrable functions $\varphi:[0,1]\to\mathbb{R}^\gamma$ and $\psi:[0,1]\to\mathbb{R}^\delta$, and a Borel regular vector-valued measure $\mu=(\mu_1,\dots,\mu_\kappa)$, not all zero, such that $\lambda\geq 0$,*

$$\alpha\geq 0,\qquad \alpha^T\widehat{a}=0, \tag{4.36}$$

$$\varphi(t)\geq 0,\qquad \varphi^T(t)\widehat{g}(t)=0\qquad \text{for a.e. } t\in[0,1], \tag{4.37}$$

$$\mu\geq 0,\qquad \int_0^1 \widehat{k}(t)d\mu(t)=0, \tag{4.38}$$

(4.10), (4.11), (4.12), (4.13) *hold, and*

$$\begin{aligned}&\big(\lambda\widehat{\ell}''+\alpha^T\widehat{a}''+\beta^T\widehat{b}''\big)(\delta x(0),\delta x(1))+\int_0^1\widehat{k}''(t;\delta x(t))d\mu(t)\\ &+\int_0^1\widehat{\mathcal{H}}''\Big(t,p(t)+\int_{]t,1]}\widehat{k}_x^T(t)d\mu(t),\varphi(t),\psi(t);\delta x(t),\delta u(t)\Big)dt\geq\sum_{i=1}^{\kappa}2\int_0^1\sigma_{\widehat{k}_i,\widehat{k}_{ix}\delta x}(t)\,d\mu_i(t),\end{aligned} \tag{4.39}$$

where $\sigma_{a,b}$ is defined by (2.14).

Remark 4.2. In the very special case when $g(t,x,u)=u$, the first-order necessary conditions in the above corollary form a special case of [Cla83, Theorem 5.2.1] and [Gir72]. On the other hand, when no state constraints are present, the first-order part of this corollary generalizes the results in [MOS98] and [OS95]. When only equality control constraints are present, the statement of Corollary 4.1 has its exact parallel in [ZZ88] for the case where the state is piecewise smooth and the control is piecewise continuous.

Proof. Define R, Q, and S by (4.28). Using the product rule and Theorem 2.2, it follows that $(\widetilde{R}_3)$ implies (R_3). Similarly, due to the product rule and Theorems 2.2 and 2.1, it follows that conditions $(\widetilde{C}_2)$, $(\widetilde{C}_5)$, and $(\widetilde{C}_7)$ are equivalent to (C_2), (C_5), and (C_7), respectively. Thus, all the assumptions of Theorem 4.1 are satisfied, and hence we also have its conclusions.

We can see that (4.7), (4.8), and (4.9) are equivalent to (4.36), (4.37), and (4.38), respectively. By the second part of Corollary 2.1, the first and third terms on the right-hand side of (4.14) vanish. By Corollary 2.2 the second term of (4.14) reduces to the right-hand side of (2.15). Therefore, (4.14) reduces to (4.39). ∎

5. Example. Consider the problem

$$(\mathcal{C})\qquad \begin{array}{ll}\text{Minimize} & x_3(1)\\ \text{subject to} & \begin{cases}\dot{x_1}(t)=u_1(t)\text{ for a.e. } t\in[-1,1],\\ \dot{x_2}(t)=u_2(t)\text{ for a.e. } t\in[-1,1],\\ \dot{x_3}(t)=x_1^3(t)+\zeta(t)u_2(t)\text{ for a.e. } t\in[-1,1],\\ x_1(-1)=x_2(-1)=x_3(-1)=0,\\ x_1(1)=x_2(1)=0,\\ -x_2(t)-(x_1(t)-t)^2-x_1^2(t)\leq 0\text{ for } t\in[-1,1],\end{cases}\end{array}$$

where

$$\zeta(t) := \begin{cases} 0 & \text{if } t \in [-1,0), \\ -1 & \text{if } t \in [0,1]. \end{cases}$$

The Hamiltonian of this problem is

$$\mathcal{H}(t,x,u,p) := p_1u_1 + p_2u_2 + p_3(x_1^3 + \zeta(t)u_2).$$

This problem is a special case of $(\widetilde{\mathcal{CP}})$, where (i), (iv), and (v) are absent and, for $x^T = (x_1, x_2, x_3)$ and $u^T = (u_1, u_2)$, we have

$$b^T(x(-1), x(1)) = (x^T(-1), x_1(1), x_2(1)), \qquad f^T(t,x,u) = (u_1, u_2, x_1^3 + \zeta(t)u_2),$$

$$k(t,x) = -x_2 - (x_1 - t)^2 - x_1^2.$$

One would like to find out whether the admissible pair $(\widehat{x};\widehat{u})^T = (0,0,0;0,0)$ is a good candidate for weak local minimality in $(\mathcal{C})$. For this reason, we shall check whether the first- and second-order necessary conditions presented by Corollary 4.1 hold true for this candidate.

We have $\widehat{k}(t) = -t^2$ and $\widehat{k}_x(t) = (2t, -1, 0)$. Now set

$$\lambda := 1, \quad p^T(t) :\equiv (0,1,1), \quad \mu := \delta_0 \text{ (the Dirac measure concentrated at 0)}.$$

Then, by replacing the left endpoint 0 of the base interval in Theorem 4.1 and Corollary 4.1 by -1, one can check that these multipliers (that are not all zero) uniquely (up to a nonzero constant multiple) satisfy (4.10), (4.13), and (4.38). By choosing the multipliers $(\beta_1, \dots, \beta_5)$ (that correspond to the endpoint conditions) properly, (4.11) and (4.12) can also be satisfied.

Define, for $t \in [-1,1]$,

$$\begin{aligned} &\delta x^T(t) = (\delta x_1(t), \delta x_2(t), \delta x_3(t)) := (1 - |t|, 0, 0) \\ \text{and} \quad &\delta u^T(t) = (\delta u_1(t), \delta u_2(t)) := (-\operatorname{sign}(t), 0). \end{aligned}$$

It follows that this choice of δx satisfies (4.34) and, for $M = 4$, (4.35). One can also check that all the remaining criticality conditions (and regularity conditions) are also satisfied. Therefore, $(\delta x, \delta u)$ is a regular critical direction for problem $(\mathcal{C})$.

It remains to check the inequality (4.39). Note that the first and the third terms there vanish. Using the definition of $\sigma_{\widehat{k},\widehat{k}_x\delta x}$ in (2.14), inequality (4.39) simplifies to

$$-4[\delta x_1(0)]^2 \geq -2[\delta x_1(0)]^2,$$

which fails to hold, since $\delta x_1(0) = 1$. Therefore, given that the conclusion of Corollary 4.1 is not valid, the pair $(\widehat{x};\widehat{u})^T = (0,0,0;0,0)$ is *not* a weak local minimum for $(\mathcal{C})$.

REFERENCES

[BT80] A. Ben-Tal, *Second order theory of extremum problems*, in Extremal Methods and System Analysis, A. V. Fiacco and K. Kortanek, eds., Springer-Verlag, Berlin, 1980, pp. 336–356.

[BTZ82] A. Ben-Tal and J. Zowe, *A unified theory of first and second-order conditions for extremum problems in topological vector spaces*, Math. Programming Stud., 19 (1982), pp. 39–76.

[Cla83] F. H. Clarke, *Optimization and Nonsmooth Analysis*, Canad. Math. Soc. Ser. Monogr. Adv. Texts, John Wiley, New York, 1983.

[Com90] R. Cominetti, *Metric regularity, tangent sets, and second-order optimality conditions*, Appl. Math. Optim., 21 (1990), pp. 265–287.

[DM63] A. Ya. Dubovitskii and A. A. Milyutin, *Extremum problems with constraints*, Soviet Math. Dokl., 4 (1963), pp. 452–455.

[DM65] A. Ya. Dubovitskii and A. A. Milyutin, *Second variations in extremal problems with constraints*, Dokl. Akad. Nauk SSSR, 160 (1965), pp. 18–21.

[Gir72] I. V. Girsanov, *Lectures on Mathematical Theory of Extremum Problems*, Lecture Notes in Econom. and Math. Systems 67, Springer-Verlag, Berlin, 1972.

[IT79] A. D. Ioffe and V. M. Tihomirov, *Theory of Extremal Problems*, North-Holland, Amsterdam, 1979.

[Kaw88a] H. Kawasaki, *An envelope like effect of infinitely many inequality constraints on second-order necessary conditions for minimization problems*, Math. Programming, 41 (1988), pp. 73–96.

[Kaw88b] H. Kawasaki, *The upper and second-order directional derivatives for a sup-type function*, Math. Programming, 41 (1988), pp. 327–339.

[Kaw91] H. Kawasaki, *Second order necessary optimality conditions for minimizing a sup type function*, Math. Programming, 49 (1991), pp. 213–229.

[Kaw92] H. Kawasaki, *Second order necessary and sufficient optimality conditions for minimizing a sup type function*, Appl. Math. Optim., 26 (1992), pp. 195–220.

[MZ79] H. Maurer and J. Zowe, *First and second order necessary and sufficient conditions for infinite dimensional programming problems*, Math. Programming, 16 (1979), pp. 98–110.

[MOS98] A. A. Milyutin and N. P. Osmolovskii, *Calculus of Variations and Optimal Control*, Transl. Math. Monogr. 180, AMS, Providence, RI, 1998.

[OS95] N. P. Osmolovskii, *Quadratic conditions for nonsingular extremals in optimal control (A theoretical treatment)*, Russ. J. Math. Phys., 2 (1995), pp. 487–516.

[PZ94a] Zs. Páles and V. Zeidan, *Nonsmooth optimum problems with constraints*, SIAM J. Control Optim., 32 (1994), pp. 1476–1502.

[PZ94b] Zs. Páles and V. Zeidan, *First- and second-order necessary conditions for control problems with constraints*, Trans. Amer. Math. Soc., 346 (1994), pp. 421–453.

[PZ96] Zs. Páles and V. Zeidan, *Second-order necessary conditions for nonsmooth optimum problems with constraints*, in World Congress of Nonlinear Analysts '92, Vols. 1–4, de Gruyter, Berlin, 1996, pp. 2337–2346.

[PZ98] Zs. Páles and V. Zeidan, *Optimum problems with certain lower semicontinuous set-valued constraints*, SIAM J. Optim., 8 (1998), pp. 707–727.

[PZ99a] Zs. Páles and V. Zeidan, *Characterization of closed and open C-convex sets in $\mathcal{C}(T,\mathbb{R}^r)$*, Acta Sci. Math. (Szeged), 65 (1999), pp. 339–357.

[PZ99b] Zs. Páles and V. Zeidan, *On L^1-closed decomposable sets in L_∞*, in Systems Modelling and Optimization (Detroit, MI, 1997), Chapman & Hall/CRC, Boca Raton, FL, 1999, pp. 198–206.

[PZ99c] Zs. Páles and V. Zeidan, *Characterization of L^1-closed decomposable sets in L^∞*, J. Math. Anal. Appl., 238 (1999), pp. 491–515.

[PZ00] Zs. Páles and V. Zeidan, *Optimum problems with measurable set-valued constraints*, SIAM J. Optim., 11 (2000), pp. 426–443.

[PZ01] Zs. Páles and V. Zeidan, *The critical tangent cone in second-order conditions for optimal control*, in Proceedings of the 2nd World Congress of Nonlinear Analysts, Catania, Italy, 2000, V. Lakshmikantham, ed., Nonlinear Anal., 47 (2001), pp. 1149–1161.

[SZ] G. Stefani and P. Zezza, *Optimal control problems with mixed state-control constraints: Necessary conditions*, J. Math. Systems Estim. Control, 2 (1992), pp. 155–189.

[ZZ88] V. Zeidan and P. Zezza, *The conjugate point condition for smooth control sets*, J. Math. Anal. Appl., 132 (1988), pp. 572–589.

SIAM J. OPTIM.
Vol. 14, No. 2, pp. 359–379

AN EXTENDED EXTREMAL PRINCIPLE WITH APPLICATIONS TO MULTIOBJECTIVE OPTIMIZATION*

BORIS S. MORDUKHOVICH†, JAY S. TREIMAN‡, AND QIJI J. ZHU‡

Abstract. We develop an extended version of the extremal principle in variational analysis that can be treated as a variational counterpart to the classical separation results in the case of nonconvex sets and which plays an important role in the generalized differentiation theory and its applications to optimization-related problems. The main difference between the conventional extremal principle and the extended version developed below is that, instead of the translation of sets involved in the extremal systems, we allow deformations. The new version seems to be more flexible in various applications and covers, in particular, multiobjective optimization problems with general preference relations. In this way we obtain new necessary optimality conditions for constrained problems of multiobjective optimization with nonsmooth data and also for multiplayer multiobjective games.

Key words. variational analysis, extended extremal principle, generalized differentiation, constrained multiobjective optimization, multiplayer games, necessary optimality conditions

AMS subject classifications. 49J52, 90C29, 90D06

DOI. 10.1137/S1052623402414701

1. Introduction. It is well known that separation theorems for convex sets play a fundamental role in many aspects of nonlinear analysis and optimization. The whole of convex analysis and its applications to constrained optimization and economics revolves around using separation theorems. The convex separation principle is very useful in the study of problems with nonconvex and nonsmooth initial data, being applied to their convex approximations built by means of convex tangent cones and directional derivatives. However, there is a large class of optimization-related and economic problems where the use of convex approximations either is impossible or does not lead to satisfactory results. An adequate approach to such problems is offered by the basic tools of modern variational analysis, in particular, by the so-called *extremal principle*, which can be viewed as a variational counterpart to the convex separation principle in nonconvex settings. We refer the reader to [13] and the bibliography therein for the history, motivations, and applications of the extremal principle in variational analysis.

The conventional extremal principle applies to locally extremal points of set systems, which naturally appear not only in optimization-related problems but also in nonvariational settings as well, e.g., in generalized differential calculus; see [13]. Roughly speaking, a common point of sets is *locally extremal* if these sets can be locally pushed apart by a (linear) *small translation* in such a way that the resulting sets have empty intersection. The extended extremal principle developed in this paper applies to a system of sets and set-valued mappings that allow a (nonlinear) local *deformation*, rather than translation, to end with empty intersection. Such ex-

*Received by the editors September 15, 2002; accepted for publication (in revised form) March 19, 2003; published electronically August 22, 2003.
http://www.siam.org/journals/siopt/14-2/41470.html

†Department of Mathematics, Wayne State University, Detroit, MI 48202 (boris@math.wayne.edu). The research of this author was partly supported by the National Science Foundation under grant DMS-0072179.

‡Department of Mathematics, Western Michigan University, Kalamazoo, MI 49008 (treiman@wmich.edu, zhu@wmich.edu). The research of these authors was partly supported by the National Science Foundation under grant DMS-0102496.

tended extremal systems naturally appear in problems of *multiobjective optimization* with general preference relations. These provide a major motivation for our study; see [26].

We derive two versions of the extended extremal principle: approximate (fuzzy) and exact. Then we apply the extremal principle to necessary optimality conditions for general problems of multiobjective optimization with equality, inequality, and geometric constraints in infinite-dimensional spaces. The forms of the results obtained depend on the assumptions imposed on the initial data. In particular, there are essential differences between problems with Lipschitzian and non-Lipschitzian data. Note that some of our results are new even for multiobjective problems with smooth initial data. We also give applications of the extremal principle to game-theoretical problems involving multiplayer multiobjective game situations.

The rest of this paper is organized as follows. Section 2 contains basic definitions and preliminary material from nonsmooth variational analysis. In section 3 we discuss extended extremal systems and their relations to multiobjective optimization, game problems, and optimal control. Section 4 is devoted to the approximate and exact versions of the extended extremal principle. In section 5 we derive various forms of necessary optimality conditions for nonsmooth problems of multiobjective optimization with equality, inequality, and geometric constraints. The concluding section 6 contains applications of the extended extremal principle to multiobjective games with many players.

Note that another extension of the conventional extremal principle has been recently developed in [10]. It applies to systems of sets that may not be extremal in the sense of the original definition and corresponds to the setting when the relations of the conventional extremal principle are necessary and sufficient for such an extended extremality.

Throughout the paper we use standard notation. Given a Banach space X, we denote by X^* its topological dual with the canonical dual pairing $\langle\cdot,\cdot\rangle$; the same symbol $\|\cdot\|$ is used for denoting the norm on X and for the corresponding dual norm on X^*; $\mathbb{B}_X$ and $\mathbb{B}_{X^*}$ stand for the closed unit balls in the space and dual space in question; and w^* denotes the weak* topology on the dual space. For a set-valued mapping (multifunction) $F\colon X \rightrightarrows X^*$, the expressions

$$\operatorname*{Lim\,sup}_{x\to\bar{x}} F(x) := \big\{x^*\in X^*\;\big|\;\; \text{there exist sequences } \; x_k\to\bar{x} \;\text{ and }\; x_k^*\xrightarrow{w^*}x^* \;\text{ with }\; x_k^*\in F(x_k) \;\text{ for all }\; k\in\mathbb{N}\big\},$$

$$\operatorname*{Lim\,inf}_{x\to\bar{x}} F(x) := \big\{x^*\in X^*\;\big|\;\; \text{for all sequence } \; x_k\to\bar{x},\; \exists\, x_k^*\xrightarrow{w^*}x^* \;\text{ with }\; x_k^*\in F(x_k) \;\text{ for all }\; k\in\mathbb{N}\big\}$$

signify, respectively, the *sequential* Painlevé–Kuratowski upper/outer and lower/inner limits in the norm topology in X and the weak* topology in X^*; $\mathbb{N}:=\{1,2,\ldots\}$.

2. Generalized differential constructions and preliminaries. In this section we review generalized differential constructions of nonsmooth variational analysis and their basic properties, which are widely used in what follows. Although most definitions in this and subsequent sections hold in any Banach space (even in more general settings in some cases), the basic properties of normals and subgradients employed below require the Asplund structure on the spaces in question, which we assume unless otherwise stated. Recall that a Banach space X is *Asplund* if every convex

continuous function on it is generically Fréchet differentiable. This is a broad class of Banach spaces including all spaces with Fréchet differentiable renorms and bump functions and hence any reflexive space. There are many geometric characterizations of Asplund spaces, one of which is that X is Asplund if and only if every separable subspace of it has a separable dual. Although the Asplund property is closely related to Fréchet-like differentiability, there are Asplund spaces that fail to admit even a Gâteaux differentiable renorm. The reader can find more information on Asplund spaces in [20] and the references therein.

Given an extended-real-valued function $\varphi\colon X \to \overline{\mathbb{R}} := (\infty, \infty]$ finite at $\bar{x}$ and a nonempty set $\Omega \subset X$, let us define the basic subdifferential and normal cone constructions of our study and applications in this paper. We refer the reader to [12], [22] and to [4], [15] for more details on these constructions in, respectively, finite and infinite dimensions.

Let $\varphi\colon X \to \overline{\mathbb{R}}$ be lower semicontinuous around $\bar{x}$. The *Fréchet subdifferential* of φ at $\bar{x}$ is

$$\widehat{\partial}\varphi(\bar{x}) := \Big\{x^* \in X^* \Big| \liminf_{x\to\bar{x}} \frac{\varphi(x) - \varphi(\bar{x}) - \langle x^*, x - \bar{x}\rangle}{\|x - \bar{x}\|} \geq 0\Big\}. \tag{2.1}$$

The *limiting subdifferential* and the *singular subdifferential* of φ at $\bar{x}$ are defined, respectively, by

$$\partial\varphi(\bar{x}) := \operatorname*{Lim\,sup}_{x \overset{\varphi}{\to} \bar{x}} \widehat{\partial}\varphi(x), \tag{2.2}$$

$$\partial^\infty\varphi(\bar{x}) := \operatorname*{Lim\,sup}_{\substack{x \overset{\varphi}{\to} \bar{x}\\ \lambda\downarrow 0}} \lambda\widehat{\partial}\varphi(x), \tag{2.3}$$

where $x \overset{\varphi}{\to} \bar{x}$ means that $x \to \bar{x}$ with $\varphi(x) \to \varphi(\bar{x})$. Note that $\widehat{\partial}\varphi(\bar{x})$ (resp., $\partial\varphi(\bar{x})$) reduces to the classical Fréchet derivative (resp., strict derivative) of φ at $\bar{x}$ if φ is Fréchet differentiable (resp., strictly differentiable) at this point. On the other hand, $\partial^\infty\varphi(\bar{x}) = \{0\}$ if φ is locally Lipschitzian around $\bar{x}$.

Let $\Omega \subset X$ be locally closed around $\bar{x} \in \Omega$; i.e., Ω is closed at x whenever x is near $\bar{x}$. Then the *Fréchet normal cone* $\widehat{N}(\bar{x}; \Omega)$ and the *limiting normal cone* $N(\bar{x}; \Omega)$ to Ω at $\bar{x}$ are defined by

$$\widehat{N}(\bar{x}; \Omega) := \left\{x^* \in X^* \Big| \limsup_{x \overset{\Omega}{\to} \bar{x}} \frac{\langle x^*, x - \bar{x}\rangle}{\|x - \bar{x}\|} \leq 0\right\}, \tag{2.4}$$

$$N(\bar{x}; \Omega) := \operatorname*{Lim\,sup}_{x \overset{\Omega}{\to} \bar{x}} \widehat{N}(x; \Omega), \tag{2.5}$$

where $x \overset{\Omega}{\to} \bar{x}$ stands for $x \to \bar{x}$ with $x \in \Omega$. One clearly has

$$\widehat{N}(\bar{x}; \Omega) = \widehat{\partial}\delta(\bar{x}; \Omega), \quad N(\bar{x}; \Omega) = \partial\delta(\bar{x}; \Omega),$$

where $\delta(\cdot; \Omega)$ is the indicator function of Ω. It is known also that

$$\begin{aligned} \partial\varphi(\bar{x}) &= \big\{x^* \in X^* \,\big|\, (x^*, -1) \in N((\bar{x}, \varphi(\bar{x})); \operatorname{epi}\varphi)\big\},\\ \partial^\infty\varphi(\bar{x}) &= \big\{x^* \in X^* \,\big|\, (x^*, 0) \in N((\bar{x}, \varphi(\bar{x})); \operatorname{epi}\varphi)\big\} \end{aligned} \tag{2.6}$$

if φ is lower semicontinuous around $\bar{x}$, and that

$$\begin{aligned}\partial\varphi(\bar{x}) &= \big\{x^* \in X^* \;\big|\; (x^*,-1) \in N((\bar{x},\varphi(\bar{x})); \operatorname{gph}\varphi)\big\},\\ \partial^\infty\varphi(\bar{x}) \cup \partial^\infty(-\varphi)(\bar{x}) &= \big\{x^* \in X^* \;\big|\; (x^*,0) \in N((\bar{x},\varphi(\bar{x})); \operatorname{gph}\varphi)\big\}\end{aligned} \tag{2.7}$$

if φ is continuous around this point; see [15], [18], and the references therein.

Next we present an important *fuzzy sum rule* for Fréchet subgradients, obtained in [6] in the framework of Asplund spaces. A prototype of this result first appeared in [8]; see also [4] and its references for more discussions and generalizations.

PROPOSITION 2.1. *Let $\varphi_i: X \to \overline{\mathbb{R}}$, $i = 1,\dots,n \ge 2$, be lower semicontinuous around $\bar{x}$ such that all but one of them are Lipschitz continuous around this point. Assume that the sum $\sum_{i=1}^n \varphi_i$ attains a local minimum at $\bar{x}$. Then for any $\varepsilon > 0$ there are $x_i \in \bar{x} + \varepsilon\mathbb{B}_X$ with $|\varphi_i(x_i) - \varphi_i(\bar{x})| \le \varepsilon$, $i = 1,\dots,n$, such that*

$$0 \in \sum_{i=1}^n \widehat{\partial}\varphi_i(x_i) + \varepsilon\mathbb{B}_{X^*}.$$

Another calculus result we use in this paper is the following *fuzzy chain rule* for Fréchet subgradients of compositions $(\varphi \circ f)(x) = \varphi(f(x))$ in Asplund spaces given in [16]. Recall that $\langle y^*, f\rangle(x) := \langle y^*, f(x)\rangle$ for a single-valued mapping $f: X \to Y$.

PROPOSITION 2.2. *Let $f: X \to Y$ and $\varphi: Y \to \mathbb{R}$ be locally Lipschitzian around the points under consideration. Then for any $x^* \in \widehat{\partial}(\varphi \circ f)(\bar{x})$ and any $\varepsilon > 0$ there are $x \in \bar{x} + \varepsilon\mathbb{B}_X$, $y \in f(x) + \varepsilon\mathbb{B}_Y$, and $y^* \in \widehat{\partial}\varphi(y)$ such that $\|f(x) - f(\bar{x})\| \le \varepsilon$ and*

$$x^* \in \widehat{\partial}\langle y^*, f\rangle(x) + \varepsilon\mathbb{B}_{X^*}.$$

Finally in this section, let us consider the problem of mathematical programming with equality, inequality, and set (geometric) constraints:

$$\begin{array}{lll} \mathcal{P} \quad \text{Minimize} & \varphi_0(x) & \\ \text{subject to} & \varphi_i(x) \le 0, & i = 1,\dots,m,\\ & \varphi_i(x) = 0, & i = m+1,\dots,n,\\ & x \in C, & \end{array}$$

where $\varphi_i: X \to \overline{\mathbb{R}}$, $i = 1,\dots,n$, and $C \subset X$. To simplify notation, we define the quantities τ_i, $i = 1,\dots,n$, as in [3]; namely, $\tau_i := 1$ for $i = 1,\dots,m$ and $\tau_i \in \{-1,1\}$ for $i = m+1,\dots,n$. The following necessary optimality conditions for $\mathcal{P}$ in a *weak fuzzy* form of Lagrange multipliers are obtained in [19] for non-Lipschitzian problems in Asplund spaces; cf. also [3] and [17].

PROPOSITION 2.3. *Let $\bar{x}$ be a local solution to problem $\mathcal{P}$, where φ_i are lower semicontinuous around $\bar{x}$ for $i = 0,1,\dots,m$ and continuous around this point for $i = m+1,\dots,n$, and where C is locally closed. Assume that*

$$\liminf_{x\to\bar{x}} \operatorname{dist}(0; \widehat{\partial}\varphi_i(x)) > 0, \quad i = 1,\dots,m, \quad \textit{and} \tag{2.8}$$

$$\liminf_{x\to\bar{x}} \operatorname{dist}(0; \widehat{\partial}\varphi_i(x) \cup \widehat{\partial}(-\varphi_i)(x)) > 0, \quad i = m+1,\dots,n. \tag{2.9}$$

Then for any $\varepsilon > 0$ and any weak* neighborhood V of the origin in X^* there are multipliers $\lambda_i \ge 0$, $i = 1,\dots,n$, not all zero, and points $x_i \in \bar{x} + \varepsilon\mathbb{B}_X$, $i = 0,1,\dots,n+1$, such that

$$|\varphi_i(x_i) - \varphi_i(\bar{x})| \le \varepsilon, \quad i = 0,1,\dots,n, \quad \textit{and}$$

$$0 \in \widehat{\partial}\varphi_0(x_0) + \sum_{i=1}^{n} \lambda_i \widehat{\partial}(\tau_i\varphi_i)(x_i) + \widehat{N}(x_{n+1}; C) + V$$

with some τ_i *described above.*

3. Extended extremal systems. In this section we define and illustrate the notion of *extremal systems of multifunctions*, which extend the one for systems of sets. First let us recall that $\bar{x} \in \Omega_1 \cap \Omega_2$ is a (locally) *extremal point* of sets Ω_1 and Ω_2 in a normed space X if there exists a neighborhood U of $\bar{x}$ such that for every $\varepsilon > 0$ there is a vector $a \in \varepsilon\mathbb{B}_X$ with

$$(\Omega_1 + a) \cap \Omega_2 \cap U = \emptyset; \tag{3.1}$$

see [13] for extremal systems of finitely many sets and more discussions.

The concept of extremal points captures the essential geometry in various optimization problems and has many applications as mentioned above. The following is a typical situation in optimization that leads to extremal points.

Example 3.1. Let $\varphi_i\colon X \to \overline{\mathbb{R}}$, $i = 1, 2$, be lower semicontinuous functions. If $\varphi_1 + \varphi_2$ attains a local minimum at $\bar{x}$, then the sets

$$\Omega_1 := \operatorname{epi}(\varphi_1 - \varphi_1(\bar{x})) \quad \text{and} \quad \Omega_2 := \operatorname{hypo}(\varphi_2(\bar{x}) - \varphi_2)$$

have an extremal point at $(\bar{x}, 0)$. Indeed, $(\Omega_1 + (0, \alpha)) \cap \Omega_2 \cap (U \times \mathbb{R}) = \emptyset$ for all $\alpha > 0$, where U is a neighborhood of $\bar{x}$.

Note that the condition $\Omega_1 \cap \Omega_2 = \{\bar{x}\}$ does not necessarily imply that $\bar{x}$ is an extremal point of the set system (Ω_1, Ω_2) even when it is a boundary point for each of the sets. This is illustrated by the following example.

Example 3.2. Let

$$\Omega_1 := \bigcup_{k=1}^{\infty} \Big((1/k, 0) + \frac{1}{4k^2}\mathbb{B}_{R^2}\Big) \cup \{(0, 0)\}, \quad \Omega_2 := \mathbb{R}^2 \setminus \bigcup_{k=1}^{\infty} \Big((1/k, 0) + \frac{1}{4k^2 - 1}\mathbb{B}_{R^2}\Big).$$

Then $\Omega_1 \cap \Omega_2 = (0, 0)$, while $\bar{x} = (0, 0)$ is not an extremal point of (Ω_1, Ω_2), since $(\Omega_1 + (\alpha, \beta)) \cap \Omega_2 \neq \emptyset$ for any $(\alpha, \beta) \neq (0, 0)$.

Recently, to prove necessary conditions for a multiobjective optimal control problem in [26], a construction similar to the extremal system of sets was used, where the set Ω_1 is *deformed* rather than *translated* as in (3.1). This motivates us to extend the concept of extremal points for sets to the one for multifunctions. In this extension the translation of sets is replaced by the deformation of sets, which is more flexible for applications.

Let us define extended extremal systems and their local extremal points. In what follows we omit for simplicity the adjectives "extended" and "local" relative to extremal systems and points, since we are not going to consider any other extremal concepts in the rest of the paper.

Definition 3.3. *Let* $S_i\colon M_i \rightrightarrows X$, $i = 1, \dots, p$, *be multifunctions from metric spaces* M_i *with metrics* d_i *into a Banach space* X. *We say that* $\bar{x}$ *is an* extremal point *of the system* $(S_1, \dots, S_p)$ *at* $(\bar{s}_1, \bar{s}_2, \dots, \bar{s}_p)$, *provided that*

$$\bar{x} \in S_1(\bar{s}_1) \cap S_2(\bar{s}_2) \cap \dots \cap S_p(\bar{s}_p),$$

and there exists a neighborhood U *of* $\bar{x}$ *such that for every* $\varepsilon > 0$ *there is* $(s_1, s_2, \dots, s_p) \in M_1 \times \dots \times M_p$ *with*

$$d(s_i, \bar{s}_i) \le \varepsilon, \quad \operatorname{dist}(\bar{x}; S_i(s_i)) \le \varepsilon \ \text{ for } \ i = 1, \dots, p, \quad \text{and}$$

$$S_1(s_1) \cap S_2(s_2) \cap \cdots \cap S_p(s_p) \cap U = \emptyset.$$

In this case $(S_1, \ldots, S_p)$ *is called the* extremal system.

It is easy to see that extremal systems of multifunctions from Definition 3.3 contain as a special case extremal systems of sets $\Omega_1, \Omega_2 \subset X$ defined in the beginning of this section. To capture it, we put $M_1 := X$, $M_2 := \{0\}$, $S_1(s_1) := \Omega_1 + s_1$, and $S_2(0) := \Omega_2$ in Definition 3.3.

The next example shows that extremal systems involving deformations of sets as in Definition 3.3 cannot be reduced to those obtained by their translations.

Example 3.4. Consider the moving sets

$$S_1(s_1) := \Big\{(x,y) \in \mathbb{R}^2 \;\Big|\; |x| - 2|y| \ge s_1\Big\},$$
$$S_2(s_2) := \Big\{(x,y) \in \mathbb{R}^2 \;\Big|\; |y| - 2|x| \ge s_2\Big\},$$

which can be viewed as deformations of the initial sets $\Omega_1 := S_1(0)$ and $\Omega_2 := S_2(0)$. One can check that $(0,0)$ is an extremal point of (S_1, S_2) in the sense of Definition 3.3, while $(0,0)$ is not an extremal point of $\{\Omega_1, \Omega_2\}$ in the sense of translation (3.1).

Next we consider an example of extended extremal systems, which demonstrates a prime motivation for our study. Let $\prec$ be an arbitrary preference for elements of a Banach space X. For any $x \in X$ we define the level (or sublevel) set with respect to $\prec$ by

$$\mathcal{L}(x) := \{y \in X |\; y \prec x\}$$

and call this preference to be *locally satiated* around $\bar{x}$ if $x \in \operatorname{cl} \mathcal{L}(x)$ for all x in some neighborhood of $\bar{x}$. We say that $\prec$ is *almost transitive* provided that for each $y \prec x$ and $z \in \operatorname{cl} \mathcal{L}(y)$ one has $z \prec x$. Such an extended preference concept covers many conventional and nonconventional preference relations in multiobjective optimization and economics and cannot be generally described in terms of utility functions; see [26] for more discussions and references.

Example 3.5. Let X be a Banach space with a locally satiated and almost transitive preference $\prec$. Take a continuous mapping $f\colon Z \to X$ on another Banach space Z and consider the following *multiobjective optimization* problem:

$$\begin{aligned} &\text{Minimize} \quad f(z) \\ &\text{subject to} \quad z \in C, \end{aligned}$$

where "minimization" is understood with respect to the preference $\prec$. Namely, we say that $\bar{z}$ is a (local) *solution* to the above problem if there is no $z \in C$ near $\bar{z}$ such that $f(z) \prec f(\bar{z})$. Now let us relate this solution to an extremal point in the sense of Definition 3.3. Put

$$M_1 := \mathcal{L}(f(\bar{z})) \cup \{f(\bar{z})\},\ M_2 := \{0\},\ S_1(s_1) := C \times \operatorname{cl} \mathcal{L}(s_1),\ S_2 := \big\{(z, f(z))\big|\; z \in Z\big\}.$$

Then $(\bar{z}, f(\bar{z}))$ is an *extremal point* of the system (S_1, S_2) at $(f(\bar{z}), 0)$. Indeed, suppose that it is not the case, i.e., for any neighborhood U of $(\bar{z}, f(\bar{z}))$ there is $s_1 \in M_1 \setminus \{f(\bar{z})\}$ close to $f(\bar{z})$ satisfying

$$S_1(s_1) \cap S_2 \cap U \ne \emptyset.$$

Then there exists z close to $\bar{z}$ with $(z, f(z)) \in S_1(s_1) = C \times \operatorname{cl} \mathcal{L}(s_1)$. Hence $z \in C$ and $f(z) \prec f(\bar{z})$, which is a contradiction.

Our next example of extremal systems concerns *game theory*; see also section 6 for more in this direction.

Example 3.6. Consider a *two-player game*, where players A and B have strategy sets $C \subset X$ and $D \subset Y$ as closed subsets of Banach spaces. Given a *payoff* function $\varphi\colon X \times Y \to \mathbb{R}$, we examine a standard game situation, where the objective of player A is to maximize the payoff while that of B is to minimize it. In other words, we consider the game problem

$$\max_{x\in C} \varphi(x,y) \quad \text{and} \quad \min_{y\in D} \varphi(x,y)$$

and define its solution as a *saddle point* $(\bar{x},\bar{y}) \in C \times D$ satisfying

$$\varphi(x,\bar{y}) \le \varphi(\bar{x},\bar{y}) \le \varphi(\bar{x},y) \ \text{ whenever } \ (x,y) \in C \times D.$$

To reduce the saddle point $(\bar{x},\bar{y})$ to an extremal point of some system from Definition 3.3, we form a set-valued mapping

$$S_1(\alpha,\beta) := C \times [\alpha,\infty) \times D \times (-\infty,\beta]$$

on $[\varphi(\bar{x},\bar{y}),\infty) \times (-\infty,\varphi(\bar{x},\bar{y})]$ and a set

$$S_2 := \operatorname{hypo}\varphi(\cdot,\bar{y}) \times \operatorname{epi}\varphi(\bar{x},\cdot).$$

Then one has $(\bar{x},\varphi(\bar{x},\bar{y}),\bar{y},\varphi(\bar{x},\bar{y})) \in S_1(\varphi(\bar{x},\bar{y}),\varphi(\bar{x},\bar{y})) \cap S_2$ and

$$S_1(\alpha,\beta) \cap S_2 = \emptyset \ \text{ for any } \ (\alpha,\beta) \in [\varphi(\bar{x},\bar{y}),\infty) \times (-\infty,\varphi(\bar{x},\bar{y})] \backslash \{(\varphi(\bar{x},\bar{y}),\varphi(\bar{x},\bar{y}))\}.$$

Thus $(\bar{x},\varphi(\bar{x},\bar{y}),\bar{y},\varphi(\bar{x},\bar{y}))$ is an extremal point for the system (S_1,S_2) at $(\varphi(\bar{x},\bar{y}),$ $\varphi(\bar{x},\bar{y}))$. Actually we have the more precise *decoupling conclusion*: the point $(\bar{x},\varphi(\bar{x},\bar{y}))$ is an extremal point for the system $(S_1(\alpha,\varphi(\bar{x},\bar{y})),\operatorname{hypo}\varphi(\cdot,\bar{y}))$, and the point $(\bar{y},\varphi(\bar{x},\bar{y}))$ is an extremal point for the system $(S_1(\varphi(\bar{x},\bar{y}),\beta),\operatorname{epi}\varphi(\bar{x},\cdot))$.

Our final example in this section concerns *optimal control* of systems governed by ordinary differential equations with endpoint constraints.

Example 3.7. Consider the following *time optimal control problem*: minimize the transient time T subject to the endpoint constraint $x(T) = 0$ over absolutely continuous trajectories $x\colon [0,T] \to \mathbb{R}^n$ satisfying

$$\dot{x}(t) = f(x(t),u(t)), \quad x(0) = x_0, \quad u(t) \in U \ \text{ a.e. } \ t \in [0,T]. \tag{3.2}$$

Let $\overline{T}$ be the optimal time in the above problem. Define $M_1 := (0,\infty)$, $M_2 := \{0\} \subset \mathbb{R}$, and $S_2 := \{0\} \subset \mathbb{R}^n$, and $S_1\colon M_1 \rightrightarrows \mathbb{R}^n$ is a *reachable set* multifunction with

$$S_1(s_1) := \Big\{ x(s_1) \in \mathbb{R}^n \Big| \ x(\cdot) \ \text{ is feasible in (3.2) on } [0,s_1] \Big\}.$$

Then one can check that $0 \in \mathbb{R}^n$ is an extremal point of the system (S_1,S_2) at $(\overline{T},0)$ in the sense of Definition 3.3.

4. Extended extremal principle. This section is mostly devoted to the formulation and proof of the extended extremal principle in both fuzzy/approximate and limiting/exact forms. We also consider some auxiliary material needed for these purposes. Let us start with the following approximate version.

THEOREM 4.1. *Let $S_i\colon M_i \rightrightarrows X$ be multifunctions from metric spaces M_i with metric d_i into an Asplund space X for $i = 1,\dots,p$. Assume that $\bar{x}$ is an extremal*

point of the system $(S_1,\dots,S_p)$ *at* $(\bar s_1,\dots,\bar s_p)$*, where each* S_i *is closed-valued around* $\bar s_i$*. Then for every* $\varepsilon>0$ *there are* $s_i\in M_i$*,* $x_i\in S_i(s_i)$*, and* $x_i^*\in X^*$*,* $i=1,\dots,p$*, such that*

$$d(s_i,\bar s_i)\le\varepsilon,\quad \|x_i-\bar x\|\le\varepsilon,\quad x_i^*\in\widehat N(x_i;S_i(s_i))+\varepsilon\mathbb{B}_{X^*},\tag{4.1}$$

$$\|x_1^*\|+\cdots+\|x_p^*\|=1,\quad \textit{and}\quad x_1^*+\cdots+x_p^*=0.\tag{4.2}$$

Proof. Let U be a neighborhood of $\bar x$ from the definition of the extremal point; for simplicity take $U:=\bar x+r\mathbb{B}_X$. Picking an arbitrary ε, we choose

$$\varepsilon'<\min\left\{\varepsilon^2/(5\varepsilon+12p^2+\varepsilon^2),r^2/4\right\}$$

and take $s_1,\dots,s_p$ from Definition 3.3 corresponding to ε'. Denote $\Omega:=S_1(s_1)\times\cdots\times S_p(s_p)$ and form the function

$$\varphi(y_1,\dots,y_p):=\sum_{i,j=1}^p\|y_i-y_j\|+\delta((y_1,\dots,y_p);\Omega),\quad (y_1,\dots,y_p)\in U^p.\tag{4.3}$$

From the construction of (4.3) one has that φ is lower semicontinuous and positive on the complete metric space U^p. On the other hand, we may choose $y_i'\in S_i(s_i)$ satisfying

$$\|y_i'-y_j'\|\le\operatorname{dist}(\bar x;S_i(s_i))+\operatorname{dist}(\bar x;S_j(s_j))+\varepsilon'\le 3\varepsilon'.$$

This gives $\varphi(y_1',\dots,y_p')\le 3p^2\varepsilon'<\varepsilon^2/4$. By the Ekeland variational principle [5] applied to (4.3) one has $x_i'\in y_i'+\varepsilon/2\mathbb{B}_X\subset\bar x+\varepsilon\mathbb{B}_X$, $i=1,\dots,p$, such that the perturbed function

$$\sum_{i,j=1}^p\|y_i-y_j\|+\frac{\varepsilon}{2}\sum_{i=1}^p\|y_i-x_i'\|+\delta((y_1,\dots,y_p);\Omega)\tag{4.4}$$

attains its global minimum at $(x_1',\dots,x_p')$ on U^p. Assume that $U^p=X^p$ without loss of generality and denote

$$\psi(y_1,\dots,y_p):=\sum_{i,j}^p\|y_i-y_j\|,\quad (y_1,\dots,y_p)\in X^p.$$

One clearly has $\psi(x_1',\dots,x_p')>0$. Now applying the fuzzy sum rule from Proposition 2.1 to (4.4) and taking into account that

$$\widehat\partial\delta((y_1,\dots,y_p);\Omega)=\widehat N(y_1;S_1(y_1))\times\cdots\times\widehat N(y_p;S_p(s_p))\ \text{ for any }\ y_i\in S(s_i),$$

we find $x_i\in S_i(s_i)\cap(x_i'+\varepsilon'\mathbb{B}_X)\subset(\bar x+\varepsilon\mathbb{B}_X)$, $z_i\in x_i'+\varepsilon'\mathbb{B}_X$, $i=1,\dots,p$, and $(x_1^*,\dots,x_p^*)\in\widehat\partial\psi(z_1,\dots,z_p)$ such that

$$0\in(x_1^*,\dots,x_p^*)+\widehat N(x_1;S_1(s_1))\times\cdots\times\widehat N(x_p;S_p(s_p))+\varepsilon'(p+1)\mathbb{B}_{(X^p)^*}.$$

The latter relations clearly imply that

$$-x_i^*\in\widehat N(x_i;S_i(s_i))+\varepsilon\mathbb{B}_{X^*}\ \text{ whenever }\ i=1,\dots,p$$

for the chosen ε; so we get the inclusion in (4.1) just by changing the sign of x_i^*, $i = 1, \dots, p$.

Shrinking ε' further if necessary, we can make $\psi(z_1, \dots, z_p) > 0$. Observe that $(x_1^*, \dots, x_p^*) \in \widehat{\partial}\psi(z_1, \dots, z_p)$ yields

$$\begin{aligned}\langle x_1^* + \cdots + x_p^*, h\rangle &\le \liminf_{t\to 0} \frac{\psi(z_1 + th, \dots, z_p + th) - \psi(z_1, \dots, z_p)}{t} \\ &= \liminf_{t\to 0} \frac{\sum_{i,j=1}^p \|(z_i + th) - (z_j + th)\| - \sum_{i,j=1}^p \|z_i - z_j\|}{t} = 0\end{aligned}$$

for any unit vector $h \in X$. This gives the second relation (Euler equation) in (4.2). Since we have already proved (4.1), it remains to show that

$$\|x_1^*\| + \cdots + \|x_p^*\| \ge 1, \tag{4.5}$$

which implies the first relations in (4.2) by normalization. To prove (4.5), we first observe that ψ is positive homogeneous. Then using $(x_1^*, \dots, x_p^*) \in \widehat{\partial}\psi(z_1, \dots, z_p)$, one has

$$\sum_{i=1}^p \langle x_i^*, -z_i\rangle \le \liminf_{t\to 0} \frac{\psi(z_1 - tz_1, \dots, z_p - tz_p) - \psi(z_1, \dots, z_p)}{t} = -\psi(z_1, \dots, z_p).$$

Since $x_1^* = -\sum_{i=2}^p x_i^*$, one has

$$\begin{aligned}&\psi(z_1, \dots, z_p) \le \sum_{i=1}^p \langle x_i^*, z_i\rangle = \sum_{i=2}^p \langle x_i^*, z_i - z_1\rangle \\ &\le \max\Big\{\|x_i^*\| \;\Big|\; i = 2, \dots, p\Big\} \sum_{i=2}^p \|z_i - z_1\| \le \max\Big\{\|x_i^*\| \;\Big|\; i = 1, \dots, p\Big\}\psi(z_1, \dots, z_p).\end{aligned}$$

Since $\psi(z_1, \dots, z_p) > 0$, we get from here that $\max\{\|x_i^*\| \mid i = 1, \dots, p\} \ge 1$, which implies (4.5) and ends the proof of the theorem. □

Remark 4.2. In fact, the Asplund property of the space X in Theorem 4.1 is not only sufficient but also *necessary* for the fulfillment of the extended extremal principle formulated here. Indeed, the extended extremal principle implies the conventional extremal principle for fixed sets, which is known to be a characterization of Asplund spaces; see [14] and also [25], where the reader can find other equivalencies between basic results in variational analysis.

Next we are going to derive the *exact/limiting form* of the extended extremal principle. This requires some additional assumptions on the set-valued mappings involved in extremal systems. First we need to define one more construction concerning generalized normals to *moving* sets.

DEFINITION 4.3. *Let $S\colon Z \rightrightarrows X$ be a set-valued mapping from a metric space Z into a Banach space X, and let $(\bar z, \bar x) \in \operatorname{gph} S$. Then*

$$\widetilde{N}(\bar x; S(\bar z)) := \operatorname*{Lim\,sup}_{(z,x) \overset{gph S}{\to} (\bar z, \bar x)} \widehat{N}(x; S(z)) \tag{4.6}$$

is the extended normal cone *to $S(\bar z)$ at $\bar x$. The mapping S is* normally semicontinuous *at $(\bar z, \bar x)$ if $\widetilde{N}(\bar x; S(\bar z)) = N(\bar x; S(\bar z))$.*

Observe that one always has the inclusion

$$N(\bar{x};S(\bar{z})) := \operatorname*{Lim\,sup}_{x \overset{S(\bar{z})}{\to} \bar{x}} \widehat{N}(x;S(\bar{z})) \subset \widetilde{N}(\bar{x};S(\bar{z})); \tag{4.7}$$

thus the normal semicontinuity of S at $(\bar{z},\bar{x})$ corresponds to the opposite inclusion in (4.7). This property was studied and used in [12] under the name of "normal semicontinuity" and then in [26] under the name of "regularity." It is easy to see that the inclusion in (4.7) may be strict in very simple situations (e.g., when $S(z)$ is a singleton around $\bar{z}$ while $N(\bar{x};S(\bar{z})) \neq X^*$). An interesting example of violating the normal semicontinuity is given in [1] for a mapping $S(z) = \operatorname{cl}\mathcal{L}(z)$ generated by level sets of the preference determined by a Lipschitz continuous utility function on $\mathbb{R}^2$.

Let us present some sufficient conditions for the normal semicontinuity of set-valued mappings. The next proposition corresponds to [12, Proposition 5.1], established in finite dimensions.

PROPOSITION 4.4. *Let $S\colon Z \rightrightarrows X$ be a multifunction from a metric space Z into a Banach space X. Then S is normally semicontinuous at $(\bar{z},\bar{x}) \in \operatorname{gph} S$ in the following two cases:*

(i) *$S(z) = g(z) + \Omega$, where $\Omega \subset X$ is an arbitrary nonempty set and $g\colon Z \to X$ is a continuous mapping.*

(ii) *S is convex-valued near $\bar{z}$ and inner semicontinuous at this point, i.e.,*

$$S(\bar{z}) \subset \operatorname*{Lim\,inf}_{z\to\bar{z}} S(z).$$

Proof. In case (i) the normal semicontinuity property follows from the definition of the limiting normal cone (2.5). Note that this case is sufficient for applications to the limiting extremal principle involving the translation of fixed sets.

Let us consider case (ii). Taking $x^* \in \widetilde{N}(\bar{x};S(\bar{z}))$, we find sequences $x_k \to \bar{x}$, $z_k \to \bar{z}$, and $x_k^* \overset{w^*}{\to} x^*$ such that $x_k^* \in \widehat{N}(x_k;S(z_k))$ for all $k \in \mathbb{N}$. It is well known that for convex sets the Fréchet normal cone agrees with the normal cone of convex analysis. Thus the latter inclusion is equivalent to

$$\langle x_k^*, u - x_k\rangle \le 0 \ \text{ for all } \ u \in S(z_k). \tag{4.8}$$

Let us show that the inner semicontinuity assumption in (ii) implies that

$$\langle x^*, u - \bar{x}\rangle \le 0 \ \text{ for all } \ u \in S(\bar{z}), \tag{4.9}$$

which means that $x^* \in N(\bar{x};S(\bar{z}))$, since the limiting normal cone also agrees with the normal cone of convex analysis for convex sets.

Indeed, assume on the contrary that (4.9) is violated at some $\bar{u} \in S(\bar{z})$, i.e., $\langle x^*, \bar{u} - \bar{x}\rangle > 0$. Using the inner semicontinuity of S at $\bar{z}$, for the given $\bar{u}$ and the sequence $z_k \to \bar{z}$ we find a sequence $u_k \to \bar{u}$ such that $u_k \in S(z_k)$ for all $k \in \mathbb{N}$. We have the representation

$$\langle x_k^*, u_k - x_k\rangle = \langle x^*, \bar{u} - \bar{x}\rangle + \Big[\langle x_k^* - x^*, \bar{u} - \bar{x}\rangle + \langle x_k^*, u_k - \bar{u}\rangle - \langle x_k^*, x_k - \bar{x}\rangle\Big].$$

One can see that all the terms in the square brackets tend to zero as $k \to \infty$ due to the corresponding convergence of $x_k,\ u_k, x_k^*$ and the boundedness of $\{x_k^*\}$. So we arrive at

$$\langle x_k^*, u_k - x_k\rangle > 0 \ \text{ for large } k \in \mathbb{N},$$

which contradicts (4.8) and completes the proof of the proposition. □

Remark 4.5. Recently Lionel Thibault [23] obtained other sufficient conditions ensuring the normal semicontinuous property of set-valued mappings. In particular, he proved this property for inner semicontinuous mappings whose images (near reference points) are *uniformly prox-regular* in the sense of [21] in Hilbert spaces.

To proceed toward the limiting extremal principle, we need one more property of set-valued mappings, which is needed in the case of infinite-dimensional image spaces.

DEFINITION 4.6. *We say that* $S\colon Z \rightrightarrows X$ *is* imagely sequentially normally compact (ISNC) *at* $(\bar{z},\bar{x}) \in \operatorname{gph} S$ *if for any sequences* (z_k, x_k, x_k^*) *satisfying*

$$x_k^* \in \widehat{N}(x_k; S(z_k)), \quad (x_k, z_k) \overset{\operatorname{gph} S}{\to} (\bar{x},\bar{z}), \quad x_k^* \overset{w^*}{\to} 0$$

one has $\|x_k^*\| \to 0$.

This property obviously holds when X is finite-dimensional. For constant mappings it reduces to the *sequential normal compactness* (SNC) property of sets; see [13] and the references therein. The latter property is closely related to the compactly epi-Lipschitzian (CEL) property of [2]; see [9] and [7] for more details and recent developments. The same kinds of relationships hold in the case of set-valued mappings from Definition 4.6 under some inner semicontinuity (uniformity) conditions on S around $(\bar{z},\bar{x})$. Note that S is surely ISNC at $(\bar{z},\bar{x})$ if the following condition holds (cf. [11] in the case of fixed sets): there exist $\gamma,\sigma > 0$ and a compact set $C \subset X$ such that, for any $(z,x) \in \operatorname{gph} S \cap \big((\bar{z},\bar{x}) + \gamma\mathbb{B}_{Z\times X}\big)$, one has

$$\widehat{N}(x; S(z)) \subset \Big\{ x^* \in X^* \Big|\ \sigma\|x^*\| \le \max_{c\in C} |\langle x^*, c\rangle| \Big\}.$$

Note also that the ISNC property of S is generally different from the SNC property of multifunctions, which means that the graph of S is SNC at the reference point; see [13].

Now we are ready to formulate and prove the limiting extremal principle for extremal systems of multifunctions.

THEOREM 4.7. *Let* $S_i\colon M_i \rightrightarrows X$, $i = 1,\dots,p$, *be multifunctions from metric spaces* M_i *into an Asplund space* X. *Assume that* $\bar{x}$ *is an extremal point of the system* $(S_1,\dots,S_p)$ *at* $(\bar{s}_1,\dots,\bar{s}_p)$, *where each* S_i *is closed-valued around* $\bar{s}_i$ *and all but one of them are ISNC at the corresponding points* $(\bar{s}_i,\bar{x})$ *of their graphs. Then there are*

$$x_i^* \in \widetilde{N}(\bar{x}; S_i(\bar{s}_i)), \quad i = 1,\dots,p, \tag{4.10}$$

not all zero, satisfying the generalized Euler equation

$$x_1^* + \dots + x_p^* = 0. \tag{4.11}$$

Proof. It easily follows from Theorem 4.1 that for any $k \in \mathbb{N}$ there are s_{ik} with $d(s_{ik},\bar{s}_i) \le \frac{1}{k}$, $x_{ik} \in \bar{x} + \frac{1}{k}\mathbb{B}_X$, and $x_{ik}^* \in \widehat{N}(x_{ik}; S_i(s_{ik}))$, $i = 1,\dots p$, such that

$$\|x_{1k}^*\| + \dots + \|x_{pk}^*\| \ge 1 - 1/k \quad \text{and} \quad \|x_{1k}^* + \dots + x_{pk}^*\| \le 1/k. \tag{4.12}$$

By normalization if necessary we can always select bounded sequences $\{x_{ik}^*\}$, $i = 1,\dots,p$, satisfying (4.12). It is well known that bounded sets are sequentially weak* compact in X^* when X is Asplund; see, e.g., [20]. Thus, without loss of generality, we may assume that there are $x_i^* \in X^*$ such that $x_{ik}^* \overset{w^*}{\to} x_i^*$ as $k \to \infty$ for all $i = 1,\dots,p$.

Now passing to the limit as $k \to \infty$ and using definition (4.6), we arrive at the desired relationships (4.10) and (4.11). It remains to show that $x_1^*, \ldots, x_p^*$ are not equal to zero simultaneously.

Suppose on the contrary that all x_i^* are zero and assume for definiteness that the first $p-1$ mappings S_i are ISNC at $(\bar{s}_i, \bar{x})$, $i = 1, \ldots, p-1$. Then $\|x_{ik}^*\| \to 0$ as $k \to \infty$ for $i = 1, \ldots, p-1$. Passing to the limit at the second relation in (4.12), we conclude that $\|x_{pk}^*\| \to 0$ as well. But this clearly contradicts the first relation in (4.12) for large $k \in \mathbb{N}$, which ends the proof of the theorem. □

It is known [4] that the SNC property is essential for the fulfillment of the limiting extremal principle even in the case of fixed convex sets in infinite dimensions when both cones (2.5) and (4.6) reduce to the normal cone of convex analysis. Note also that the extended normal cone (4.6) cannot be generally replaced in (4.10) by the limiting one (2.5) unless the corresponding mapping S_i is assumed to be normally semicontinuous. Indeed, consider the extremal system of mappings (S_1, S_2) from Example 3.4. One can easily check that S_1 and S_2 are not normally semicontinuous at the origin and that

$$N(0; S_1(0)) \cap \big[-N(0; S_2(0)) \big] = \{0\}.$$

Hence an analogue of Theorem 4.7 with $\widetilde{N}$ replaced by N does not hold for this extremal system.

5. Necessary conditions in multiobjective optimization. This section is completely devoted to applications of the extended extremal principle to problems of multiobjective optimization. We start with the multiobjective problem formulated in Example 3.5, which contains only set/geometric constraints. Necessary conditions for this problem can be derived directly from the extremal principle. Then we consider a more general multiobjective problem with additional functional (equality and inequality) constraints and obtain various forms of necessary optimality conditions for this problem depending on the assumptions made on its data. To do this, we use the reduction to problems with no functional constraints and the advanced tools of generalized differentiation discussed in section 2. Recall that all the spaces considered below are assumed to be Asplund.

The initial multiobjective problem of our study is

$$\begin{aligned} \mathcal{M}_0 \qquad & \text{Minimize} \quad f(x) \\ & \text{subject to} \quad x \in C, \end{aligned}$$

where $f\colon X \to Y$ is a vector-valued objective function that is "minimized" with respect to the general preference $\prec$ from Example 3.5. That is, $\bar{x}$ is a (local) *solution* to $\mathcal{M}_0$, provided that $\bar{x} \in C$ and there is no any other element $x \in C$ close to $\bar{x}$ with $f(x) \prec f(\bar{x})$. For simplicity we consider global solutions to $\mathcal{M}_0$, although the following proposition and subsequent results hold for local solutions as well.

PROPOSITION 5.1. *Let $\bar{x}$ be a solution to $\mathcal{M}_0$, where f is assumed to be locally Lipschitzian around $\bar{x}$. Then for any $\varepsilon > 0$ there are $(x_0, x_1, y_0, y_1, x^*, y^*) \in X^2 \times Y^2 \times X^* \times Y^*$ satisfying*

$$x_0, x_1 \in \bar{x} + \varepsilon \mathbb{B}_X, \quad y_0, y_1 \in f(\bar{x}) + \varepsilon \mathbb{B}_Y, \quad x^* \in \widehat{N}(x_1; C), \quad y^* \in \widehat{N}(y_1; \operatorname{cl} \mathcal{L}(y_0))$$

with $\|y^*\| = 1$, and

$$0 \in x^* + \widehat{\partial} \langle y^*, f \rangle (x_0) + \varepsilon \mathbb{B}_{X^*}. \tag{5.1}$$

Proof. As in Example 3.5, we define

$$M_1 := \mathcal{L}(f(\bar{x})) \cup \{f(\bar{x})\},\ M_2 := \{0\},\ S_1(y) := C \times \operatorname{cl}\mathcal{L}(y),\ S_2 := \big\{(x', f(x'))\big|\ x' \in X\big\}$$

and observe that $(\bar{x}, f(\bar{x}))$ is an extremal point of the system (S_1, S_2) at $(f(\bar{x}), 0)$. Given $\varepsilon > 0$ and a Lipschitz constant $L > 0$ of f around $\bar{x}$, we choose

$$\varepsilon' := \min\big\{2\varepsilon L/(1+L),\ \varepsilon/2,\ 1/8(2+L),\ 1/2\big\}$$

and employ the extremal principle from Theorem 4.1. This gives $y_0 \in f(\bar{x}) + \varepsilon'\mathbb{B}_Y$, $(x_i, y_i) \in (\bar{x}, f(\bar{x})) + \varepsilon'\mathbb{B}_{X\times Y}$ for $i = 1, 2$, and

$$(x_1^*, y_1^*) \in \widehat{N}((x_1, y_1); S_1(y_0)), \quad (x_2^*, y_2^*) \in \widehat{N}((x_2, y_2); S_2)$$

satisfying the relations

$$\|(x_1^*, y_1^*)\| + \|(x_2^*, y_2^*)\| \ge 1 - \varepsilon' \ge 1/2, \quad \|(x_1^*, y_1^*) + (x_2^*, y_2^*)\| \le \varepsilon'. \tag{5.2}$$

By the definition of Fréchet normals (2.4) we have

$$0 \ge \langle x_2^*, x - x_2\rangle + \langle y_2^*, y - y_2\rangle - \varepsilon'\|(x - x_2, y - y_2)\|$$

for $(x, y) \in S_2$ sufficiently close to (x_2, y_2). Observing that $y_2 = f(x_2)$ and $y = f(x)$, we conclude that the function

$$\varphi(x) := -\langle x_2^*, x - x_2\rangle - \langle y_2^*, f(x) - f(x_2)\rangle + \varepsilon'\|(x - x_2, f(x) - f(x_2))\|$$

attains its local minimum at $x = x_2$. Now it follows from Proposition 2.1 that there is $x_0 \in x_2 + \varepsilon'\mathbb{B}_X \subset \bar{x} + 2\varepsilon'\mathbb{B}_X$ satisfying

$$0 \in x_2^* + \widehat{\partial}\langle y_2^*, f\rangle(x_0) + (2+L)\varepsilon'\mathbb{B}_{X^*}. \tag{5.3}$$

Using (5.3) and the second relation in (5.2), we get

$$0 \in x_1^* + \widehat{\partial}\langle y_1^*, f\rangle(x_0) + 2(12+L)\varepsilon'\mathbb{B}_{X^*}, \tag{5.4}$$

which implies $\|y_1^*\| \ge 1/4(1+L)$ due to (5.2), (5.4), and the choice of $\varepsilon' \le 1/8(2+L)$. Indeed, for any $\widetilde{x}^* \in \widehat{\partial}\langle y_1^*, f\rangle(x_0)$ we have $\|\widetilde{x}^*\| \le L\|y_1^*\|$ and then

$$\begin{aligned} 1/2 \le \|(x_1^*, y_1^*)\| &= \|x_1^*\| + \|y_1^*\| = \|\widetilde{x}^* + 2(2+L)\varepsilon' e^*\| + \|y_1^*\| \\ &\le L\|y_1^*\| + 2(2+L)\varepsilon' + \|y_1^*\| \end{aligned}$$

with some $e^* \in \mathbb{B}_{X^*}$. Solving the latter for $\|y_1^*\|$, we get

$$\|y_1^*\| \ge \frac{1/2 - 2(2+L)\varepsilon'}{(1+L)} \ge \frac{1}{4(1+L)},$$

which gives the required estimate. Now dividing (5.4) by $\|y_1^*\|$ and then putting $x^* := x_1^*/\|y_1^*\| \in \widehat{N}(x_1; C)$ and $y^* := y_1^*/\|y_1^*\| \in \widehat{N}(y_1; \operatorname{cl}\mathcal{L}(y_0))$, we finally arrive at (5.1) and finish the proof of this proposition. □

Remark 5.2. Let the preference $\prec$ be defined by a *cone* K, i.e., $x \prec y$ if and only if $y - x \in K$. Then $\mathcal{L}(y) = y - K$, and changing y reduces to translating the level sets. This implies that the sets $S_1(f(\bar{x}))$ and S_2 from the proof of Proposition 5.1

form the extremal system in the sense of (3.1), and thus one can use the conventional extremal principle [13]. However, this is not the case for the general preferences under consideration, which require the full strength of the extended extremal principle.

Next let us consider the main multiobjective optimization problem studied in this paper:

$$\begin{aligned} \mathcal{M} \qquad \text{Minimize } & f(x) \\ & \varphi_i(x) \le 0, \quad i = 1, \dots, m, \\ & \varphi_i(x) = 0, \quad i = m+1, \dots, n, \\ & x \in C, \end{aligned}$$

where $f\colon X \to Y$ is minimized with respect to the given preference $\prec$ on Y as in $\mathcal{M}_0$. For convenience we first present the following simple lemma concerning Fréchet normals to epigraphs that is used in the proof of the subsequent theorem.

LEMMA 5.3. *Let $\varphi\colon X \to \overline{\mathbb{R}}$ be a lower semicontinuous function finite at $\bar{x}$, and let $\bar{\nu} \ge \varphi(\bar{x})$. Then one has*

$$\widehat{N}((\bar{x}, \bar{\nu}); \operatorname{epi}\varphi) \subset \widehat{N}((\bar{x}, \varphi(\bar{x})); \operatorname{epi}\varphi).$$

If in addition φ is strictly differentiable at $\bar{x}$, then

$$\widehat{N}((\bar{x}, \varphi(\bar{x})); \operatorname{epi}\varphi) = N((\bar{x}, \varphi(\bar{x})); \operatorname{epi}\varphi) = \Big\{\lambda(\nabla\varphi(\bar{x}), -1)\Big|\ \lambda \ge 0\Big\}. \tag{5.5}$$

Proof. Take $(x^*, \lambda) \in \widehat{N}((\bar{x}, \bar{\nu}); \operatorname{epi}\varphi)$. By definition (2.4) one has

$$\langle (x^*, \lambda), (x - \bar{x}, \nu - \bar{\nu})\rangle + o\big(\|(x - \bar{x}, \nu - \bar{\nu})\|\big) \le 0 \tag{5.6}$$

whenever (x, ν) is sufficiently close to $(\bar{x}, \bar{\nu})$. Pick $(z, \mu) \in \operatorname{epi}\varphi$ close to $(\bar{x}, \varphi(\bar{x}))$; hence $(z, \bar{\nu} + \mu - \varphi(\bar{x})) \in \operatorname{epi}\varphi$ is close to $(\bar{x}, \bar{\nu})$. Substituting $(x, \nu) = (z, \bar{\nu} + \mu - \varphi(\bar{x}))$ into (5.6), we get

$$\langle (x^*, \lambda), (z - \bar{x}, \mu - \varphi(\bar{x}))\rangle + o\big(\|(z - \bar{x}, \mu - \varphi(\bar{x}))\|\big) \le 0$$

for all (z, μ) sufficiently close to $(\bar{x}, \varphi(\bar{x}))$, meaning that $(x^*, \lambda) \in \widehat{N}((\bar{x}, \varphi(\bar{x})); \operatorname{epi}\varphi)$.

If φ is strictly differentiable at $\bar{x}$, then it is locally Lipschitzian around this point, and hence $\partial^\infty\varphi(\bar{x}) = \{0\}$. Moreover, $\widehat{\partial}\varphi(\bar{x}) = \partial\varphi(\bar{x}) = \{\nabla\varphi(\bar{x})\}$ in this case. This implies (5.5) due to (2.6) and the well-known representation

$$\widehat{\partial}\varphi(\bar{x}) = \big\{x^* \in X^*\big|\ (x^*, -1) \in \widehat{N}((\bar{x}, \varphi(\bar{x})); \operatorname{epi}\varphi)\big\},$$

which finishes the proof. □

The following theorem gives *fuzzy* necessary optimality conditions for the multiobjective problem $\mathcal{M}$ with non-Lipschitzian functional constraints in terms of Fréchet normals to graphs and epigraphs of constraint functions.

THEOREM 5.4. *Let $\bar{x}$ be a (local) solution to $\mathcal{M}$, where C is locally closed, f is locally Lipschitzian, and φ_i are lower semicontinuous for $i = 1, \dots, m$ and continuous for $i = m+1, \dots, n$ around $\bar{x}$. Then for any $\varepsilon > 0$ there are $x_0 \in \bar{x} + \varepsilon\mathbb{B}_X$,*

$$\begin{aligned} & (x_i, \varphi_i(x_i)) \in (\bar{x}, \varphi_i(\bar{x})) + \varepsilon\mathbb{B}_{X\times\mathbb{R}}, \quad i = 1, \dots, n, \\ & x_{n+1} \in C \cap (\bar{x} + \varepsilon\mathbb{B}_X), \quad y_0, y_1 \in f(\bar{x}) + \varepsilon\mathbb{B}_Y, \\ & \lambda_0 \in [0, 1], \quad (x_i^*, -\lambda_i) \in \widehat{N}((x_i, \varphi_i(x_i)); \operatorname{epi}\varphi_i),\ i = 1, \dots, m, \end{aligned} \tag{5.7}$$

$$(x_i^*,-\lambda_i)\in\widehat{N}((x_i,\varphi_i(x_i));\operatorname{gph}\varphi_i),\ i=m+1,\ldots,n,\quad x_{n+1}^*\in\widehat{N}(x_{n+1};C), \tag{5.8}$$

and $y^*\in\widehat{N}(y_1;\operatorname{cl}\mathcal{L}(y_0))$ *satisfying*

$$0\in\lambda_0\widehat{\partial}\langle y^*,f\rangle(x_0)+\sum_{i=1}^{n+1}x_i^*+\varepsilon\mathbb{B}_{X^*} \tag{5.9}$$

with the nontriviality conditions

$$\lambda_0+\sum_{i=1}^{n}\|(x_i^*,\lambda_i)\|+\|x_{n+1}^*\|=1,\quad \|y^*\|=1. \tag{5.10}$$

Proof. Let Ω be the set of all points in X satisfying the constraints in problem $\mathcal{M}$. Then $\bar{x}$ is a solution to the initial multiobjective problem $\mathcal{M}_0$ with the only geometric constraint $x\in\Omega$. Applying Proposition 5.1 to the latter problem, we find

$$x_0,\hat{x}\in\bar{x}+(\varepsilon/3)\mathbb{B}_X,\quad y_0,y_1\in f(\bar{x})+(\varepsilon/3)\mathbb{B}_Y,\quad x^*\in\widehat{N}(\hat{x};\Omega),\quad y^*\in\widehat{N}(y_1;\operatorname{cl}\mathcal{L}(y_0))$$

with $\|y^*\|=1$ and

$$0\in x^*+\widehat{\partial}\langle y^*,f\rangle(x_0)+(\varepsilon/3)\mathbb{B}_{X^*}. \tag{5.11}$$

From the definition of $\widehat{N}(\hat{x};\Omega)$ in (2.4) one has

$$\langle x^*,x-\hat{x}\rangle-(\varepsilon/3)\|x-\hat{x}\|\le 0\ \text{ for all }\ x\in\Omega\ \text{ near }\ \hat{x}.$$

Then $\hat{x}$ is a local solution to the standard minimization problem with a Lipschitzian objective and nonsmooth constraints:

$$\begin{aligned}&\text{Minimize } -\langle x^*,x-\hat{x}\rangle+(\varepsilon/3)\|x-\hat{x}\|\\ &\text{subject to } \varphi_i(x)\le 0,\quad i=1,\ldots,m,\\ &\qquad\quad \varphi_i(x)=0,\quad i=m+1,\ldots,n,\\ &\qquad\quad x\in C.\end{aligned}$$

Choose $\varepsilon'<\varepsilon/3(n+1)$ and apply [13, Theorem 5.1(i)] to the latter problem. In this way, taking Lemma 5.3 into account, we have

$$\begin{aligned}&(x_i,\varphi_i(x_i))\in(\bar{x},\varphi_i(\bar{x}))+\varepsilon'\mathbb{B}_{X\times\mathbb{R}},\quad i=1,\ldots,n,\\ &\quad x_{n+1}\in C\cap(\bar{x}+\varepsilon'\mathbb{B}_X),\quad y_0,y_1\in f(\bar{x})+\varepsilon'\mathbb{B}_Y,\end{aligned}$$

$\lambda_0\ge 0$, $(x_i^*,\lambda_i)\in X^*\times\mathbb{R}$, $i=1,\ldots,n$, and $x_{n+1}^*\in\widehat{N}(x_{n+1};C)$ satisfying (5.7) and (5.8) and such that

$$\lambda_0x^*\in\sum_{i=1}^{n+1}x_i^*+\Big(\frac{\varepsilon}{3}+\varepsilon'(n+1)\Big)\mathbb{B}_{X^*} \tag{5.12}$$

with the nontriviality condition in (5.10) obtained by normalization. Rescaling λ_0 if necessary, we have $\lambda_0\in[0,1]$. Finally, multiplying (5.11) by λ_0 and combining this with (5.12), we arrive at (5.9) and complete the proof of the theorem. □

Next we obtain *exact/limiting* necessary conditions for the multiobjective problem $\mathcal{M}$ with non-Lipschitzian constraints using the limiting normal and subgradient constructions defined in (2.2), (2.5), and (4.6).

THEOREM 5.5. *Let $\bar{x}$ be a solution to $\mathcal{M}$. Suppose, in addition to the assumptions of Theorem* 5.4, *that* $\dim Y < \infty$ *and all but one of the sets* $\operatorname{epi}\varphi_i$ *for* $i = 1, \dots, m$, $\operatorname{gph}\varphi_i$ *for* $i = m+1, \dots, n$, *and C are sequentially normally compact at the points* $(\bar{x}, \varphi_i(\bar{x}))$ *and* $\bar{x}$, *respectively. Then there are* $\lambda_0 \in [0,1]$,

$$\begin{aligned} &y^* \in \widetilde{N}(f(\bar{x}); \operatorname{cl}\mathcal{L}(f(\bar{x}))), \quad (x_i^*, -\lambda_i) \in N((\bar{x}, \varphi_i(\bar{x})); \operatorname{epi}\varphi_i),\ i = 1, \dots, m, \\ &(x_i^*, -\lambda_i) \in N((\bar{x}, \varphi_i(\bar{x})); \operatorname{gph}\varphi_i),\ i = m+1, \dots, n, \quad x_{n+1}^* \in N(\bar{x}; C) \end{aligned} \tag{5.13}$$

such that $\|y^*\| = 1$, $(\lambda_0, \dots, \lambda_n, x_1^*, \dots, x_{n+1}^*) \neq 0$, *and*

$$0 \in \lambda_0 \partial\langle y^*, f\rangle(\bar{x}) + \sum_{i=1}^{n+1} x_i^*. \tag{5.14}$$

Proof. Employing Theorem 5.4, we find sequences $x_k \to \bar{x}$, $\lambda_{0k} \to \lambda_0$, $x_{ik} \overset{\varphi_i}{\to} \bar{x}$ for $i = 1, \dots, n$, $x_{(n+1)k} \overset{C}{\to} \bar{x}$, $y_{ik} \to f(\bar{x})$ for $i = 0, 1$, $(x_{ik}^*, \lambda_{ik}) \overset{w^*}{\to} (x_i^*, \lambda_i)$ for $i = 1, \dots, n$, $x_{(n+1)k}^* \overset{w^*}{\to} x_{n+1}^*$, $y_k^* \to y^*$, and $x_k^* \in \widehat{\partial}\langle y_k^*, f\rangle(x_k)$ satisfying the relations

$$\begin{aligned} &y_k^* \in \widehat{N}(y_{1k}; \operatorname{cl}\mathcal{L}(y_{0k})) \ \text{ with } \ \|y_k^*\| = 1, \quad x_{(n+1)k}^* \in \widehat{N}(x_{(n+1)k}; C), \\ &(x_{ik}^*, -\lambda_{ik}) \in \widehat{N}((x_{ik}, \varphi_i(x_{ik})); \operatorname{epi}\varphi_i) \ \text{ for } \ i = 1, \dots, m, \\ &(x_{ik}^*, -\lambda_{ik}) \in \widehat{N}((x_{ik}, \varphi_i(x_{ik})); \operatorname{gph}\varphi_i) \ \text{ for } \ i = m+1, \dots, n, \end{aligned} \tag{5.15}$$

$$\|\lambda_{0k} x_k^* + x_{1k}^* + \cdots + x_{nk}^* + x_{(n+1)k}^*\| \to 0 \ \text{ as } \ k \to \infty, \ \text{ and} \tag{5.16}$$

$$\lambda_{0k} + \sum_{i=1}^{n} \|(x_{ik}^*, \lambda_{ik})\| + \|x_{(n+1)k}^*\| \geq 1 \ \text{ for all } \ k \in \mathbb{N}. \tag{5.17}$$

Since X is Asplund and the sequence $\{x_k^*\}$ is bounded (due to the Lipschitz continuity of f around $\bar{x}$), it is weak* sequentially compact. Thus we may assume that $x_k^* \overset{w^*}{\to} x^* \in X^*$ as $k \to \infty$. Let us show that $x^* \in \partial\langle y^*, f\rangle(\bar{x})$.

Indeed, it is easy to observe, since f is locally Lipschitzian, that the inclusion $x_k^* \in \widehat{\partial}\langle y_k^*, f\rangle(x_k)$ is equivalent to

$$(x_k^*, -y_k^*) \in \widehat{N}((x_k, f(x_k)); \operatorname{gph} f), \quad k \in \mathbb{N}.$$

Passing there to the limit as $k \to \infty$, we get the relation

$$(x^*, -y^*) \in N((\bar{x}, f(\bar{x})); \operatorname{gph} f),$$

which is equivalent to $x^* \in \partial\langle y^*, f\rangle(\bar{x})$; see [15, Theorem 5.2]. Passing to the limit in (5.15) and (5.16) as $k \to \infty$ and using the definitions of the normal cones (2.4) and (4.6), we arrive at (5.13) and (5.14). It is clear that $\|y^*\| = 1$, since Y is finite-dimensional. It remains to show that $(\lambda_0, \dots, \lambda_n, x_1^*, \dots, x_{n+1}^*) \neq 0$ under the SNC assumptions of the theorem.

Suppose the contrary and assume for definiteness that the sets $\operatorname{epi}\varphi_i$ for $i = 1,\dots,m$ and $\operatorname{gph}\varphi_i$ for $i = m+1,\dots,n$ are sequentially normally compact at $(\bar{x},\varphi_i(\bar{x}))$. Then $\|(x_{ik}^*,\lambda_i)\| \to 0$ as $k \to \infty$ for all $i = 1,\dots,n$. Since $\lambda_{0k} \to 0$, we get from (5.16) that $\|x_{n+1}^*\| \to 0$ as well. But this clearly contradicts (5.17) and completes the proof of the theorem. □

Note that Theorem 5.5 can also be proved by using the limiting extremal principle of Theorem 4.7. Now we present two important corollaries of Theorem 5.5. The first one gives necessary optimality conditions for $\mathcal{M}$ in terms of limiting subgradients and singular subgradients of the constraint functions and is actually equivalent to Theorem 5.5 due to subgradient representations of limiting normals to graphs and epigraphs. In what follows we use the quantities τ_i defined in section 2 before the formulation of Proposition 2.3.

COROLLARY 5.6. *Let $\bar{x}$ be a solution to $\mathcal{M}$ under the assumptions of Theorem* 5.5. *Then the following alternative holds:*

(i) *either there exist $x_i^* \in \partial^\infty(\tau_i\varphi_i)(\bar{x})$, $i = 1,\dots,n$, and $x_{n+1}^* \in N(\bar{x};C)$, not all zero, satisfying*

$$x_1^* + \dots + x_{n+1}^* = 0, \quad \textit{or}$$

(ii) *there exist $\lambda_i \ge 0$, $i = 0,\dots,n$, not all zero, and $y^* \in \widetilde{N}(f(\bar{x});\operatorname{cl}\mathcal{L}(f(\bar{x})))$ with $\|y^*\| = 1$ satisfying*

$$0 \in \lambda_0\partial\langle y^*,f\rangle(\bar{x}) + \sum_{i\in\{i|\ \lambda_i>0\}}^{n} \lambda_i\partial(\tau_i\varphi_i)(\bar{x}) + \sum_{i\in\{i|\ \lambda_i=0\}}^{n} \partial^\infty(\tau_i\varphi_i)(\bar{x}) + N(\bar{x};C)$$

with some quantities τ_i described in the end of section 2.

Proof. This follows directly from Theorem 5.5 due to relationships (2.6) and (2.7) between subgradients of functions and normals to their graphs and epigraphs. □

The second corollary concerns Lipschitzian constraints in problem $\mathcal{M}$ when the assumptions and relations of Theorem 5.5 can be essentially simplified.

COROLLARY 5.7. *Let $\bar{x}$ be a solution to $\mathcal{M}$, where the set C is locally closed and the functions $f\colon X \to \mathbb{R}^s$ and $\varphi_i\colon X \to \mathbb{R}$, $i = 1,\dots,n$, are Lipschitz continuous around $\bar{x}$. Then there are multipliers $\lambda_i \ge 0$, $i = 0,\dots,n$, not all zero, and $y^* \in \widetilde{N}(f(\bar{x});\operatorname{cl}\mathcal{L}(f(\bar{x})))$ with $\|y^*\| = 1$ such that*

$$\lambda_i\varphi_i(\bar{x}) = 0 \ \textit{ for } \ i = 1,\dots,m, \quad \textit{and}$$

$$0 \in \lambda_0\partial\langle y^*,f\rangle(\bar{x}) + \sum_{i=1}^{m} \lambda_i\partial\varphi_i(\bar{x}) + \sum_{i=m+1}^{n} \lambda_i\Big(\partial\varphi_i(\bar{x}) \cup \partial(-\varphi_i)(\bar{x})\Big) + N(\bar{x};C).$$

Proof. It follows from Corollary 5.6, since $\partial^\infty\varphi(\bar{x}) = \{0\}$ and since both sets $\operatorname{epi}\varphi$ and $\operatorname{gph}\varphi$ are sequentially normally compact for locally Lipschitzian functions. □

Note that in general the extended normal cone $\widetilde{N}$ to the level set of the preference in Theorem 5.5 and its corollaries cannot be replaced by the limiting normal cone N. The following example concerns multiobjective problems with no constraints in finite-dimensional spaces.

Example 5.8. Define a preference $\prec$ on $\mathbb{R}^2$ by $(x_1,x_2) \prec (y_1,y_2)$ if $|x_1| - 2|x_2| > |y_1| - 2|y_2|$. Given $f\colon \mathbb{R}^p \to \mathbb{R}^2$ as

$$f(x_1,\dots,x_p) := \left(2\operatorname{sign}(x_1)\left(\sum_{i=1}^{p} x_i^2\right)^{1/2}, \left(\sum_{i=1}^{p} x_i^2\right)^{1/2}\right),$$

we consider its optimization with respect to the preference $\prec$. Then $\bar{x} = 0$ is obviously a solution to this problem. One can easily check that Corollary 5.7 holds in this setting while its counterpart in terms of $N(\cdot; \operatorname{cl} \mathcal{L})$ does not.

Remark 5.9. When both spaces X and Y are finite-dimensional and the mapping $\operatorname{cl} \mathcal{L}(\cdot)$ is normally semicontinuous at the point $(f(\bar{x}), f(\bar{x}))$, the results of Corollaries 5.6 and 5.7 are derived in [24] by using a method similar to the proof of the extremal principle. That paper also contains qualification conditions ensuring that $\lambda_0 \neq 0$ in the above statements and discusses optimization problems with variational inequality constraints.

The results established in Theorem 5.5 and its corollaries extend to the multiobjective case the corresponding results for nonsmooth optimization problems with real-valued cost functions derived in [12] and [13] in finite-dimensional and Asplund spaces, respectively, by using the extremal principle for set systems (3.1). Finally in this section, we obtain an extension to non-Lipschitzian multiobjective problems of the results of Proposition 2.3 on necessary optimality conditions in the "weak fuzzy" form; cf. [3], [17], and [19]. The main difference between the next theorem and Theorem 5.4 is that, instead of Fréchet normal to graphs and epigraphs, we now use Fréchet subgradients of constraint functions. However, the price we pay is that, instead of a small dual ball as in (5.9), we have to involve a weak* neighborhood of the origin in the following conditions.

THEOREM 5.10. *Let $\bar{x}$ be a solution to $\mathcal{M}$, where C is locally closed, f is locally Lipschitzian, and φ_i are lower semicontinuous for $i = 1, \ldots, m$ and continuous for $i = m+1, \ldots, n$ around $\bar{x}$. Assume also that* (2.8) *and* (2.9) *are fulfilled. Then for any $\varepsilon > 0$ and any weak* neighborhood V of the origin in X^* there are $x_0 \in \bar{x} + \varepsilon \mathbb{B}_X$,*

$$\begin{aligned} (x_i, \varphi_i(x_i)) &\in (\bar{x}, \varphi_i(\bar{x})) + \varepsilon \mathbb{B}_{X \times \mathbb{R}}, \quad i = 1, \ldots, n, \\ x_{n+1} &\in C \cap (\bar{x} + \varepsilon \mathbb{B}_X), \quad y_0, y_1 \in f(\bar{x}) + \varepsilon \mathbb{B}_Y, \\ y^* &\in \widehat{N}(y_1; \operatorname{cl} \mathcal{L}(y_0)) \quad \textit{with} \quad \|y^*\| = 1, \end{aligned}$$

and multipliers $\lambda_i \geq 0$, $i = 1, \ldots, n$, not all zero, such that

$$0 \in \widehat{\partial} \langle y^*, f \rangle(x_0) + \sum_{i=1}^{n} \lambda_i \widehat{\partial}(\tau_i \varphi_i)(x_i) + \widehat{N}(x_{n+1}; C) + V \tag{5.18}$$

with some quantities τ_i described in the end of section 2.

Proof. Let Ω be a set of feasible points for problem $\mathcal{M}$. Then this problem can be rewritten in the form of problem $\mathcal{M}_0$ from the beginning of this section. Applying Proposition 5.1, for any $\varepsilon > 0$ we find $x_0, \hat{x} \in \bar{x} + \varepsilon \mathbb{B}_X$, $y_0, y_1 \in f(\bar{x}) + \varepsilon \mathbb{B}_Y$, $x^* \in \widehat{N}(\hat{x}; \Omega)$, and $y^* \in \widehat{N}(y_1; \operatorname{cl} \mathcal{L}(y_0))$ with $\|y^*\| = 1$ such that inclusion (5.1) holds. Now taking an arbitrary small $\varepsilon' > 0$ and using construction (2.4) for $x^* \in \widehat{N}(\hat{x}; \Omega)$, we have

$$\langle x^*, x - \hat{x} \rangle - \varepsilon' \|x - \hat{x}\| \leq 0 \ \text{ for all } \ x \in \Omega \cap U,$$

where $U \subset X$ is an appropriate neighborhood of $\hat{x}$. Thus $\hat{x}$ is a local solution to the constrained minimization problem $\mathcal{P}$ from section 2 with

$$\varphi_0(x) := -\langle x^*, x - \hat{x} \rangle + \varepsilon' \|x - \hat{x}\|.$$

Employing Proposition 2.3 and then Lemma 5.3, for any weak* neighborhood $V \subset X^*$ of the origin we get, under assumptions (2.8) and (2.9), points $(x_i, \varphi_i(x_i)) \in$

$(\hat{x},\varphi_i(\hat{x}))+\varepsilon'\mathbb{B}_{X\times\mathbb{R}}$, $i=1,\dots,n$, and $x_{n+1}\in\hat{x}+\varepsilon'\mathbb{B}_X$ as well as multipliers $\lambda_i\ge 0$, $i=1,\dots,n$, not all zero, satisfying the inclusion

$$x^*\in\sum_{i=1}^n\lambda_i\widehat{\partial}(\tau_i\varphi_i)(x_i)+\widehat{N}(x_{n+1};C)+V.$$

Substituting this into (5.1), we finish the proof. □

6. Multiplayer games. In this concluding section of the paper we briefly consider some applications of the extended extremal principle to a class of multiobjective games with many players. These can be roughly described as games with n players, where each player wants to choose a strategy $\bar{x}_i$ from a space X_i such that they $\prec_i$ optimize (with respect to the preference $\prec_i$ on Y) an objective function $f:X_1\times X_2\times\cdots\times X_n\to Y$ given all other players choices $\bar{x}_j$, $j\ne i$.

This is a general game setting that covers, in particular, the case when each of the players can have a different objective function $f_i\colon X_1\times\cdots\times X_n\to Y_i$. In the latter case one has $f:=(f_1,\dots,f_n)\colon X_1\times\cdots\times X_n\to Y:=Y_1\times\cdots\times Y_n$ with the ordering $\prec_i$ on Y defined by

$$y\prec_i v\ \text{ for }\ y,v\in Y\quad\text{provided that}\quad y_i\prec_i v_i\ \text{ for }\ y_i,v_i\in Y_i.$$

It is well known that an essential concept in all game theory is that of a *saddle point*. Let us give a generalized version of this concept for the above multiobjective setting, where $\prec$ stands for $(\prec_1,\dots,\prec_n)$.

DEFINITION 6.1. *A point $\bar{x}=(\bar{x}_1,\dots,\bar{x}_n)$ is a* local $\prec$-saddle point *of $f\colon X_1\times\cdots\times X_n\to Y$ if for each $i=1,\dots,n$ there is a neighborhood U_i of $\bar{x}_i$ such that*

$$f(\bar{x})\prec_i f(\bar{x}_1,\dots,\bar{x}_{i-1},x_i,\bar{x}_{i+1},\dots,\bar{x}_n)\quad\text{for all}\quad x_i\in U_i.$$

This can be different from the usual saddle point concept regardless of the space, as the following example shows.

Example 6.2. Consider the mapping $f:\mathbb{R}^4\to\mathbb{R}^2$ given by

$$f(x,y,u,v):=(x^2+u,-y^2-e^v).$$

Let us group the variables so that x and y are for player one and u and v are for player two. This means that $X_1=X_2=Y=\mathbb{R}^2$. The order $\prec_1$ on $Y=\mathbb{R}^2$ for player one is that $(w,z)\prec_1(w_1,z_1)$ if $w<w_1$ and $z\ge z_1$ or $w\le w_1$ and $z>z_1$. The order $\prec_2$ on $Y=\mathbb{R}^2$ for player two is that $(w,z)\prec_2(w_1,z_1)$ if $w<w_1$ and $z<z_1$. This is a *mixture of Pareto and weak Pareto optimality*. Any point of the form $(0,0,u,v)$ is a *$\prec$-saddle point* for these orderings.

Now we present necessary optimality conditions for multiobjective games (in the sense of finding saddle points from Definition 6.1) under additional constraints. For simplicity we consider only the case of geometric constraints. Given $f\colon X_1\times\cdots\times X_n\to Y$ and $\prec_i$ as above, we impose constraints $x_i\in C_i\subset X_i$ for each $i=1,\dots,n$ and consider a (local) $\prec$-saddle point $\bar{x}$ for game $\mathcal{G}$ under these constraints. Due to Definition 6.1 (taking constraints into account) we have the following multiobjective optimization problem for each player i:

$$\begin{aligned}\mathcal{M}_i\quad&\text{Minimize } f(\bar{x}_1,\dots,\bar{x}_{i-1},x_i,\bar{x}_{i+1},\dots,\bar{x}_n)\\&\text{subject to}\quad x_i\in C_i,\end{aligned}$$

where "minimization" is understood with respect to the preference $\prec_i$ on Y.

Denote $f_i(x_i) := f(\bar{x}_1, \dots, \bar{x}_{i-1}, x_i, \bar{x}_{i+1}, \dots, \bar{x}_n)$ and consider the level sets $\mathcal{L}_i(y)$ induced by the preferences $\prec_i$ on Y. Employing the results of section 5 based on the extended extremal principle, we get the following necessary optimality conditions for constrained multiobjective games.

THEOREM 6.3. *Let $\bar{x}$ be a local $\prec$-saddle point for game $\mathcal{G}$, where the preferences $\prec_i$ are locally satiated and almost transitive. Assume that the spaces $X_1, \dots, X_n$ and Y are Asplund, that f is locally Lipschitzian around $\bar{x}$, and that C_i are locally closed around $\bar{x}_i$ for all i. Then for any $\varepsilon > 0$ there exist points*

$$u_i, x_i \in \bar{x}_i + \varepsilon \mathbb{B}_{X_i}, \quad y_i, z_i \in f_i(\bar{x}_i) + \varepsilon \mathbb{B}_Y, \quad x_i^* \in \widehat{N}(x_i; C_i), \quad y_i^* \in \widehat{N}(z_i; \operatorname{cl} \mathcal{L}_i(y_i))$$

with $\|y_i^*\| = 1$ satisfying

$$0 \in x_i^* + \widehat{\partial} \langle y_i^*, f_i \rangle (u_i) + \varepsilon \mathbb{B}_{X_i^*} \quad \text{for each} \quad i = 1, \dots, n.$$

If in addition $\dim Y < \infty$, then there are $y_i^* \in \widetilde{N}(f_i(\bar{x}_i); \operatorname{cl} \mathcal{L}_i(f_i(\bar{x}_i)))$ with $\|y_i^*\| = 1$ such that

$$\partial \langle y_i^*, f_i \rangle (\bar{x}_i) \cap \big(- N(\bar{x}_i; C) \big) \neq \emptyset, \quad i = 1, \dots, n.$$

Proof. This follows from Proposition 5.1 and Theorem 5.5 applied to the multiobjective optimization problems $\mathcal{M}_i$ corresponding to the above definition of local $\prec$-saddle points for game $\mathcal{G}$. □

Acknowledgments. The authors are indebted to Alexander Kruger, Hristo Sendov, and two anonymous referees for valuable comments and remarks that helped them improve the original version of the paper.

REFERENCES

[1] S. BELLAASSALI AND A. JOURANI, *Necessary optimality conditions in multiobjective dynamic optimization*, SIAM J. Control Optim., to appear.

[2] J. M. BORWEIN AND H. M. STROJWAS, *Tangential approximations*, Nonlinear Anal., 9 (1985), pp. 1347–1366.

[3] J. M. BORWEIN, J. S. TREIMAN, AND Q. J. ZHU, *Necessary conditions for constrained optimization problems with semicontinuous and continuous data*, Trans. Amer. Math. Soc., 350 (1998), pp. 2409–2429.

[4] J. M. BORWEIN AND Q. J. ZHU, *A survey of subdifferential calculus with applications*, Nonlinear Anal., 38 (1999), pp. 687–773.

[5] I. EKELAND, *On the variational principle*, J. Math. Anal. Appl., 47 (1974), pp. 324–353.

[6] M. FABIAN, *Subdifferentiability and trustworthiness in the light of a new variational principle of Borwein and Preiss*, Acta Univ. Carolina, 30 (1989), pp. 51–56.

[7] M. FABIAN AND B. S. MORDUKHOVICH, *Sequential normal compactness versus topological normal compactness in variational analysis*, Nonlinear Anal., 54 (2003), pp. 1057–1067.

[8] A. D. IOFFE, *Calculus of Dini subdifferentials of functions and contingent derivatives of set-valued maps*, Nonlinear Anal., 8 (1984), pp. 517–539.

[9] A. D. IOFFE, *Coderivative compactness, metric regularity and subdifferential calculus*, in Constructive, Experimental, and Nonlinear Analysis, M. Théra, ed., CMS Conf. Proc. 27, AMS, Providence, RI, 2000, pp. 123–164.

[10] A. Y. KRUGER, *Strict (ε, δ)-semidifferentials and extremality conditions*, Optimization, 51 (2002), pp. 539–554.

[11] P. D. LOEWEN, *Limits of Fréchet normals in nonsmooth analysis*, in Optimization and Nonlinear Analysis, A. Ioffe et al. eds., Pitman Res. Notes Math. Ser. 244, Longman, Harlow, UK, 1992, pp. 178–188.

[12] B. S. MORDUKHOVICH, *Approximation Methods in Problems of Optimization and Control*, Nauka, Moscow, 1988.

[13] B. S. Mordukhovich, *The extremal principle and its applications to optimization and economics*, in Optimization and Related Topics, A. Rubinov and B. Glover, eds., Appl. Optim. 47, Kluwer Academic, Dordrecht, The Netherlands, 2001, pp. 343–369.
[14] B. S. Mordukhovich and Y. Shao, *Extremal characterizations of Asplund spaces*, Proc. Amer. Math. Soc., 124 (1996), pp. 197–205.
[15] B. S. Mordukhovich and Y. Shao, *Nonsmooth sequential analysis in Asplund spaces*, Trans. Amer. Math. Soc., 348 (1996), pp. 1235–1280.
[16] B. S. Mordukhovich and Y. Shao, *Fuzzy calculus for coderivatives of multifunctions*, Nonlinear Anal., 29 (1997), pp. 605–626.
[17] B. S. Mordukhovich and B. Wang, *Necessary suboptimality and optimality conditions via variational principles*, SIAM J. Control Optim., 41 (2002), pp. 623–640.
[18] H. V. Ngai and M. Théra, *Metric regularity, subdifferential calculus and applications*, Set-Valued Anal., 9 (2001), pp. 187–216.
[19] H. V. Ngai and M. Théra, *A fuzzy necessary optimality condition for non-Lipschitz optimization in Asplund spaces*, SIAM J. Optim., 12 (2002), pp. 656–668.
[20] R. R. Phelps, *Convex Functions, Monotone Operators and Differentiability*, 2nd ed., Springer-Verlag, Berlin, 1993.
[21] R. A. Poliquin, R. T. Rockafellar, and L. Thibault, *Local differentiability of distance functions*, Trans. Amer. Math. Soc., 332 (2000), pp. 5231–5249.
[22] R. T. Rockafellar and R. J.-B. Wets, *Variational Analysis*, Springer-Verlag, Berlin, 1998.
[23] L. Thibault, *personal communication.*
[24] J. J. Ye and Q. J. Zhu, *Multiobjective optimizations problems with variational inequality constraints*, Math. Programming, to appear.
[25] Q. J. Zhu, *The equivalence of several basic theorems for subdifferentials*, Set-Valued Anal., 6 (1998), pp. 171–185.
[26] Q. J. Zhu, *Hamiltonian necessary conditions for a multiobjective optimal control problem with endpoint constraints*, SIAM J. Control Optim., 39 (2000), pp. 97–112.

SIAM J. OPTIM.
Vol. 14, No. 2, pp. 380–401

LIMITED-MEMORY REDUCED-HESSIAN METHODS FOR LARGE-SCALE UNCONSTRAINED OPTIMIZATION*

PHILIP E. GILL† AND MICHAEL W. LEONARD†

Abstract. Limited-memory BFGS quasi-Newton methods approximate the Hessian matrix of second derivatives by the sum of a diagonal matrix and a fixed number of rank-one matrices. These methods are particularly effective for large problems in which the approximate Hessian cannot be stored explicitly.

It can be shown that the conventional BFGS method accumulates approximate curvature in a sequence of expanding subspaces. This allows an approximate Hessian to be represented using a smaller *reduced* matrix that increases in dimension at each iteration. When the number of variables is large, this feature may be used to define *limited-memory* reduced-Hessian methods in which the dimension of the reduced Hessian is limited to save storage. Limited-memory reduced-Hessian methods have the benefit of requiring half the storage of conventional limited-memory methods.

In this paper, we propose a particular reduced-Hessian method with substantial computational advantages compared to previous reduced-Hessian methods. Numerical results from a set of unconstrained problems in the CUTE test collection indicate that our implementation is competitive with the limited-memory codes L-BFGS and L-BFGS-B.

Key words. unconstrained optimization, quasi-Newton methods, BFGS method, reduced-Hessian methods, conjugate-direction methods

AMS subject classifications. 65K05, 90C30

DOI. 10.1137/S1052623497319973

1. Introduction.

BFGS quasi-Newton methods have proved reliable and efficient for the unconstrained minimization of a smooth nonlinear function $f : \mathbb{R}^n \to \mathbb{R}$. However, the need to store an $n \times n$ approximate Hessian has limited their application to problems with a small-to-moderate number of variables (say, less than 500). For larger n it is necessary to use methods that do not require the storage of a full $n \times n$ matrix. Sparse quasi-Newton updates can be applied if the Hessian has a significant number of zero entries (see, e.g., Powell and Toint [29], Fletcher [10]). However, if the Hessian is dense, as is often the case for certain subproblems arising in nonlinearly constrained optimization, other methods must be used. Such methods include conjugate-gradient methods, limited-memory quasi-Newton methods, and limited-memory reduced-Hessian quasi-Newton methods.

Conjugate-gradient methods require storage for only a few n-vectors (see, e.g., Gill, Murray, and Wright [17, pp. 144–150]). These methods can be equivalent to the BFGS quasi-Newton method on a quadratic function, but they are generally acknowledged to be less robust on general nonlinear problems (see Gill and Murray [14] for some numerical comparisons). Limited-memory quasi-Newton methods also require storage of few n-vectors but have a more explicit relationship with quasi-Newton methods. Limited-memory methods exploit the fact that the approximate Hessian (or its inverse) can be written as the sum of a diagonal matrix and a number

*Received by the editors April 17, 1997; accepted for publication (in revised form) March 19, 2003; published electronically October 2, 2003.

http://www.siam.org/journals/siopt/14-2/31997.html

†Department of Mathematics, University of California, San Diego, La Jolla, CA 92093-0112 (pgill@ucsd.edu, mleonard@na-net.ornl.gov). The research of the first author was supported by National Science Foundation grants DMI-9424639, CCR-9896198, and DMS-9973276 and Office of Naval Research grant N00014-96-1-0274. The research of the second author was supported by National Science Foundation grant DMI-9424639.

of rank-one matrices. This allows the search direction to be calculated as a simple linear combination of the vectors that define each rank-one update. The idea of a limited-memory method is to store a fixed number m ($m \ll n$) of pairs of update vectors and to discard older pairs as new ones are computed. These methods appeared in the early 1980s (see, e.g., Shanno [30] and Nocedal [26]), and they have now been developed to a considerable level of sophistication (see Byrd, Nocedal, and Schnabel [5] and Kaufman [19]). Limited-memory approximate Hessians may be used directly in a conventional quasi-Newton method, or they may be used as preconditioners for a nonlinear conjugate-gradient method (see, e.g., Buckley [2, 3], Gill, Murray, and Wright [17, pp. 151–152], Morales and Nocedal [22], and Nazareth [25]).

A different approach has been taken by Fenelon [8] and Siegel [33], who independently proposed methods in which the curvature is accumulated in a subspace spanned by a set of m independent vectors. These *reduced-Hessian* methods exploit the fact that quasi-Newton methods accumulate approximate curvature in a sequence of expanding subspaces (see Gill and Leonard [13]). Reduced-Hessian methods represent the approximate Hessian using a smaller *reduced* matrix that increases in dimension at each iteration. This reduced matrix incorporates curvature information that has been accumulated during earlier iterations and allows the search direction to be calculated from a linear system that is smaller than that used in conventional methods.

In this paper we propose the limited-memory method L-RHR, which may be viewed as a limited-memory variant of the reduced-Hessian method RHR of Gill and Leonard [13]. L-RHR has two features in common with the limited-memory method of Siegel [33]: a basis of search directions is maintained for the sequence of m-dimensional subspaces, and an implicit orthogonal decomposition is used to define an orthonormal basis for each subspace. However, L-RHR is different from Seigel's algorithm in several ways: (i) L-RHR updates the Cholesky factor of the reduced Hessian instead of updating an explicit reduced inverse Hessian; (ii) the formulation of L-RHR as a modification of RHR allows the application of *Hessian reinitialization*, which is shown to greatly enhance performance on large problems; and (iii) L-RHR employs *selective basis reorthogonalization* to improve robustness for moderate values of the subspace dimension. Property (i) implies that in exact arithmetic, even if implemented without Hessian reinitialization and selective reorthogonalization, the L-RHR iterates are different from those of Siegel's method (see section 3.6). Properties (i)–(iii) not only provide substantial improvements in efficiency compared to Siegel's method, but also make reduced-Hessian methods competitive with the state-of-the-art limited-memory method L-BFGS-B of Zhu et al. [34]. L-RHR requires the storage of an $n \times m$ matrix, two $m \times m$ nonsingular upper-triangular matrices, and a fixed number of n- and m-vectors. For a given m, this is approximately half the storage required for L-BFGS-B to represent essentially the same amount of second-derivative information. Moreover, L-RHR requires fewer floating-point operations per iteration, which results in smaller overall computation times on many problems.

The paper is organized as follows. In section 2 we briefly review various theoretical aspects of reduced-Hessian quasi-Newton methods, including the definition of Algorithm RHR, a reduced-Hessian method with Hessian reinitialization. Algorithm RHR provides the theoretical framework for the limited-memory algorithm L-RHR proposed in section 3. We give algorithms for maintaining both gradient- and search-direction subspace bases, and it is shown that L-RHR has the property of finite termination on a strictly convex quadratic function. To simplify the discussion, the algorithms of sections 2–3 are stated with the assumption that all computations are performed in

exact arithmetic. The effects of rounding error and the use of reorthogonalization are discussed in sections 4.1, 4.2, and 4.3. Finally, section 5 includes some numerical results obtained when various limited-memory reduced-Hessian algorithms are applied to test problems from the CUTE test collection of Bongartz et al. [1]. It is shown that reinitialization and selective reorthogonalization (in conjunction with an explicit factorization for the subspace basis) give, respectively, significantly fewer function evaluations and increased robustness compared to Siegel's method. Section 5 also includes comparisons of L-RHR with two alternative implementations of the conventional limited-memory BFGS method.

Unless explicitly indicated otherwise, $\|\cdot\|$ denotes the vector two-norm or its subordinate matrix norm.

2. Motivation. The BFGS method generates a sequence of iterates $\{x_k\}$ such that $x_{k+1} = x_k + \alpha_k p_k$, where p_k is the search direction and α_k is a scalar step length. The search direction satisfies $H_k p_k = -\nabla f(x_k)$, where H_k is an approximate Hessian. The application of the BFGS update to H_k gives a matrix H_k' such that

$$H_k' = H_k - \frac{1}{\delta_k^T H_k \delta_k} H_k \delta_k \delta_k^T H_k + \frac{1}{\gamma_k^T \delta_k} \gamma_k \gamma_k^T, \tag{2.1}$$

where $\delta_k = x_{k+1} - x_k$, $g_k = \nabla f(x_k)$, and $\gamma_k = g_{k+1} - g_k$. A conventional BFGS method then defines $H_{k+1} = H_k'$ (another choice for H_{k+1} is discussed in section 2.2). If H_0 is symmetric and positive definite, and if α_k is such that the approximate curvature $\gamma_k^T \delta_k$ is positive, then H_k is symmetric positive definite for all $k \geq 0$. Conditions imposed on the step length by practical step-length algorithms can ensure both positivity of the approximate curvature and sufficient descent. This is the case, for example, for any α_k satisfying the Wolfe conditions

$$f(x_k + \alpha_k p_k) \leq f(x_k) + \mu \alpha_k g_k^T p_k \quad \text{and} \quad g_{k+1}^T p_k \geq \eta g_k^T p_k, \tag{2.2}$$

where the constants μ and η are chosen so that $0 \leq \mu < \eta < 1$ and $\mu < \frac{1}{2}$.

The need to solve a linear system for p_k makes it convenient to use the upper-triangular Cholesky factor C_k such that $H_k = C_k^T C_k$. In this case, the Cholesky factor C_{k+1} of H_{k+1} is obtained from a rank-one change to C_k (see Dennis and Schnabel [7]). We omit the details of this procedure and simply write $C_{k+1} = \boldsymbol{update}(C_k, \delta_k, \gamma_k)$.

2.1. Reduced-Hessian methods. Reduced-Hessian methods provide an alternative way of implementing the BFGS method. Let $\mathcal{G}_k$ denote the subspace $\mathcal{G}_k = \text{span}\{g_0, g_1, \ldots, g_k\}$, and let $\mathcal{G}_k^\perp$ denote the orthogonal complement of $\mathcal{G}_k$ in $\mathbb{R}^n$. Reduced-Hessian methods are based on the following result (see, e.g., Fletcher and Powell [11], Fenelon [8], and Siegel [33]).

LEMMA 2.1. *Consider the BFGS method applied to a general nonlinear function. If $H_0 = \sigma I$ ($\sigma > 0$) and $H_k p_k = -g_k$, then $p_k \in \mathcal{G}_k$ for all k. Moreover, if $z \in \mathcal{G}_k$ and $w \in \mathcal{G}_k^\perp$, then $H_k z \in \mathcal{G}_k$ and $H_k w = \sigma w$.*

Let r_k denote $\dim(\mathcal{G}_k)$, and let B_k (B for "basis") denote an $n \times r_k$ matrix whose columns form a basis for $\mathcal{G}_k$. An orthonormal basis Z_k can be defined from the QR decomposition $B_k = Z_k T_k$, where T_k is a nonsingular upper-triangular matrix. Let the $n - r_k$ columns of W_k define an orthonormal basis for $\mathcal{G}_k^\perp$. If Q_k is the orthogonal matrix $Q_k = \begin{pmatrix} Z_k & W_k \end{pmatrix}$, then the transformation $x = Q_k x_Q$ defines a transformed approximate Hessian $Q_k^T H_k Q_k$ and a transformed gradient $Q_k^T g_k$. If $H_0 = \sigma I$ ($\sigma > 0$), it follows from (2.1) and Lemma 2.1 that the transformation induces a block-diagonal

structure, with

$$Q_k^T H_k Q_k = \begin{pmatrix} Z_k^T H_k Z_k & 0 \\ 0 & \sigma I_{n-r_k} \end{pmatrix} \quad \text{and} \quad Q_k^T g_k = \begin{pmatrix} Z_k^T g_k \\ 0 \end{pmatrix}. \tag{2.3}$$

The positive-definite matrix $Z_k^T H_k Z_k$ is known as a reduced approximate Hessian (or just reduced Hessian). The vector $Z_k^T g_k$ is known as a reduced gradient.

If we write the equation for the search direction as $(Q_k^T H_k Q_k) Q_k^T p_k = -Q_k^T g_k$, it follows from (2.3) that

$$p_k = Z_k q_k, \text{ where } q_k \text{ satisfies } Z_k^T H_k Z_k q_k = -Z_k^T g_k. \tag{2.4}$$

If the Cholesky factorization $Z_k^T H_k Z_k = R_k^T R_k$ is known, q_k can be computed from the forward substitution $R_k^T d_k = -Z_k^T g_k$ and back-substitution $R_k q_k = d_k$. The practical benefit of this approach is that, if $k \ll n$, the matrices Z_k and R_k require much less storage than H_k.

There are a number of alternative choices for B_k (see Gill and Leonard [13, Theorem 2.3]). Both Fenelon and Siegel propose that B_k be formed from a linearly independent subset of $\{g_0, g_1, \ldots, g_k\}$. With this choice, the orthonormal basis can be accumulated columnwise as the iterations proceed using Gram–Schmidt orthogonalization (see, e.g., Golub and Van Loan [18, pp. 218–220]). During iteration k, the number of columns of Z_k either remains unchanged or increases by one, depending on the value of the scalar ρ_{k+1} such that $\rho_{k+1} = \|(I - Z_k Z_k^T) g_{k+1}\|$. If $\rho_{k+1} = 0$, the new gradient has no component outside range(Z_k) and g_{k+1} is said to be *rejected*. Thus, if $\rho_{k+1} = 0$, Z_k already provides a basis for $\mathcal{G}_{k+1}$ with $r_{k+1} = r_k$ and $Z_{k+1} = Z_k$. Otherwise, $r_{k+1} = r_k + 1$ and the gradient g_{k+1} is said to be *accepted*. In this case, Z_k gains a new column z_{k+1} defined by the identity $\rho_{k+1} z_{k+1} = (I - Z_k Z_k^T) g_{k+1}$. The calculation of z_{k+1} also provides the r_k-vector $u_k = Z_k^T g_{k+1}$ and the scalar $z_{k+1}^T g_{k+1}$ $(= \rho_{k+1})$, which are the components of the reduced gradient $Z_{k+1}^T g_{k+1}$ for the next iteration. For simplicity, we write $(Z_{k+1}, u_k, \rho_{k+1}, r_{k+1}) = \textbf{\textit{orthog}}(Z_k, g_{k+1}, r_k)$ in the algorithms that follow. This orthogonalization procedure requires approximately $2nr_k$ flops. Gram–Schmidt orthogonalization may be considered as an algorithm for computing the QR decomposition of B_k without storing T_k. Suppose that at the start of iteration k there exists a nonsingular T_k with $B_k = Z_k T_k$. If g_{k+1} is accepted, then

$$B_{k+1} = \begin{pmatrix} B_k & g_{k+1} \end{pmatrix} = \begin{pmatrix} Z_k & z_{k+1} \end{pmatrix} \begin{pmatrix} T_k & Z_k^T g_{k+1} \\ 0 & \rho_{k+1} \end{pmatrix} = Z_{k+1} T_{k+1}, \tag{2.5}$$

where the last equality defines T_{k+1}, which is nonsingular since $\rho_{k+1} \neq 0$. Otherwise, $T_{k+1} = T_k$.

Definition (2.4) of each search direction implies that $p_j \in \mathcal{G}_k$ for all $0 \leq j \leq k$. This leads naturally to another basis for $\mathcal{G}_k$ based on orthogonalizing the search directions $p_0, p_1, \ldots, p_k$. The next theorem implies that the columns of Z_k constitute an orthonormal basis for $\mathcal{P}_k$, the span of *all* search directions $\{p_0, p_1, \ldots, p_k\}$ (for a proof, see Gill and Leonard [13]).

THEOREM 2.2. *If $H_0 = \sigma I$ $(\sigma > 0)$, then the subspaces $\mathcal{G}_k$ and $\mathcal{P}_k$ generated by the gradients and search directions of the conventional BFGS method are identical.*

This result implies that Z_k can be generated from either gradients or search directions, a point that will be used to advantage in section 3.

Given Z_{k+1} and H_k, the calculation of the search direction for the next iteration requires the Cholesky factor of $Z_{k+1}^T H_{k+1} Z_{k+1}$.[1] This factor can be obtained from R_k in a two-step process without the need to know H_k. The first step, which is not needed if g_{k+1} is rejected, is to compute the factor R_k' of $Z_{k+1}^T H_k Z_{k+1}$. This step involves adding a row and column to R_k to account for the new last column of Z_{k+1}. It follows from Lemma 2.1 and (2.3) that

$$Z_{k+1}^T H_k Z_{k+1} = \begin{pmatrix} Z_k^T H_k Z_k & Z_k^T H_k z_{k+1} \\ z_{k+1}^T H_k Z_k & z_{k+1}^T H_k z_{k+1} \end{pmatrix} = \begin{pmatrix} Z_k^T H_k Z_k & 0 \\ 0 & \sigma \end{pmatrix},$$

giving an expanded block-diagonal factor R_k' defined by

$$R_k' = \begin{cases} R_k & \text{if } r_{k+1} = r_k; \\ \begin{pmatrix} R_k & 0 \\ 0 & \sigma^{1/2} \end{pmatrix} & \text{if } r_{k+1} = r_k + 1. \end{cases} \tag{2.6}$$

The algorithm that defines R_k' from R_k will be denoted by ***expand*** for obvious reasons. This expansion also involves vectors $v_k = Z_k^T g_k$, $u_k = Z_k^T g_{k+1}$, and $q_k = Z_k^T p_k$, which are updated to give $v_k' = Z_{k+1}^T g_k$, $u_k' = Z_{k+1}^T g_{k+1}$, and $q_k' = Z_{k+1}^T p_k$. As both p_k and g_k lie in range(Z_k), if g_{k+1} is accepted, the vectors v_k' and q_k' are trivially defined from v_k and q_k by appending a zero component (see (2.3)). Similarly, the vector u_k' is formed from u_k and ρ_{k+1}. If g_{k+1} is rejected, $v_k' = v_k$, $u_k' = u_k$, and $q_k' = q_k$. In either case, v_{k+1} is equal to u_k' and need not be calculated at the start of iteration $k+1$ (see Algorithm 2.1 below).

The second step of the modification alters R_k' to reflect the BFGS update to H_k. This update gives a modified factor $R_k'' = \textbf{\textit{update}}(R_k', s_k, y_k)$, where $s_k = Z_{k+1}^T(x_{k+1} - x_k) = \alpha_k q_k'$, and $y_k = Z_{k+1}^T(g_{k+1} - g_k) = u_k' - v_k'$.

2.2. Reinitialization. The initial approximate Hessian can greatly influence the practical performance of quasi-Newton methods. The usual choice $H_0 = \sigma I$ ($\sigma > 0$) can result in many iterations and function evaluations—especially if the iterates tend toward a minimizer at which the Hessian of f is ill-conditioned (see, e.g., Powell [27] and Siegel [33]). This is sometimes associated with "stalling" of the iterates, a phenomenon that can greatly increase the overall cpu time for solution (or termination). The form of the transformed Hessian $Q_k^T H_k Q_k$ (see (2.3)) reveals the influence of H_0 on the approximate Hessian. In particular, the scale factor σ represents the approximate curvature along all directions in $\mathcal{G}_k^{\perp}$. However, in the reduced-Hessian formulation, this initial approximate curvature is not installed until the end of iteration k, when it is used in the ***expand*** procedure according to (2.6). Our idea is to replace σ whenever g_{k+1} is accepted with a value more representative of the approximated curvature. This has the effect of *reinitializing* the approximate curvature along z_{k+1} and is meant to alleviate inefficiencies resulting from poor choices of H_0. An estimate σ_k of the approximate curvature is maintained and updated as new curvature information is obtained. Some popular choices for σ_k are considered by Leonard [20, pp. 44–48]. In section 5 we discuss values that have been proposed for limited-memory methods.

The initial approximate curvature can be reinitialized by using some σ_{k+1} in place of σ_k in the ***expand*** procedure. Gill and Leonard [13] show that reinitialization can

[1] As mentioned earlier, H_{k+1} is usually H_k', which is defined by (2.1). However, we will implicitly alter H_k' further, as described in section 2.2.

be done either before or after the ***expand***, but they recommend the latter because it results in a simpler convergence result. Here, the reinitialization is performed after ***update*** to be consistent with that article. The procedure ***reinitialize*** involves simply changing the trailing diagonal element of R_k'' from $\sigma_k^{1/2}$ to $\sigma_{k+1}^{1/2}$ whenever g_{k+1} is accepted.

2.3. Summary. We conclude this section by defining a generic reduced-Hessian method that is the basis of the limited-memory method proposed in section 3. As described above, the reduced-Hessian method involves four main procedures: an *orthogonalize*, which determines Z_{k+1} using the Gram–Schmidt QR process; an *expand*, which increases the order of the reduced Hessian by one; an *update*, which applies a BFGS update directly to the reduced Hessian; and a *reinitialize*, which reinitializes the last diagonal of the reduced-Hessian factor.

ALGORITHM 2.1. (RHR) REDUCED-HESSIAN METHOD WITH REINITIALIZATION.
Choose x_0 and σ_0 ($\sigma_0 > 0$);
$k = 0$; $r_0 = 1$; $g_0 = \nabla f(x_0)$;
$Z_0 = \begin{pmatrix} g_0/\|g_0\| \end{pmatrix}$; $R_0 = \begin{pmatrix} \sigma_0^{1/2} \end{pmatrix}$; $v_0 = \|g_0\|$;
while not converged **do**
  Solve $R_k^T d_k = -v_k$; $R_k q_k = d_k$;
  $p_k = Z_k q_k$;
  Find α_k satisfying the Wolfe conditions (2.2);
  $x_{k+1} = x_k + \alpha_k p_k$; $g_{k+1} = \nabla f(x_k + \alpha_k p_k)$;
  $(Z_{k+1}, u_k, \rho_{k+1}, r_{k+1}) = \boldsymbol{orthog}(Z_k, g_{k+1}, r_k)$;
  $(R_k', u_k', v_k', q_k') = \boldsymbol{expand}(R_k, u_k, v_k, q_k, \rho_{k+1}, \sigma_k)$;
  $s_k = \alpha_k q_k'$; $y_k = u_k' - v_k'$; $R_k'' = \boldsymbol{update}(R_k', s_k, y_k)$;
  Compute σ_{k+1}; $R_{k+1} = \boldsymbol{reinitialize}(R_k'', \sigma_{k+1})$;
  $v_{k+1} = u_k'$;
  $k = k + 1$;
end do

When no reinitialization is done, this algorithm generates the same iterates as the conventional BFGS method with $H_0 = \sigma_0 I$. In exact arithmetic, the methods differ only in the storage needed and the number of operations per iteration. It can be shown that both with and without reinitialization, the algorithm retains two important properties of the BFGS method: it has quadratic termination, and it converges globally and Q-superlinearly on strongly convex functions (see Gill and Leonard [13]).

Algorithm RHR implicitly defines a full-sized BFGS approximate Hessian H_k. Let Z_k and R_k be defined at the start of the kth iteration, and let $Q_k = \begin{pmatrix} Z_k & W_k \end{pmatrix}$ denote an orthogonal matrix whose first r_k columns are the columns of Z_k. The full-sized approximate Hessian is given by

$$H_k = Q_k \begin{pmatrix} R_k^T & 0 \\ 0 & \sigma_k^{1/2} I_{n-r_k} \end{pmatrix} \begin{pmatrix} R_k & 0 \\ 0 & \sigma_k^{1/2} I_{n-r_k} \end{pmatrix} Q_k^T. \tag{2.7}$$

Given R_k, Z_k, and any n-vector v, the identity

$$H_k v = Z_k R_k^T R_k Z_k^T v + \sigma_k (I - Z_k Z_k^T) v$$

implies that products $H_k v$ can be calculated. This allows the reduced-Hessian approach to be used in constrained optimization algorithms that use H_k as an operator via products of the form $H_k v$ (see, e.g., Gill, Murray, and Saunders [15]).

3. A limited-memory reduced-Hessian method. In this section we propose a limited-memory method that may be viewed as a reduced-Hessian method in which only the most recent curvature information is retained. As in Algorithm RHR, a triangular factor of the reduced Hessian is updated and reinitialized at each iteration—the crucial difference is that the number of basis vectors (and hence the dimension of the reduced Hessian) is limited by a preassigned value m. For problems with many variables, a choice of $m \ll n$ gives significant savings in storage compared to conventional quasi-Newton methods.

A simple limited-memory version of RHR can be defined by discarding the oldest gradient when the storage limit is reached. However, algorithms based on this idea have proved to be inefficient in practice. One explanation of this inefficiency is that discarding the oldest gradient invalidates RHR's property of finite termination on a quadratic function (see Theorem 3.1 and the concluding remarks of section 3.7). There is considerable numerical evidence that quadratic termination is beneficial when minimizing general functions; see, e.g., Siegel [31] and Leonard [20]. The limited-memory method proposed here retains the property of finite termination by following Siegel's suggestion of using a basis of search directions rather than gradients. This strategy is sufficient to maintain quadratic termination when the oldest basis vector is discarded (see section 3.3).

An important feature of the method is that it is necessary to store and update the triangular factor T_k associated with the orthogonal factorization $B_k = Z_k T_k$. In practice, we store T_k and either Z_k or B_k.

3.1. The search-direction basis and its factorization. We start by describing how the orthogonal factorization is maintained as directions are added to the basis. Initially, this procedure is described in the context of building an m-dimensional basis before a search direction is discarded, where m is assumed to satisfy $m \geq 2$. The usual context is to add and remove a vector at every iteration. The procedure for removing a direction from the basis is described in section 3.2. To simplify the discussion, we assume that every gradient is accepted.

In order to allow for the fact that Z_k is used in the equations that define p_k, the gradient g_k is used as a temporary basis vector until it can be replaced by p_k. This implies that the kth iteration involves three basis matrices: B_k, B_k', and B_k''. The starting basis is $B_k = (p_0 \quad \cdots \quad p_{k-1} \quad g_k)$. The matrix B_k' is obtained from B_k by replacing g_k by p_k as soon as it is computed, and B_k'' is found by adding the accepted gradient to B_k'. The matrices B_k and B_k' differ by a single column, yet, by Theorem 2.2, their columns span the same subspace during the build process.

The procedure starts with $B_0 = \begin{pmatrix} g_0 \end{pmatrix}$, $T_0 = \begin{pmatrix} \|g_0\| \end{pmatrix}$, and $Z_0 = \begin{pmatrix} z_0 \end{pmatrix}$, where $z_0 = g_0/\|g_0\|$. Once p_0 is calculated, it is swapped into the basis to give $B_0' = \begin{pmatrix} p_0 \end{pmatrix}$ and $T_0' = \begin{pmatrix} \|p_0\| \end{pmatrix}$. After the line search, g_1 is accepted (by assumption) and we define

$$B_0'' = \begin{pmatrix} p_0 & g_1 \end{pmatrix}, \quad T_0'' = \begin{pmatrix} \|p_0\| & u_0 \\ 0 & \rho_1 \end{pmatrix}, \quad \text{and} \quad Z_0' = \begin{pmatrix} z_0 & z_1 \end{pmatrix}$$

(see (2.5) and recall that $u_k = Z_k^T g_{k+1}$). With our assumption that $m \geq 2$, no vector need be discarded and these matrices define B_1, T_1, and Z_1. The kth iteration ($1 \leq k \leq m-1$) proceeds in a similar way, with

$$B_k = \begin{pmatrix} p_0 & \cdots & p_{k-1} & g_k \end{pmatrix}, \quad T_k = \begin{pmatrix} \underline{T}_{k-1}' & v_k \end{pmatrix}, \quad \underline{T}_{k-1}' = \begin{pmatrix} T_{k-1}' \\ 0 \end{pmatrix},$$

and $Z_k = (z_0 \quad \cdots \quad z_{k-1} \quad z_k)$ (the form of the last column of T_k follows from the definition of v_k as $Z_k^T g_k$). Once p_k is computed, it is swapped with g_k in the basis with no computation required, yielding

$$B_k' = \begin{pmatrix} p_0 & \cdots & p_{k-1} & p_k \end{pmatrix} \quad \text{and} \quad T_k' = \begin{pmatrix} \underline{T}_{k-1}' & q_k \end{pmatrix},$$

where $q_k = Z_k^T p_k$. The matrix Z_k is unchanged. While building the basis, the last component of q_k is nonzero and the swap can always be done (see Leonard [20, pp. 94–99]). After the line search, g_{k+1} is accepted and the orthogonalization procedure yields

$$B_k'' = \begin{pmatrix} B_k' & g_{k+1} \end{pmatrix} = \begin{pmatrix} p_0 & \cdots & p_{k-1} & p_k & g_{k+1} \end{pmatrix}, \qquad T_k'' = \begin{pmatrix} T_k' & u_k \\ 0 & \rho_{k+1} \end{pmatrix},$$

and $Z_k' = (z_0 \quad \cdots \quad z_{k-1} \quad z_k \quad z_{k+1})$. These matrices are then passed to iteration $k+1$ as B_{k+1}, T_{k+1}, and Z_{k+1}, respectively.

3.2. Discarding the oldest basis vector. Now suppose that $k = m-1$. Given the assumption that every gradient is accepted, there are $m+1$ vectors in the basis at the end of this iteration. At this point, p_0, the oldest search direction, must be discarded before starting iteration $k+1$. This gives the new basis

$$B_{k+1} = \begin{pmatrix} p_1 & \cdots & p_k & g_{k+1} \end{pmatrix}.$$

(On the other hand, if at least one gradient is rejected, then no vector is discarded at iteration m. In this case, B_{k+1} contains r_{k+1} ($r_{k+1} \leq m$) linearly independent vectors consisting of at most one gradient (the vector g_{k+1}) and a linearly independent set of search directions. If g_{k+1} is rejected, B_{k+1} will consist of r_{k+1} linearly independent search directions.) Discarding a vector from the basis will decrease the rank by one. Hence, a symbol r_k' is needed for the intermediate rank determined by the orthogonalization procedure ***orthog***. The final rank r_{k+1} is then either $r_k' - 1$ or r_k' depending upon whether or not a basis vector is discarded.

When the oldest direction p_0 is discarded, the removal of its associated column from the basis must be reflected in all factorizations associated with B_k''. To simplify the description, the subscript k is suppressed, and a bar is used to denote quantities with subscript $k+1$.

The relationship between the old and new bases B'' and $\bar{B}$ ($= B_{k+1}$) is given by $B'' = \begin{pmatrix} p_0 & \bar{B} \end{pmatrix}$, where $\bar{B}$ is $n \times m$. Associated with $\bar{B}$, we require $\bar{Z}$ and $\bar{T}$ such that $\bar{B} = \bar{Z}\bar{T}$. Moreover, the change from Z' to $\bar{Z}$ induces a corresponding change to the Cholesky factor. If R''' denotes the factor defined by the ***reinitialize*** procedure, then we require the factor $\bar{R}$ such that $\bar{R}^T\bar{R} = \bar{Z}^T H'' \bar{Z}$, where H'' is defined as in (2.7) but in terms of $\bar{\sigma}$, Z', and R'''. The matrix H'' is the H_{k+1} defined in section 2.2.

Daniel et al. [6] give the following method for updating Z' and T''. Given any orthogonal S, the orthogonal factorization of B'' may be written as $B'' = Z'T'' = Z'S^T S T'' = Z_S T_S$ (say). The matrix S is constructed so that T_S has the partitioned form

$$T_S = \begin{pmatrix} t & \bar{T} \\ \tau & 0 \end{pmatrix},$$

where $\bar{T}$ is the desired $m \times m$ upper-triangular matrix, t is an m-vector, and τ is a scalar. In particular, $S = P_{m,m+1} P_{m-1,m} \cdots P_{12}$, where $P_{i,i+1}$ is an $(m+1) \times (m+1)$

plane rotation in the $(i, i+1)$ plane that annihilates the $(i+1, i+1)$ element of $P_{i-1,i} \cdots P_{12} T''$.

The matrix $\bar{Z}$ consists of the first m columns of Z_S, i.e., $Z_S = \begin{pmatrix} \bar{Z} & z \end{pmatrix}$, where z is an n-vector. From the definition of B'', we have

$$B'' = \begin{pmatrix} p_0 & \bar{B} \end{pmatrix} = Z_S T_S = \begin{pmatrix} \bar{Z}t + \tau z & \bar{Z}\bar{T} \end{pmatrix},$$

and it follows that $\bar{B} = \bar{Z}\bar{T}$ is the required orthogonal factorization.

Next, we propose how to update R''' when the first column of B'' is discarded. The old and new orthogonal bases are related by the identity $\bar{Z} = Z_S E_m$, where E_m comprises the first m columns of the identity matrix of order $m+1$. From the definitions of $\bar{Z}$ and H'', the new reduced Hessian is given by

$$\bar{Z}^T H'' \bar{Z} = E_m^T Z_S^T H'' Z_S E_m = E_m^T S Z'^T H'' Z' S^T E_m = E_m^T S R'''^T R''' S^T E_m.$$

In general, $R''' S^T$ is not upper triangular, but it may be restored to upper-triangular form by a second sweep of plane rotations $\widetilde{S}$. The $(m+1) \times (m+1)$ matrix $\widetilde{S}$ is orthogonal and is chosen so that $\widetilde{S} R''' S^T$ is upper triangular. If $R_S = \widetilde{S} R''' S^T$ denotes the resulting product, then $\bar{Z}^T H'' \bar{Z} = E_m^T R_S^T R_S E_m$, which implies that the leading $m \times m$ block of R_S is the required factor $\bar{R}$. Note that at this point, we can define $\bar{H}$ $(= H_{k+1})$ in terms of $\bar{R}$, $\bar{Z}$, and $\bar{\sigma}$.

The matrix $\widetilde{S}$ is the product $\widetilde{P}_{m,m+1} \cdots \widetilde{P}_{23} \widetilde{P}_{12}$, where $\widetilde{P}_{i,i+1}$ is an $(m+1) \times (m+1)$ plane rotation in the $(i, i+1)$ plane that annihilates the $(i, i+1)$ element of $\widetilde{P}_{i-1,i} \cdots \widetilde{P}_{12} R''' P_{12}^T \cdots P_{i,i+1}^T$. In practice, the two sweeps $\widetilde{S}$ and S are interlaced so that only $\mathcal{O}(m^2)$ operations are required.

It remains to show how u' $(= Z'^T \bar{g})$ is updated, thereby avoiding the mn operations necessary to compute the new reduced gradient u'' $(= \bar{Z}^T \bar{g})$ from scratch. The identity $u'' = \bar{Z}^T \bar{g} = (Z_S E_m)^T \bar{g} = E_m^T S (Z'^T \bar{g}) = E_m^T S u'$ implies that u'' comprises the first m components of Su'.

3.3. Comparison of the bases. We now revert to using subscripts to denote iteration indices. Under the assumption that every gradient is accepted, the gradient and search-direction bases at the start of iteration $m-1$ are given by $G_{m-1} = \begin{pmatrix} g_0 & g_1 & \cdots & g_{m-1} \end{pmatrix}$ and $P_{m-1} = \begin{pmatrix} p_0 & p_1 & \cdots & p_{m-2} & g_{m-1} \end{pmatrix}$. Theorem 2.2 implies that $\text{range}(G_{m-1}) = \text{range}(P_{m-1})$, and we can expect that the value of p_{m-1} is independent of the choice of basis. However, the following argument shows that this is not necessarily true for p_m, and hence the gradient and search-direction bases are not necessarily the same in the limited-memory context.

At the end of iteration $m-1$, both bases will have $m+1$ vectors. In the limited-memory context, the oldest basis vector must be discarded, giving bases $G_m = \begin{pmatrix} g_1 & g_2 & \cdots & g_m \end{pmatrix}$ and $P_m = \begin{pmatrix} p_1 & p_2 & \cdots & p_{m-1} & g_m \end{pmatrix}$. These bases do not include g_0 and p_0 (which is parallel to g_0), respectively. Note that p_{m-1} is not necessarily in $\text{range}(G_m)$ because p_{m-1} may have a nonzero component of g_0, which has been discarded from the gradient basis. Since $p_{m-1} \in \text{range}(P_m)$ by construction, it follows that $\text{range}(G_m) \neq \text{range}(P_m)$. We will discuss a specific implication of this phenomenon in section 3.7.

3.4. An implicit representation of Z. When a basis vector is discarded, the application of the plane rotations to the right of Z_k' requires approximately $4mn$ operations. Although it is possible to reduce this to $3mn$ operations (see Daniel et al. [6]), the update to Z_k' dominates the time to perform an iteration and reduces the

efficiency compared to other methods. For example, the *total* number of operations for an iteration of the limited-memory method of Nocedal [26] is approximately $4mn$.

Our limited-memory reduced-Hessian method is substantially faster if, as proposed by Siegel, the basis matrix B_k is stored instead of Z_k. In this case, products involving Z_k are computed as needed using T_k and B_k, and the number of operations required to drop a column from the basis is reduced to $\mathcal{O}(r_k^2)$.

With an implicit definition of Z_k, the orthogonalization procedure becomes a method for updating the orthogonal factorization of B_k *without storing* Z_k. Given B_k and a new gradient g_{k+1}, the first step is to compute $u_k = Z_k^T g_{k+1}$ from the equations $T_k'^T u_k = B_k'^T g_{k+1}$. Once u_k is known, ρ_{k+1} can be computed from the identity $\rho_{k+1}^2 = \|g_{k+1}\|^2 - \|u_k\|^2$. The updated triangular factor T_k'' is defined by augmenting T_k' by a column formed from $u_k = Z_k^T g_{k+1}$ and ρ_{k+1} (see (2.5)). Note that the column z_{k+1} is not needed. The implicit form of Z_k reduces the cost of the orthogonalization procedure by half to approximately nr_k operations.

3.5. The limited-memory algorithm. We have described a reduced-Hessian limited-memory algorithm that needs three procedures in addition to those needed by Algorithm RHR: a *swap*, which replaces an accepted gradient g_k with p_k in the definition of B_k, giving a basis defined by B_k' and T_k'; a new *orthogonalize*, which orthogonalizes g_{k+1} with respect to Z_k, giving a new orthonormal basis defined by B_k'' and T_k''; and a *discard*, which drops the oldest search direction from the basis. As in Algorithm RHR, statements of the form $(B_k', T_k') = \boldsymbol{swap}(B_k, T_k)$ indicate computed quantities and their dependencies associated with a given procedure. Similarly, the results of the implicit orthogonalization and discard procedures are denoted by $(B_k'', T_k'', u_k, \rho_{k+1}, r_k') = \boldsymbol{iorthog}(B_k', T_k', g_{k+1}, r_k)$ and $(B_{k+1}, T_{k+1}, R_{k+1}, u_k'') = \boldsymbol{drop}(B_k'', T_k'', R_k''', u_k')$.

ALGORITHM 3.1. (L-RHR) LIMITED-MEMORY VERSION OF ALGORITHM RHR.
Choose x_0, σ_0 ($\sigma_0 > 0$), and m ($m \geq 2$);
$k = 0$; $r_0 = 1$; $g_0 = \nabla f(x_0)$;
$B_0 = \begin{pmatrix} g_0 \end{pmatrix}$; $T_0 = \begin{pmatrix} \|g_0\| \end{pmatrix}$; $v_0 = \|g_0\|$; $R_0 = \begin{pmatrix} \sigma_0^{1/2} \end{pmatrix}$;
while not converged **do**
  Solve $R_k^T d_k = -v_k$; $R_k q_k = d_k$;
  Solve $T_k w = q_k$; $p_k = B_k w$;
  if g_k was accepted **then** $(B_k', T_k') = \boldsymbol{swap}(B_k, T_k)$;
  Find α_k satisfying the Wolfe conditions (2.2);
  $x_{k+1} = x_k + \alpha_k p_k$; $g_{k+1} = \nabla f(x_k + \alpha_k p_k)$;
  $(B_k'', T_k'', u_k, \rho_{k+1}, r_k') = \boldsymbol{iorthog}(B_k', T_k', g_{k+1}, r_k)$;
  $(R_k', u_k', v_k', q_k') = \boldsymbol{expand}(R_k, u_k, v_k, q_k, \rho_{k+1}, \sigma_k)$;
  $s_k = \alpha_k q_k'$; $y_k = u_k' - v_k'$; $R_k'' = \boldsymbol{update}(R_k', s_k, y_k)$;
  Compute σ_{k+1}; $R_k''' = \boldsymbol{reinitialize}(R_k'', \sigma_{k+1})$;
  if r_k' equals $m+1$ **then**
    $(B_{k+1}, T_{k+1}, R_{k+1}, u_k'') = \boldsymbol{drop}(B_k'', T_k'', R_k''', u_k')$; $r_{k+1} = m$;
  else
    $R_{k+1} = R_k'''$; $B_{k+1} = B_k''$; $T_{k+1} = T_k''$; $u_k'' = u_k'$; $r_{k+1} = r_k'$;
  end if
  $v_{k+1} = u_k''$;
  $k = k + 1$;
end do

Iteration k of Algorithm L-RHR requires $2nr_k+2n+\mathcal{O}(r_k^2)$ operations. (This total includes the work required for the ***swap***, ***iorthog***, ***update***, and ***drop*** procedures but does not include any overhead incurred during the line search.)

3.6. Keeping the reduced Hessian vs. the reduced inverse Hessian. If Siegel's algorithm and an un-reinitialized version of L-RHR are applied with the same line search, then the same search directions and subspace bases are generated for the first m iterations. During these iterations the full $n\times n$ Hessian of L-RHR (see (2.7)) is the inverse of the full inverse Hessian of Siegel's method. However, once a basis vector is discarded, the methods generate different search directions. In both algorithms, the updating procedures associated with a discard relegate curvature information associated with the oldest basis vector to the last row and column of their respective reduced matrices. The off-diagonal entries of this row and column are replaced by zero, and the diagonal element is set to either σ or $1/\sigma$ depending on the method. At this point the two basis matrices generate the same subspace, but because the leading $m\times m$ principal submatrices of a symmetric matrix and its inverse are not generally the inverse of each other, the full Hessian of L-RHR is no longer the inverse of the full inverse Hessian of Siegel's method. At the next iteration, the methods generate different search directions, and subsequent bases and reduced matrices are no longer related.

3.7. Finite termination on quadratics. Next we briefly discuss the properties of Algorithm L-RHR when it is applied with an exact line search to a strictly convex quadratic function.

THEOREM 3.1. *Consider Algorithm* L-RHR *implemented with an exact line search and $\sigma_0 = 1$. If this algorithm is applied to a strictly convex quadratic function, then R_k, B_k, and T_k $(k \geq 1)$ satisfy*

$$R_k = \begin{pmatrix} a_l/h_l & b_l & 0 & \cdots & 0 \\ & a_{l+1} & b_{l+1} & \ddots & \vdots \\ & & \ddots & \ddots & 0 \\ & & & a_{k-1} & b_{k-1} \\ & & & & \sigma_k^{1/2} \end{pmatrix}, \quad B_k = \begin{pmatrix} p_l & p_{l+1} & \cdots & p_{k-1} & g_k \end{pmatrix},$$

and

$$T_k = \begin{pmatrix} h_l d_l & h_l t_{l,l+1} & h_l t_{l,l+2} & \cdots & h_l t_{l,k-1} & 0 \\ & d_{l+1} & t_{l+1,l+2} & \cdots & t_{l+1,k-1} & 0 \\ & & d_{l+2} & & \vdots & \vdots \\ & & & \ddots & t_{k-2,k-1} & 0 \\ & & & & d_{k-1} & 0 \\ & & & & & \|g_k\| \end{pmatrix},$$

where $l = \max\{0, k-m+1\}$ and the scalars a_j, b_j, t_{ij}, d_j, and h_j are given by

$$a_j = \frac{\|g_j\|}{(y_j^T s_j)^{1/2}}, \quad b_j = -\frac{\|g_{j+1}\|}{(y_j^T s_j)^{1/2}}, \quad t_{ij} = -\frac{\|g_j\|^2}{\sigma_j \|g_i\|}, \quad d_j = -\frac{\|g_j\|}{\sigma_j}, \quad h_j = \delta_j \frac{\|p_j\|\sigma_j}{\|g_j\|}$$

with $\delta_j = 1$ if $j = 0$, and $\delta_j = -1$ otherwise. Furthermore, the search directions are given by

$$p_0 = -g_0; \qquad p_k = -\frac{1}{\sigma_k} g_k + \beta_{k-1} p_{k-1}, \quad \beta_{k-1} = \frac{\sigma_{k-1}}{\sigma_k} \frac{\|g_k\|^2}{\|g_{k-1}\|^2}, \quad k \geq 1.$$

Proof. See Leonard [20]. ∎

COROLLARY 3.2. *If Algorithm* L-RHR *is used to minimize a strictly convex quadratic under the conditions of Theorem* 3.1*, then the method converges to the minimizer in at most n iterations.*

Proof. We show by induction that the search directions are parallel to the conjugate-gradient directions $\{d_k\}$. Specifically, $\sigma_k p_k = d_k$ for all k. This is true for $k = 0$ since $1 \cdot p_0 = -g_0 = d_0$. Assume that $\sigma_{k-1} p_{k-1} = d_{k-1}$. Using Theorem 3.1 and the inductive hypothesis, we find

$$\sigma_k p_k = -g_k + \sigma_{k-1} \frac{\|g_k\|^2}{\|g_{k-1}\|^2} p_{k-1} = -g_k + \frac{\|g_k\|^2}{\|g_{k-1}\|^2} d_{k-1} = d_k,$$

which completes the induction. The result now follows from the quadratic termination property of the conjugate-gradient method. ∎

We remark that the specific form of L-RHR discussed in Theorem 3.1 defines a "rescaled" form of the classical Fletcher–Reeves conjugate-gradient method [12].

Let p_m^G and p_m^P denote the search directions defined during iteration m of the gradient- and search-direction variants of the limited-memory algorithm. Observe that p_{m-1} is parallel to d_{m-1}, regardless of which basis is used. However, since $g_m \in \text{range}(G_m)$, but possibly $p_{m-1} \notin \text{range}(G_m)$, the vector p_m^G may not be parallel to d_m and the gradient-basis variant does not have quadratic termination.

4. Implementation details. In this section, we describe some details associated with a particular implementation of Algorithm L-RHR. We outline a method for improving the orthonormal basis and discuss a practical criterion for accepting a gradient. We also provide information about the line search, the BFGS update, and restarts.

4.1. Reorthogonalization. For general applications the implicit QR version of Z_k is recommended, with default memory size $m = 5$. For larger values of m (e.g., $m \geq 15$), it is often beneficial to use *reorthogonalization* in combination with the *explicit* QR. L-RHR employs the following reorthogonalization scheme proposed by Daniel et al. [6]. Let u_k and w_k denote the computed values of $Z_k^T g_{k+1}$ and $g_{k+1} - Z_k u_k$, respectively. The vectors u_k and w_k may be improved using one or more steps of the iterative refinement scheme:

$$\Delta u_k = Z_k^T w_k, \qquad u_k \leftarrow u_k + \Delta u_k;$$

and

$$\Delta w_k = -Z_k \Delta u_k, \qquad w_k \leftarrow w_k + \Delta w_k.$$

L-RHR uses criteria suggested by Daniel et al. [6] for invoking and terminating the reorthogonalization. Each step of reorthogonalization adds approximately $2nr_k$ operations to the cost of an iteration. Moreover, the use of an explicit Z_k requires an additional nr_k operations for the calculation of z_{k+1} and an extra $3nr_k$ operations when a basis vector is discarded (since the plane rotations associated with a discard must be applied to Z_k as well as T_k).

4.2. The criterion for gradient acceptance. In exact arithmetic, a gradient is accepted for the basis if $\rho_{k+1} > 0$, where ρ_{k+1} is the norm of $(I - Z_k Z_k^T)g_{k+1}$. This condition ensures that the basis vectors are linearly independent, and hence that T_k'' is nonsingular. When ρ_{k+1} is computed in finite-precision, gradients with small (but nonzero) ρ_{k+1} must be rejected to prevent T_{k+1} and B_{k+1} from being too ill-conditioned. In practice, an accepted gradient must satisfy $\rho_{k+1} \geq \epsilon\|g_{k+1}\|$, where ϵ is a preassigned positive constant. In the results of section 5, ϵ was set to 10^{-4}. Rounding error in the calculation of ρ_{k+1} is exacerbated by the use of an implicit form for Z_k—for example, it is necessary to reject g_{k+1} if the computed value of $\rho_{k+1}^2 = \|g_{k+1}\|^2 - \|u_k\|^2$ is negative. However, a negative computed value of ρ_{k+1}^2 rarely occurred in our experiments, and when it did, it did not prevent the method from terminating successfully (see section 5 for the criterion used). For example, a negative value was computed 268 times during the 69747 iterations in the runs of Table 5.6 below. Moreover, of the 10 problems in which a negative value occurred, all were solved successfully.

4.3. The line search, the BFGS update, and restarts. The line search is a slightly modified version of the one used in the package NPSOL [16]. It is designed to ensure that α_k satisfies the so-called strong Wolfe conditions,

$$f(x_k + \alpha_k p_k) \leq f(x_k) + \mu\alpha_k g_k^T p_k \quad \text{and} \quad |g_{k+1}^T p_k| \leq \eta |g_k^T p_k|, \tag{4.1}$$

where the constants μ and η are chosen so that $0 \leq \mu < \eta < 1$ and $\mu < \frac{1}{2}$ (see Gill et al. [16] or Fletcher [9, pp. 26–30]). The step-length parameters are $\mu = 10^{-4}$ and $\eta = 0.9$. The line search is based on using a safeguarded polynomial interpolation to find an approximate minimizer of the univariate function

$$\phi_k(\alpha) = f(x_k + \alpha p_k) - f(x_k) - \mu\alpha g_k^T p_k$$

(see Moré and Sorensen [23]). The step α_k is the first member of a minimizing sequence $\{\alpha_k^i\}$ that satisfies the Wolfe conditions. The sequence is usually started with $\alpha_k^0 = 1$ (see below).

If α_k satisfies the strong Wolfe conditions, it follows that $y_k^T s_k \geq -(1-\eta)g_k^T s_k > 0$ and the BFGS update can be applied without difficulty. On very difficult problems, however, the combination of a poor search direction and rounding error in f may prevent the line search from satisfying the line search conditions within 20 function evaluations. In this case, the search terminates with the step corresponding to the best value of f found so far. If this α_k defines a strict decrease in f, the minimization continues. In this case, the BFGS update is skipped unless $y_k^T s_k \geq \epsilon_M |g_k^T s_k|$, where ϵ_M is the machine precision. If a strict decrease is not obtained after 20 function evaluations, the algorithm is restarted with $T_k = \big(\,\|g_k\|\,\big)$, $v_k = \|g_k\|$, and $R_k = \big(\,\sigma_k^{1/2}\,\big)$. To prevent the method from degenerating into steepest descent, no more restarts are allowed until the reduced Hessian has built up to its full size of m rows and columns. In practice, a restart is rarely invoked. For example, in the experiments of Table 5.6, L-RHR used only one restart (on problem *freuroth*, which was not solved successfully). For comparison, L-BFGS-B used two restarts (on problem *bdqrtic*, again without success).

If p_k is a poorly scaled version of the steepest-descent direction, the step to a minimizer of $\phi_k(\alpha)$ may be very small relative to one, and a large number of function evaluations may be needed to find an acceptable step length. To prevent this inefficiency, the initial step for the first line search and each line search immediately

following a restart is limited so that $\alpha_k^0 \leq \min\{\Delta/\|p_k\|, 1\}$, where Δ is a preassigned constant ($\Delta = 2$ in the experiments described in the next section). This procedure ensures that the initial change in x does not exceed Δ.

5. Numerical results. In this section, we give numerical results for most of the large unconstrained problems in the CUTE[2] collection (see Bongartz et al. [1]). After some discussion of the test problems, we compare L-RHR with and without reinitialization. Next, we illustrate the differences between L-RHR and Siegel's Algorithm 6 [33], which we refer to as ALG6. This is followed by results that compare L-RHR with L-BFGS and L-BFGS-B, which are two alternative implementations of the limited-memory BFGS method. L-BFGS is based on an algorithm that maintains an implicit approximate inverse Hessian as a sequence of update pairs. L-BFGS-B employs an algorithm that updates an approximate Hessian in factored form $\theta I - WMW^T$, where θ is a scalar and WMW^T is a matrix of low rank. L-BFGS-B is intended for problems with upper and lower bounds on the variables but is also recommended over L-BFGS-B for unconstrained problems (see Zhu et al. [34]).

Throughout, we use m_{LB} to denote the number of update pairs to be kept in memory by L-BFGS and L-BFGS-B. This should not be confused with m, the number of vectors stored by L-RHR.

5.1. Test problem selection. The test set was constructed using the CUTE interactive `select` tool, which allows the identification of groups of problems with certain features. In our case, the `select` tool was first used to locate the twice-continuously differentiable unconstrained problems for which the number of variables in the data file can be varied. Of these problems, the number of variables was set to a value in the range $100 \leq n \leq 1500$ according to criteria that we discuss below. The input for the `select` tool was as follows:

```
Objective function type         : *
Constraints type                : U  (No constraints)
Regularity                      : R  (twice-cont. differentiable)
Degree of available derivatives : *
Problem interest                : *
Explicit internal variables     : *
Number of variables             : v  (variable dimension)
Number of constraints           : 0.
```

A total of 87 problems was obtained from this selection. Six fixed-dimension problems were obtained by using the `select` tool with the number of variables set as follows:

```
Number of variables             : in [    50,   1000 ].
```

Additional criteria were used to determine the suitability of these 93 problems, as we now explain.

After using the `select` tool, it remained to determine a suitable value of n for the problems with variable dimension. The value $n = 1500$ was used for the twelve problems *dixmaana–dixmaanl*, as suggested by Zhu et al. [34]. Values $n \approx 1000$ were used for most of the remaining problems, but it was necessary to choose significantly smaller values of n in some cases. The problems *chnrosnb*, *errinros*, and *watson*

[2]The version of CUTE used was obtained September 7, 2001.

have limits on the size of n, and the mandated maximum values of 50, 50, and 31, respectively, were used in these cases. It was also necessary to limit n to be less than 1000 if a problem could not be decoded using the CUTE decoder `sifdec` (compiled with the option `tobig`). For any such problem, the values $n = 300$ and $n = 100$ were tried successively until the decoding succeeded. Problems in this category were *arglina–arglinc*, *brownal*, *hilberta*, *hilbertb*, *mancino*, *penalty3*, and *sensors*. The value $n = 300$ was used for *arglina*, *brownal*, *hilberta*, and *hilbertb*. The value $n = 100$ was used for *mancino*, *penalty3*, and *sensors*. The problems *arglinb* ($n = 300$) and *arglinc* ($n = 100$) and *penalty3* ($n = 100$) were successfully decoded but were removed from the set for reasons described below.

A value of n such that $n < 1000$ was also used if both L-BFGS-B and L-RHR failed to meet the termination criterion with m, $m_{LB} = 5$, 15, 30, and 45. (The termination criterion will not be satisfied if there is a failure in the line search or 40,000 iterations are completed.) In this case, 300 and 100 were tried successively to determine an acceptable value for n. The problems *arglinb*, *arglinc*, *curly10*, *curly20*, *curly30*, *fletchbv*, *hydc20ls* (fixed $n = 99$), *indef*, *nonmsqrt*, *penalty3*, *sbrybnd*, *scosine*, *scurly10*, *scurly20*, and *scurly30* were removed from the test set since, even with $n = 100$, neither method could meet the termination criterion. The value $n = 100$ was used for *penalty2* since neither L-RHR nor L-BFGS-B could achieve the termination criterion with $n = 1000$ or $n = 300$.

These selection criteria had the effect of removing 15 problems from the list generated by the select tool. This left 78 problems suitable for testing. For completeness, we list the problems not already mentioned, with their associated values of n. There were 44 variable-dimension problems with $n = 1000$: *arwhead*, *bdqrtic*, *broydn7d*, *brybnd*, *chainwoo*, *cosine*, *cragglvy*, *dixon3dq*, *dqdrtic*, *dqrtic*, *edensch*, *engval1*, *extrosnb*, *fletcbv2*, *fletcbv3*, *fletchcr*, *freuroth*, *genhumps*, *genrose*, *liarwhd*, *morebv*, *ncb20*, *ncb20b*, *noncvxu2*, *noncvxun*, *nondia*, *nondquar*, *penalty1*, *powellsg*, *power*, *quartc*, *schmvett*, *sinquad*, *sparsine*, *sparsqur*, *spmsrtls*, *srosenbr*, *testquad*, *tointgss*, *tquartic*, *tridia*, *vardim*, *vareigvl*, and *woods*. There were four problems with $n = 1024$: *fminsrf2*, *fminsurf*, *msqrtals*, and *msqrtbls*. The remaining three variable-dimension problems were *eigenals* ($n = 1056$), *eigenbls* ($n = 1056$), and *eigencls* ($n = 1122$). Finally, the names and numbers of variables of the five fixed-dimension problems included in the test set were *deconvu* ($n = 61$), *eg2* ($n = 1000$), *tointgor* ($n = 50$), *tointpsp* ($n = 50$), and *tointqor* ($n = 50$).

All runs were made on a Sun UltraSPARC-IIi (single cpu at 333MHz) with 256MB of RAM. The algorithms L-RHR, L-BFGS-B, and L-BFGS are coded in Fortran and were compiled using g77. ALG6 is coded in C and was compiled using gcc. Full compiler optimization was used in all cases. The caption of each table specifies the amount of limited memory used and indicates whether or not reinitialization and/or reorthogonalization was used. All methods were terminated when $\|g_k\|_\infty < 10^{-5}$, as proposed by Zhu et al. [34].

5.2. L-RHR with and without reinitialization. Table 5.1 gives the results of running L-RHR both with and without reinitialization. Without reinitialization, the parameter σ was fixed at $y_0^T s_0/\|s_0\|^2$ (see (2.6)).[3] This is the scheme proposed by Siegel [33]. With reinitialization, $\sigma_0 = 1$ and $\sigma_k = y_k^T y_k / y_k^T s_k$ ($k \geq 1$), which are the reciprocals of the parameters used by Liu and Nocedal [21]. Of the 78 problems

[3]The steepest-descent direction is used for the first iteration. After the first step, σ is set to $y_0^T s_0/\|s_0\|^2$ and R is defined accordingly.

attempted, L-RHR with reinitialization solved 74 problems satisfactorily and reduced the gradient to within two orders of magnitude of the 10^{-5} target value on three others (*bdqrtic*, *freuroth*, and *noncvxun*). The algorithm was unable to reduce $\|g_k\|_\infty$ below 1.5×10^{-2} for *fletcbv3*. Without reinitialization, L-RHR was able to solve only 70 problems and required considerably more function and gradient evaluations on almost every problem attempted. (The additional four unsolved problems were *chainwoo*, *cragglvy*, *edensch*, and *penalty2*.) Table 5.1 gives the total number of iterations, function evaluations and cpu seconds for L-RHR with and without reinitialization on the 70 problems that could be solved by both versions. These results indicate that reinitialization provides substantial practical benefits and indicates an advantage of L-RHR compared to Siegel's method, which does not include reinitialization. A direct comparison between L-RHR and Siegel's method is given in the next section.

TABLE 5.1
L-RHR[a] *with and without reinitialization on* 70 *CUTE problems.*

L-RHR	Itns	Fncs	Cpu	Fail
with reinitialization	65115	66914	1567	4
without reinitialization	83611	107253	1884	8

[a] $m = 5$; without reorthogonalization.

5.3. L-RHR compared with ALG6. Next we compare L-RHR with Siegel's ALG6. In the first set of runs, ALG6 uses the recommended value of $\epsilon = 10^{-3}$ for the gradient acceptance parameter (see [32, p. 8]). The line search in ALG6 is a slightly modified version of the one in Powell's Fortran package TOLMIN [28]. It attempts to satisfy the Wolfe conditions (see (2.2)) with $\mu = 10^{-2}$ and $\eta = 0.9$ but allows $f(x_{k+1}) \geq f(x_k)$ to within a small tolerance. ALG6 is not optimized for cpu time, and it may be possible to improve the performance by making appropriate changes to the code. However, the *relative* differences in cpu times are unlikely to be altered by recoding because many of the run times are dominated by the cumulative cost of the function evaluations (see section 5.5).

Table 5.2 gives the results of comparing ALG6 with a version of L-RHR implemented *without* reinitialization. Algorithm L-RHR succeeded on 70 of the 78 test problems and reduced the gradient norm to at most 10^{-3} on seven others: *bdqrtic*, *chainwoo*, *cragglvy*, *edensch*, *freuroth*, *noncvxun*, and *penalty2*. On the other unsuccessful case, *fletcbv3*, the final gradient norm was 1.1×10^{-1}. ALG6 succeeded on 74 out of the 78 problems. On *arwhead*, *bdqrtic*, and *noncvxun*, ALG6 was able to reduce the gradient norm to at most 10^{-3}. On *fletcbv3*, ALG6 reduced the gradient norm to 4.7×10^{-2}. Table 5.2 summarizes the results for the 69 problems that both methods were able to solve successfully. If the L-RHR line search is made to conform to ALG6 by allowing $f(x_{k+1}) \geq f(x_k)$ to within a prescribed tolerance, then L-RHR is able to solve three more problems, *chainwoo*, *edensch*, and *penalty2*.

TABLE 5.2
L-RHR[a] *compared with* ALG6[b] *on* 69 *CUTE problems.*

Method	Itns	Fncs	Cpu	Fail
L-RHR	83601	107237	1884	8
ALG6	101959	194091	5744	4

[a] $m = 5$; with no reinitialization and no reorthogonalization; $\epsilon = 10^{-4}$.
[b] $m = 5$; with Siegel's version of the TOLMIN line search; $\epsilon = 10^{-3}$.

Since the L-RHR and ALG6 directions have similar definitions when L-RHR does not use reinitialization, it might seem surprising that L-RHR requires significantly fewer function evaluations than ALG6. This phenomenon can be partly explained by differences in the line search and the different choice of ϵ. To illustrate these effects, Table 5.3 gives a comparison between L-RHR and ALG6 when both algorithms are implemented with the NPSOL line search and $\epsilon = 10^{-4}$. Note that the number of function evaluations for ALG6 decreases dramatically, though the stricter requirement that $f(x_{k+1}) < f(x_k)$ results in a few more failures.

TABLE 5.3
L-RHR[a] *compared with* ALG6[b] *on* 69 *CUTE problems.*

Method	Itns	Fncs	Cpu	Fail
L-RHR	82363	105997	1884	8
ALG6	86970	138512	3620	8

[a] $m = 5$; with no reinitialization and no reorthogonalization; $\epsilon = 10^{-4}$.
[b] $m = 5$; with the line search from NPSOL; $\epsilon = 10^{-4}$.

The results are closer, but Table 5.3 illustrates that the methods are still generating different iterates. This is because, in the limited-memory context, an algorithm based on updating a reduced Hessian is fundamentally different from an algorithm based on updating a reduced *inverse* Hessian. The directions generated by ALG6 and an un-reinitialized version of L-RHR are only the same until a basis vector is discarded. From this point, the Hessian of L-RHR is no longer related to the inverse Hessian of ALG6 (see section 3.6). For example, Table 5.4 illustrates that if L-RHR and ALG6 are applied to problem *msqrtals* with $m = 30$, the function values and gradient norms are in close agreement at iteration 30. At the next iteration the first discard is made and most of the agreement is lost. By iteration 50, only 1 significant digit of agreement remains. The total numbers of iterations required are 2556 and 3133 for L-RHR and ALG6, respectively. It follows that L-RHR's significant advantage in the "Fncs" column of Table 5.3 results from the use of a reduced Hessian instead of a reduced inverse Hessian.

TABLE 5.4
L-RHR[a] *and* ALG6[b] *applied to msqrtals with* $m = 30$.

	L-RHR		ALG6	
k	f_k	$\|g_k\|_\infty$	f_k	$\|g_k\|_\infty$
30	0.387222166177969	0.583990052575414	0.387222166177971	0.583990052575427
31	0.341636067001799	0.569093782828364	0.341694835004789	0.569019238515868
50	0.096802831009406	0.126889757489794	0.097640019163964	0.140754040816826

[a] $m = 5$; with no reinitialization and no reorthogonalization; $\epsilon = 10^{-4}$.
[b] $m = 5$; with the line search from NPSOL; $\epsilon = 10^{-4}$.

We provide one final comparison in which L-RHR uses reinitialization. In this case, both L-RHR and ALG6 succeed on 74 problems, and there are 73 problems that are solved by both methods. Table 5.5 shows the overall results for these 73 problems. Note that, overall, the use of reinitialization by L-RHR results in significantly fewer function evaluations compared to ALG6.

TABLE 5.5
L-RHR[a] *compared with* ALG6[b] *on* 73 *CUTE problems.*

Method	Itns	Fncs	Cpu	Fail
L-RHR	69737	71782	1592	4
ALG6	103217	195405	5752	4

[a] $m = 5$; with reinitialization; without reorthogonalization.
[b] $m = 5$; with Siegel's implementation of the TOLMIN line search.

5.4. L-RHR compared with L-BFGS-B. L-RHR and Seigel's method are related to the limited-memory BFGS method of Byrd et al. [4] because all three methods consolidate the quasi-Newton updates into dense matrices. By contrast, L-BFGS keeps an implicit inverse Hessian by storing a fixed number of vector pairs (γ_k, δ_k) (see (2.1) for the definitions of γ_k and δ_k). Products of the inverse Hessian with a vector are then formed without the need to keep an explicit H_k (see Nocedal [26]). Consolidation of the updates is crucial for efficiency if a limited-memory method is to be extended to handle upper and lower bounds on the variables. In this section we compare L-RHR with Version 2.1 of the code L-BFGS-B (see Zhu et al. [34]), which is an implementation of the method of Byrd, Lu, Nocedal and Zhu. L-BFGS-B applies the strong Wolfe conditions (4.1) using the line search of Moré and Thuente [24] with line search parameters $\mu = 10^{-4}$ and $\eta = 0.9$. The memory for L-BFGS-B was limited to $m_{LB} = 5$ pairs of vectors, which is twice the storage used by L-RHR.

Table 5.6 summarizes the performance of L-RHR and L-BFGS-B on the 74 CUTE problems on which both methods succeed. On these 74 problems, L-RHR requires fewer function evaluations than L-BFGS-B on 27 problems and more function evaluations on 43 problems. L-RHR requires less cpu time than L-BFGS-B on 55 problems and more cpu time than L-BFGS-B on 16 problems. L-BFGS-B was able to solve one more problem than L-RHR, namely, *fletcbv3* (L-RHR reduced the gradient norm to 1.5×10^{-2} in this case). Neither method was able to satisfy the termination criterion on *bdqrtic*, *freuroth*, and *noncvxun*. In these cases, the final gradient norms for L-RHR (L-BFGS-B) were 2.90×10^{-4} (6.08×10^{-4}), 3.40×10^{-4} (1.60×10^{-5}), and 1.00×10^{-3} (1.66×10^{-3}), respectively. Overall, L-RHR requires a comparable number of function evaluations and has a significant advantage in terms of cpu time.

TABLE 5.6
L-RHR[a] *and* L-BFGS-B[b] *on* 74 *CUTE problems.*

Method	Itns	Fcns	Cpu	Fail
L-RHR	69747	71798	1592	4
L-BFGS-B	66717	72264	1916	3

[a] $m = 5$ n-vectors; with reinitialization; without reorthogonalization.
[b] $m_{LB} = 5$ pairs of n-vectors.

Although the values $m = 5$ and $m_{LB} = 5$ are recommended for L-RHR and L-BFGS-B, it is of interest to investigate the relative performance of the algorithms as the memory size is increased. For example, the overhead required to solve an unsymmetric $2m_{LB} \times 2m_{LB}$ system every iteration of L-BFGS-B might suggest that L-RHR would have a greater cpu time advantage as m and m_{LB} are increased. Table 5.7 gives the performance of L-RHR and L-BFGS-B with increasing memory-size parameters m and m_{LB}. When $m \geq 15$ it is recommended that L-RHR use reorthogonalization to maintain a good basis and improve robustness (see section 4.1). However, to illustrate

the difference when L-RHR with $m = 5$ does and does not use reorthogonalization, the results of Table 5.7 use reorthogonalization for all memory sizes. For $m = 5$, 15, 30, and 45, the numbers of failures for L-RHR without reorthogonalization are 4, 5, 7, and 11, respectively. With reorthogonalization, these numbers drop to 4, 4, 3, and 3, respectively. Table 5.7 provides the total number of reorthogonalizations required for each value of m. Although the amount of work per iteration is more than doubled with reorthogonalization, the cpu seconds required for $m = 5$ *decreases* from 1590 to 1580. In this case, the reductions in iterations and function evaluations compensates for the increased cost of computing the search direction.

All 78 problems were attempted, but the totals in each row include only the statistics for problems that could be solved by both methods. The cpu time advantage for L-RHR increases from 82% of the time required by L-BFGS-B to 59% as m and m_{LB} are increased from 5 to 45. However, L-BFGS-B gains an advantage in terms of the number of function evaluations. The interpretation of these results is complicated by the sharp increase in function evaluations when m is increased from 15 to 30. This increase occurs because both methods are able to solve problem *noncvxun* when $m = 30$. If *noncvxun* is removed from the problem set, a total of 69,522 (72,264), 67,293 (62,470), 64,808 (57,114), and 58,075 (54,391) evaluations are required by L-RHR (L-BFGS-B) for m (m_{LB}) with the values 5, 15, 30, and 45, respectively. These results indicate that for both methods, the number of function evaluations generally decreases as the limited-memory size is increased.

TABLE 5.7
L-RHR[a] *and* L-BFGS-B *with various memory sizes for* 78 *CUTE problems.*

	L-RHR					L-BFGS-B			
Mem	Itns	Fcns	Reors	Cpu	Fail	Itns	Fcns	Cpu	Fail
5	67573	69522	42018	1580	4	66717	72264	1916	3
15	65050	67293	57662	1951	4	57302	62470	2155	5
30	99122	101432	95445	2866	3	80669	86304	3744	2
45	78344	80550	76173	2782	3	65859	70962	4686	3

[a] With reinitialization and reorthogonalization.

5.5. L-RHR compared with L-BFGS. We conclude this section by providing a comparison of L-RHR with L-BFGS, which is a limited-memory BFGS method that maintains an implicit inverse approximate Hessian as a sequence of m_{LB} update pairs (see Nocedal [26] and Liu and Nocedal [21]). The recommended memory size is $m_{LB} = 5$. The search direction requires approximately $4nm_{LB}$ operations, which is roughly twice the work required by L-RHR. L-BFGS uses the Moré and Thuente line search with the same parameter settings as L-BFGS-B.

Table 5.8 summarizes the performance of L-RHR and L-BFGS on the 74 problems that both methods solved successfully. Of these 74 problems, L-RHR required fewer function evaluations on 27 problems and L-BFGS required fewer evaluations on 43 problems. L-RHR required less cpu time on 30 problems and more time on 41 problems. Overall, L-BFGS and L-RHR required comparable numbers of function evaluations and iterations. However, even though L-RHR requires roughly half as much work to compute the search direction, the overall cpu time was very close to that of L-BFGS. Further investigation using a performance profiler indicated that the cost of evaluating the objective for seven of the problems (*eigenals*, *eigenbls*, *eigencls*, *msqrtals*, *msqrtbls*, *ncb20*, and *ncb20b*) dominated the overall computation time. The

cpu time required by L-RHR to compute the search direction was less than 60% of that needed by L-BFGS. However, on these seven critical problems, the calculation of the search direction constitutes less than 10% of the solve time.

TABLE 5.8
L-RHR[a] *and* L-BFGS[b] *on* 74 *CUTE problems.*

Method	Itns	Fcns	Cpu	Fail
L-RHR	69747	71798	1592	4
L-BFGS	66445	71978	1670	4

[a] $m = 5$ n-vectors; with reinitialization; without reorthogonalization.
[b] $m_{LB} = 5$ pairs of n-vectors.

In order to compare L-RHR with L-BFGS as m and m_{LB} are increased, L-RHR uses reorthogonalization for improved robustness. As m and m_{LB} take on the values 5, 15, 30, and 45, the total function evaluations on the problems solved by both methods L-RHR (L-BFGS) are 69,522 (71,978), 67,293 (63,466), 101,432 (80,832), and 80,550 (88,717). As in the comparison with L-BFGS-B, L-RHR is competitive in terms of function evaluations, but in this case L-RHR requires more cpu time when $m \geq 15$. Although some of the functions dominate the overall cpu time when $m = 5$, their evaluation carries less weight with increasing m. This effect, when combined with the cost of reorthogonalization, is why L-RHR is slower than L-BFGS when $m = 45$, even though L-RHR requires significantly fewer function evaluations. We reemphasize that L-RHR requires approximately half the storage of L-BFGS.

6. Summary and conclusions. We have presented theoretical and practical details of a limited-memory reduced-Hessian method for large-scale smooth unconstrained optimization problems for which first derivatives are available. The method maintains the Cholesky factor of a reduced Hessian and requires roughly half the storage of conventional limited-memory methods.

The numerical results of section 5 confirm that L-RHR is efficient and reliable on a set of large test problems from the CUTE collection. Moreover, it is shown that Hessian reinitialization and selective reorthogonalization are vital components of an efficient and robust reduced-Hessian method.

When compared to the state-of-the-art code L-BFGS-B on our test set, L-RHR converges in less cpu time, while requiring comparable numbers of function evaluations. Compared to the code L-BFGS, L-RHR required comparable numbers of function evaluations and iterations when both algorithms were applied with their default memory sizes.

Acknowledgments. We thank Dirk Siegel for graciously providing a copy of his limited-memory code. We also appreciate many suggestions from the referees and Associate Editor Jorge Nocedal.

REFERENCES

[1] I. BONGARTZ, A. R. CONN, N. I. M. GOULD, AND P. L. TOINT, *CUTE: Constrained and unconstrained testing environment*, ACM Trans. Math. Software, 21 (1995), pp. 123–160.
[2] A. G. BUCKLEY, *A combined conjugate-gradient quasi-Newton minimization algorithm*, Math. Programming, 15 (1978), pp. 200–210.

[3] A. G. Buckley, *Extending the relationship between the conjugate-gradient and BFGS algorithms*, Math. Programming, 15 (1978), pp. 343–348.
[4] R. H. Byrd, P. Lu, J. Nocedal, and C. Zhu, *A limited memory algorithm for bound constrained optimization*, SIAM J. Sci. Comput., 16 (1995), pp. 1190–1208.
[5] R. H. Byrd, J. Nocedal, and R. B. Schnabel, *Representations of quasi-Newton matrices and their use in limited-memory methods*, Math. Programming, 63 (1994), pp. 129–156.
[6] J. W. Daniel, W. B. Gragg, L. Kaufman, and G. W. Stewart, *Reorthogonalization and stable algorithms for updating the Gram-Schmidt QR factorization*, Math. Comp., 30 (1976), pp. 772–795.
[7] J. E. Dennis, Jr. and R. B. Schnabel, *A new derivation of symmetric positive definite secant updates*, in Nonlinear Programming 4 (Madison, WI, 1980), Academic Press, New York, 1981, pp. 167–199.
[8] M. C. Fenelon, *Preconditioned Conjugate-Gradient-Type Methods for Large-Scale Unconstrained Optimization*, Ph.D. thesis, Department of Operations Research, Stanford University, Stanford, CA, 1981.
[9] R. Fletcher, *Practical Methods of Optimization*, 2nd ed., John Wiley and Sons, Chichester, New York, 1987.
[10] R. Fletcher, *An optimal positive definite update for sparse Hessian matrices*, SIAM J. Optim., 5 (1995), pp. 192–218.
[11] R. Fletcher and M. J. D. Powell, *A rapidly convergent descent method for minimization*, Comput. J., 6 (1963), pp. 163–168.
[12] R. Fletcher and C. M. Reeves, *Function minimization by conjugate gradients*, Comput. J., 7 (1964), pp. 149–154.
[13] P. E. Gill and M. W. Leonard, *Reduced-Hessian quasi-Newton methods for unconstrained optimization*, SIAM J. Optim., 12 (2001), pp. 209–237.
[14] P. E. Gill and W. Murray, *Conjugate-Gradient Methods for Large-Scale Nonlinear Optimization*, Report SOL 79-15, Department of Operations Research, Stanford University, Stanford, CA, 1979.
[15] P. E. Gill, W. Murray, and M. A. Saunders, *SNOPT: An SQP algorithm for large-scale constrained optimization*, SIAM J. Optim., 12 (2002), pp. 979–1006.
[16] P. E. Gill, W. Murray, M. A. Saunders, and M. H. Wright, *User's Guide for NPSOL (Version* 4.0*): A Fortran Package for Nonlinear Programming*, Report SOL 86-2, Department of Operations Research, Stanford University, Stanford, CA, 1986.
[17] P. E. Gill, W. Murray, and M. H. Wright, *Practical Optimization*, Academic Press, London, New York, 1981.
[18] G. H. Golub and C. F. Van Loan, *Matrix Computations*, 2nd ed., The Johns Hopkins University Press, Baltimore, MD, 1989.
[19] L. Kaufman, *Reduced storage, quasi-Newton trust region approaches to function optimization*, SIAM J. Optim., 10 (1999), pp. 56–69.
[20] M. W. Leonard, *Reduced Hessian Quasi-Newton Methods for Optimization*, Ph.D. thesis, Department of Mathematics, University of California, San Diego, 1995.
[21] D. C. Liu and J. Nocedal, *On the limited memory BFGS method for large scale optimization*, Math. Programming, 45 (1989), pp. 503–528.
[22] J. L. Morales and J. Nocedal, *Automatic preconditioning by limited memory quasi-Newton updating*, SIAM J. Optim., 10 (2000), pp. 1079–1096.
[23] J. J. Moré and D. C. Sorensen, *Newton's method*, in Studies in Numerical Analysis, MAA Stud. Math. 24, Mathematical Association of America, Washington, DC, 1984, pp. 29–82.
[24] J. J. Moré and D. J. Thuente, *Line search algorithms with guaranteed sufficient decrease*, ACM Trans. Math. Software, 20 (1994), pp. 286–307.
[25] L. Nazareth, *A relationship between the BFGS and conjugate gradient algorithms and its implications for new algorithms*, SIAM J. Numer. Anal., 16 (1979), pp. 794–800.
[26] J. Nocedal, *Updating quasi-Newton matrices with limited storage*, Math. Comput., 35 (1980), pp. 773–782.
[27] M. J. D. Powell, *Updating conjugate directions by the BFGS formula*, Math. Programming, 38 (1987), pp. 693–726.
[28] M. J. D. Powell, *TOLMIN: A Fortran Package for Linearly Constrained Optimization Calculations*, Report DAMTP/1989/NA2, Department of Applied Mathematics and Theoretical Physics, University of Cambridge, 1989.
[29] M. J. D. Powell and P. L. Toint, *On the estimation of sparse Hessian matrices*, SIAM J. Numer. Anal., 16 (1979), pp. 1060–1074.
[30] D. F. Shanno, *Conjugate-gradient methods with inexact searches*, Math. Oper. Res., 3 (1978), pp. 244–256.

[31] D. Siegel, *Modifying the BFGS Update by a New Column Scaling Technique*, Report DAMTP/1991/NA5, Department of Applied Mathematics and Theoretical Physics, University of Cambridge, 1991.

[32] D. Siegel, *Implementing and Modifying Broyden Class Updates for Large Scale Optimization*, Report DAMTP/1992/NA12, Department of Applied Mathematics and Theoretical Physics, University of Cambridge, 1992.

[33] D. Siegel, *Modifying the BFGS update by a new column scaling technique*, Math. Programming, 66 (1994), pp. 45–78.

[34] C. Zhu, R. H. Byrd, P. Lu, and J. Nocedal, *Algorithm* 778: *L-BFGS-B: Fortran subroutines for large-scale bound-constrained optimization*, ACM Trans. Math. Software, 23 (1997), pp. 550–560.

SIAM J. OPTIM.
Vol. 14, No. 2, pp. 402–414

ASYMPTOTIC BEHAVIOR OF CONTINUOUS TRAJECTORIES FOR PRIMAL-DUAL POTENTIAL-REDUCTION METHODS*

REHA H. TÜTÜNCÜ†

Abstract. This article considers continuous trajectories of the vector fields induced by primal-dual potential-reduction algorithms for solving linear programming problems. It is known that these trajectories converge to the analytic center of the primal-dual optimal face. We establish that this convergence may be tangential to the central path, tangential to the optimal face, or in between, depending on the value of the potential function parameter.

Key words. linear programming, potential functions, potential-reduction methods, central path, continuous trajectories for linear programming

AMS subject classification. 90C05

DOI. 10.1137/S1052623401394948

1. Introduction.

During the past two decades, interior-point methods (IPMs) emerged as one of the most efficient and reliable techniques for the solution of linear programming problems. The development of IPMs and their theoretical convergence analyses often rely on certain continuous trajectories associated with the given linear program. The best known examples of such trajectories are the *central path* and the *weighted centers*—the sets of minimizers of the parametrized standard and weighted logarithmic barrier functions in the interior of the feasible region.

Primal-dual variants of IPMs, which have been very successful in practical implementations, not only solve the given linear program but also its dual. If both the given linear program and its dual have strictly feasible solutions, the primal-dual central path starts from the analytic center of the primal-dual feasible set and converges to the analytic center of the optimal solution set. Similarly, weighted centers converge to weighted analytic centers. This property of the central trajectories led to the development of *path following* IPMs: algorithms that try to reach an optimal solution by generating a sequence of points that are "close" to a corresponding sequence of points on the central path (or the weighted central path) that converge to its limit point.

An alternative characterization of the central path and weighted centers can be obtained by representing them as solutions of certain differential equations. Using this perspective, Adler and Monteiro analyzed the limiting behavior of continuous trajectories associated with primal-only affine-scaling and projective-scaling algorithms as well as a primal-only potential-reduction method [1, 11, 12]. Kojima et al. [7] studied similar trajectories for primal-dual potential-reduction methods.

Potential-reduction algorithms use the following strategy: First, one defines a *potential function* that measures the quality (or potential) of any trial solution of the given problem, combining measures of proximity to the set of optimal solutions, proximity to the feasible set in the case of infeasible interior-points, and a measure of centrality within the feasible region. Potential functions are chosen such that one

*Received by the editors September 7, 2001; accepted for publication (in revised form) April 14, 2003; published electronically October 2, 2003. This research was supported in part by the NSF through grant CCR-9875559.

http://www.siam.org/journals/siopt/14-2/39494.html

†Department of Mathematical Sciences, Carnegie Mellon University, Pittsburgh, PA 15213 (reha@cmu.edu).

approaches an optimal solution of the underlying problem by reducing the potential function. Then the search for an optimal solution can be performed by progressive reduction of the potential function, leading to a potential-reduction algorithm. We refer the reader to two excellent surveys for further details on potential-reduction algorithms [2, 15].

Often, implementations of potential-reduction interior-point algorithms exhibit behavior that is similar to that of path-following algorithms. For example, they take about the same number of iterations as path-following algorithms, and they tend to converge to the analytic center of the optimal face, just like most path-following variants. Since potential-reduction methods do not generally make an effort to follow the central path, this behavior is surprising. In an effort to better understand the limiting behavior of primal-dual potential-reduction algorithms for linear programs this paper studies continuous trajectories associated with the algorithm proposed by Kojima, Mizuno, and Yoshise (KMY) [8], which uses scaled and projected steepest-descent directions for the Tanabe–Todd–Ye (TTY) primal-dual potential function [14, 16].

Using earlier results [9, 10, 7], we show that all trajectories of the vector field induced by the KMY search directions converge to the analytic center of the primal-dual optimal face. Our main results are on the direction of convergence for these trajectories. We demonstrate that their asymptotic behavior depends on the potential function parameter. There is a threshold value of this parameter—the value that makes the TTY potential function homogeneous. When the parameter is below this threshold, the centering is too strong, and the trajectories converge tangentially to the central path. When the parameter is above the threshold, trajectories converge tangentially to the optimal face. However, the direction of convergence of these trajectories depends on the initial point. At the threshold value, the behavior of the trajectories is in between these two extremes and depends on the initial point.

Following this introduction, section 2 discusses continuous trajectories associated with the KMY methods and proves their convergence. Section 3 is devoted to the analysis of the limiting behavior of these trajectories. Our notation is fairly standard: For an n-dimensional vector x, the corresponding capital letter X denotes the $n \times n$ diagonal matrix with $X_{ii} \equiv x_i$. We will use the letter e to denote a column vector with all entries equal to 1, and its dimension will be apparent from the context. We also denote the base of the natural logarithm with e, and sometimes the vector e and the scalar e appear in the same expression, but no confusion should arise. For a given matrix A, we use $\mathcal{R}(A)$ and $\mathcal{N}(A)$ to denote its range (column) and null space. For a vector-valued differentiable function $x(t)$ of a scalar variable t, we use the notation $\dot{x}$ or $\dot{x}(t)$ to denote the vector of the derivatives of its components with respect to t. For n-dimensional vectors x and s, we write xs to denote their Hadamard (componentwise) product. Also, for an n-dimensional vector x, we write x^p to denote the vector $X^p e$, where p can be fractional if $x > 0$.

2. Primal-dual potential-reduction trajectories. We consider linear programs in the following standard form:

$$\text{(LP)} \qquad \begin{array}{lrcl} \min\limits_{x} & c^T x, & & \\ & Ax & = & b, \\ & x & \geq & 0, \end{array} \tag{2.1}$$

where $A \in \Re^{m \times n}$, $b \in \Re^m$, $c \in \Re^n$ are given, and $x \in \Re^n$. Without loss of generality we assume that the constraints are linearly independent. Then the matrix A has full

row rank. Further, we assume that $0 < m < n$; $m = 0$ and $m = n$ correspond to trivial problems.

The linear programming dual of this (primal) problem is

$$\begin{array}{lll} \text{(LD)} & \max\limits_{y,s} & b^T y, \\ & & A^T y + s = c, \\ & & s \geq 0, \end{array} \tag{2.2}$$

where $y \in \Re^m$ and $s \in \Re^n$. We can rewrite the dual problem by eliminating the y variables in (2.2). This is achieved by considering G^T, a null-space basis matrix for A; that is, G is an $(n-m) \times n$ matrix with rank $n-m$, and it satisfies $AG^T = 0$, $GA^T = 0$. Note also that A^T is a null-space basis matrix for G. Further, let $d \in \Re^n$ be a vector satisfying $Ad = b$. Then (2.2) is equivalent to the following problem, which has a high degree of symmetry with (2.1):

$$\begin{array}{lll} \text{(LD}') & \min\limits_{s} & d^T s, \\ & & Gs = Gc, \\ & & s \geq 0. \end{array} \tag{2.3}$$

Let $\mathcal{F}$ and $\mathcal{F}^0$ denote the primal-dual feasible region and its relative interior:

$$\begin{aligned} \mathcal{F} &:= \{(x,s) : Ax = b, Gs = Gc, (x,s) \geq 0\}, \\ \mathcal{F}^0 &:= \{(x,s) : Ax = b, Gs = Gc, (x,s) > 0\}. \end{aligned}$$

We assume that $\mathcal{F}^0$ is nonempty. This assumption has the important consequence that the primal-dual optimal solution set Ω defined below is nonempty and bounded:

$$\Omega := \{(x,s) \in \mathcal{F} : x^T s = 0\}. \tag{2.4}$$

We also define the optimal partition $\mathcal{B} \cup \mathcal{N} = \{1, \ldots, n\}$ for future reference:

$$\begin{aligned} \mathcal{B} &:= \{j : x_j > 0 \text{ for some } (x,s) \in \Omega\}, \\ \mathcal{N} &:= \{j : s_j > 0 \text{ for some } (x,s) \in \Omega\}. \end{aligned}$$

The fact that $\mathcal{B}$ and $\mathcal{N}$ form a partition of $\{1, \ldots, n\}$ is a classical result of Goldman and Tucker. The analytic center of Ω is the point $(x^*, s^*) = ((x^*_{\mathcal{B}}, 0), (0, s^*_{\mathcal{N}}))$, where $x^*_{\mathcal{B}}$ and $s^*_{\mathcal{N}}$ are unique maximizers of the following problems:

$$\begin{array}{lll} \max & \sum\limits_{j \in \mathcal{B}} \ln x_j, & \\ & A_{\mathcal{B}} x_{\mathcal{B}} = b, & \text{and} \\ & x_{\mathcal{B}} > 0 & \end{array} \qquad \begin{array}{lll} \max & \sum\limits_{j \in \mathcal{N}} \ln s_j, & \\ & G_{\mathcal{N}} s_{\mathcal{N}} = Gc, & \\ & s_{\mathcal{N}} > 0. & \end{array} \tag{2.5}$$

The central path $\mathcal{C}$ of the primal-dual feasible set $\mathcal{F}$ is the set of points on which the componentwise product of the primal and dual variables is constant:

$$\mathcal{C} := \{(x(\mu), s(\mu)) \in \mathcal{F}^0 : x(\mu)s(\mu) = \mu e \text{ for some } \mu > 0\}. \tag{2.6}$$

The points on the central path are obtained as unique minimizers of certain barrier problems associated with the primal and dual linear programs, and they converge to the analytic center of the primal-dual optimal face; see, e.g., [18].

While the central path is the main theoretical tool in the construction of path-following algorithms, primal-dual potential-reduction algorithms for linear programming are derived using *potential functions*, i.e., functions that measure the quality (or potential) of trial solutions for the primal-dual pair of problems. The most frequently used primal-dual potential function for linear programming problems is the TTY potential function [14, 16]:

$$\Phi_\rho(x,s) := \rho \ln(x^T s) - \sum_{i=1}^{n} \ln(x_i s_i) \text{ for every } (x,s) > 0. \tag{2.7}$$

When $\rho > n$, the TTY potential function diverges to $-\infty$ along a feasible sequence $\{(x^k, s^k)\}$ only if this sequence is converging to a primal-dual optimal pair of solutions. Therefore, the primal-dual pair of linear programming problems can be solved by minimizing the TTY potential function.

KMY developed a primal-dual algorithm that monotonically reduces the TTY potential function using a scaled and projected steepest-descent search direction using a primal-dual scaling matrix [8]. In the remainder of this article, we will study continuous trajectories that are naturally associated with their algorithm. Given an iterate $(x,s) \in \mathcal{F}^0$, the search direction used by the KMY method is the solution of the following system:

$$\begin{array}{rcrcl} A\Delta x & & & = & 0, \\ & & G\Delta s & = & 0, \\ S\Delta x & + & X\Delta s & = & \dfrac{x^T s}{\rho} e - xs, \end{array} \tag{2.8}$$

where $X = \operatorname{diag}(x)$, $S = \operatorname{diag}(s)$, and e is a vector of ones of appropriate dimension. When we discuss the search direction given by (2.8) and associated trajectories, we will assume that $\rho > n$.

For any given $(x_0, s_0) \in \mathcal{F}^0$, one can associate a trajectory $\{(x(t), s(t)) : t \geq 0\}$ starting from (x_0, s_0) with the property that the tangent direction to the trajectory at any of its points coincides with the KMY direction. In other words, we consider trajectories that solve the following system of ODEs:

$$\begin{bmatrix} A & 0 \\ 0 & G \\ S & X \end{bmatrix} \begin{bmatrix} \dot{x} \\ \dot{s} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \dfrac{x^T s}{\rho} e - xs \end{bmatrix}, \tag{2.9}$$

with the initial condition $(x(0), s(0)) = (x_0, s_0)$. In [7, section 4.3], Kojima et al. study these trajectories and establish that their solution curves satisfy the following system of equations:

$$Ax(t) = b, \quad Gs(t) = Gc, \quad x(t)s(t) = w(t), \quad t \geq 0, \tag{2.10}$$

where

$$w(t) = e^{-t} w_0 + h(t) e \text{ with } w_0 = x_0 s_0 \text{ and} \tag{2.11}$$

$$h(t) = \frac{e^T w_0}{n} \left(\exp\{-(1-\beta)\, t\} - e^{-t} \right) \text{ with } \beta = \frac{n}{\rho}. \tag{2.12}$$

Since $w(0) = w_0$, we will use these two expressions interchangeably. Kojima et al. do not address the existence and uniqueness of the solutions to (2.9) rigorously, but these

results follow easily from standard theory of ODEs; see, e.g., Theorem 1 on p. 162 and Lemma on p. 171 of the textbook by Hirsch and Smale [6]. We also note that Monteiro [12] studies trajectories based on primal-only potential-reduction algorithms and obtains similar but less explicit descriptions of these trajectories.

The characterization of the potential-reduction trajectories using the system (2.10) leads to the following observations.

THEOREM 2.1. *Let $(x(t), s(t))$ for $t \geq 0$ denote the solution of the ODE* (2.9) *with the initial condition $(x(0), s(0)) = (x_0, s_0)$. Then the following statements hold:*

(i) *For $\rho > n$, $\Phi_\rho(x(t), s(t))$ is a decreasing function of t.*
(ii) *When $w_0 = x_0 s_0 = \mu e$ for some $\mu > 0$ (i.e., when (x_0, s_0) is on the central path), then $\{(x(t), s(t)) : t \geq 0\}$ is a subset of the central path $\mathcal{C}$.*
(iii) *$(x(t), s(t))$ converges to the analytic center of the primal-dual optimal face Ω as $t \to \infty$.*

Proof. Lemma 4.14 in [7] proves (i). Observing that $w_0 = \mu e$ implies $w(t) = \mu e^{-(1-\beta)t} e$ with $\beta = \frac{n}{\rho}$, (ii) follows immediately from (2.6) and (2.10). For (iii), first observe that $\frac{w(t)}{\|w(t)\|} \to \frac{e}{\sqrt{n}}$ as $t \to \infty$. Now, the proof of Theorem 9 in [9] (or Corollary 2 of Theorem 5 in [10]) immediately leads to (iii). ∎

So these trajectories converge to a unique point regardless of their starting point as they monotonically decrease the potential function. Further, they include the central path as a special case, giving a theoretical basis for the observation that central path-following search directions are often very good potential-reduction directions as well. Another related result is by Nesterov [13], who observes that the neighborhood of the central path is the region of fastest decrease for a homogeneous potential function.

A direct proof of (iii) in Theorem 2.1 can be obtained using the following result, which will also be useful in the next section.

LEMMA 2.2. *Let $(x(t), s(t))$ for $t \geq 0$ denote the solution of the ODE* (2.9) *with the initial condition $(x(0), s(0)) = (x_0, s_0)$. Then $x_\mathcal{B}(t)$ and $s_\mathcal{N}(t)$ solve the following pair of problems:*

$$\begin{array}{ccc} \max & \sum_{j \in \mathcal{B}} w_j(t) \ln x_j, & \\ & A_\mathcal{B} x_\mathcal{B} = b - A_\mathcal{N} x_\mathcal{N}(t), & \\ & x_\mathcal{B} > 0 & \end{array} \quad \text{and} \quad \begin{array}{ccc} \max & \sum_{j \in \mathcal{N}} w_j(t) \ln s_j, & \\ & G_\mathcal{N} s_\mathcal{N} = Gc - G_\mathcal{B} s_\mathcal{B}(t), & \\ & s_\mathcal{N} > 0. & \end{array} \tag{2.13}$$

Proof. We prove the optimality of $x_\mathcal{B}(t)$ for the first problem in (2.13)—the corresponding result for $s_\mathcal{N}(t)$ can be proven similarly. $x_\mathcal{B}(t)$ is clearly feasible for the given problem. It is optimal if and only if there exists $y \in \Re^m$ such that

$$w_\mathcal{B}(t) x_\mathcal{B}^{-1}(t) = A_\mathcal{B}^T y.$$

From (2.10) we obtain $w_\mathcal{B}(t) x_\mathcal{B}^{-1}(t) = s_\mathcal{B}(t)$. Note that for any s feasible for (LD′) we have that $c - s \in \mathcal{R}(A^T)$, and therefore $c_\mathcal{B} - s_\mathcal{B} \in \mathcal{R}(A_\mathcal{B}^T)$. Furthermore, since $s^* = (0, s_\mathcal{N}^*)$ is also feasible for (LD′) we must have that $c_\mathcal{B} \in \mathcal{R}(A_\mathcal{B}^T)$ and that $s_\mathcal{B}(t) \in \mathcal{R}(A_\mathcal{B}^T)$. This is exactly what we needed. ∎

3. Asymptotic analysis of the trajectories. In the previous section, we saw that all primal-dual potential-reduction trajectories $(x(t), s(t))$ that solve the differential equation (2.9) converge to the analytic center (x^*, s^*) of the primal-dual optimal face Ω regardless of the initial point of the trajectory. In this section, we investigate the direction of convergence for these trajectories. That is, we want to analyze the

limiting behavior of the normalized vectors $(\frac{\dot{x}(t)}{\|\dot{x}(t)\|}, \frac{\dot{s}(t)}{\|\dot{s}(t)\|})$. Inevitably, this analysis is quite technical.

Our strategy for this analysis is as follows. Using the optimal partition $\mathcal{B} \cup \mathcal{N}$ we express the "basic" components of the convergence directions of the trajectories in terms of the "nonbasic" ones in Lemma 3.1. Then we establish a bound on the convergence speed of the "nonbasic" components in Lemma 3.3. The dependence of the convergence direction on ρ, the potential function parameter, becomes apparent at this point, and a case analysis is required. Lemma 3.4 and Theorems 3.5 and 3.6 consider the case $\rho \leq 2n$, while Theorem 3.8 addresses the case $\rho > 2n$.

Let $\beta = \frac{n}{\rho}$, and note that $\beta \in (0, 1)$. We now introduce some notation:

$$
\begin{array}{ll}
\hat{w}_{\mathcal{B}}(t) = w_{\mathcal{B}}(t)e^{(1-\beta)t}, & \hat{w}_{\mathcal{N}}(t) = w_{\mathcal{N}}(t)e^{(1-\beta)t}, \\
d_{\mathcal{B}}(t) = \hat{w}_{\mathcal{B}}^{\frac{1}{2}}(t)x_{\mathcal{B}}^{-1}(t), & d_{\mathcal{N}}(t) = \hat{w}_{\mathcal{N}}^{\frac{1}{2}}(t)s_{\mathcal{N}}^{-1}(t), \\
D_{\mathcal{B}}(t) = \operatorname{diag}\left(d_{\mathcal{B}}(t)\right), & D_{\mathcal{N}}(t) = \operatorname{diag}\left(d_{\mathcal{N}}(t)\right), \\
d_{\mathcal{B}}^{-1}(t) = D_{\mathcal{B}}^{-1}(t)e, & d_{\mathcal{N}}^{-1}(t) = D_{\mathcal{N}}^{-1}(t)e, \\
\tilde{A}_{\mathcal{B}}(t) = A_{\mathcal{B}}D_{\mathcal{B}}^{-1}(t), & \tilde{G}_{\mathcal{N}}(t) = G_{\mathcal{N}}D_{\mathcal{N}}^{-1}(t), \\
\tilde{x}_{\mathcal{B}}(t) = D_{\mathcal{B}}(t)\dot{x}_{\mathcal{B}}(t) = d_{\mathcal{B}}(t)\dot{x}_{\mathcal{B}}(t), & \tilde{s}_{\mathcal{N}}(t) = D_{\mathcal{N}}(t)\dot{s}_{\mathcal{N}}(t) = d_{\mathcal{N}}(t)\dot{s}_{\mathcal{N}}(t), \\
u_{\mathcal{B}}(t) = \hat{w}_{\mathcal{B}}^{-\frac{1}{2}}(t)w_{\mathcal{B}}(0), & u_{\mathcal{N}}(t) = \hat{w}_{\mathcal{N}}^{-\frac{1}{2}}(t)w_{\mathcal{N}}(0).
\end{array}
$$

For our asymptotic analysis, we express "basic" components of the vectors $\dot{x}(t)$ and $\dot{s}(t)$ in terms of the "nonbasic"ones in the next lemma, which forms the backbone of our analysis.

LEMMA 3.1. *Let $(x(t), s(t))$ for $t \geq 0$ denote the solution of the ODE (2.9) with the initial condition $(x(0), s(0)) = (x_0, s_0)$. Then the following equalities hold:*

$$
D_{\mathcal{B}}(t)\dot{x}_{\mathcal{B}}(t) = -\tilde{A}_{\mathcal{B}}^{+}(t)A_{\mathcal{N}}\dot{x}_{\mathcal{N}}(t) - \frac{n \cdot e^{-\beta t}}{\rho\left(1 - e^{-\beta t}\right)}\left(I - \tilde{A}_{\mathcal{B}}^{+}(t)\tilde{A}_{\mathcal{B}}(t)\right)u_{\mathcal{B}}(t), \tag{3.1}
$$

$$
D_{\mathcal{N}}(t)\dot{s}_{\mathcal{N}}(t) = -\tilde{G}_{\mathcal{N}}^{+}(t)G_{\mathcal{B}}\dot{s}_{\mathcal{B}}(t) - \frac{n \cdot e^{-\beta t}}{\rho\left(1 - e^{-\beta t}\right)}\left(I - \tilde{G}_{\mathcal{N}}^{+}(t)\tilde{G}_{\mathcal{N}}(t)\right)u_{\mathcal{N}}(t). \tag{3.2}
$$

Here, $\tilde{A}_{\mathcal{B}}^{+}(t)$ and $\tilde{G}_{\mathcal{N}}^{+}(t)$ denote the pseudoinverse of $\tilde{A}_{\mathcal{B}}(t)$ and $\tilde{G}_{\mathcal{N}}(t)$, respectively.

Proof. We prove only the first identity; the second one follows similarly. Recall from Lemma 2.2 that $x_{\mathcal{B}}(t)$ solves the first problem in (2.13). Therefore, as in the proof of Lemma 2.2, we must have that

$$
w_{\mathcal{B}}(t)x_{\mathcal{B}}^{-1}(t) \in \mathcal{R}(A_{\mathcal{B}}^{T}). \tag{3.3}
$$

Differentiating with respect to t we obtain

$$
\begin{array}{l}
w_{\mathcal{B}}(t)x_{\mathcal{B}}^{-2}(t)\dot{x}_{\mathcal{B}}(t) - \dot{w}_{\mathcal{B}}(t)x_{\mathcal{B}}^{-1}(t) \in \mathcal{R}(A_{\mathcal{B}}^{T}) \text{ or} \\
w_{\mathcal{B}}(t)x_{\mathcal{B}}^{-1}(t)\dot{x}_{\mathcal{B}}(t) - \dot{w}_{\mathcal{B}}(t) \in \mathcal{R}(X_{\mathcal{B}}(t)A_{\mathcal{B}}^{T}).
\end{array}
$$

Observe that

$$
\dot{w}_{\mathcal{B}}(t) = -w_{\mathcal{B}}(0)e^{-t} + \dot{h}(t)e_{\mathcal{B}} = -w_{\mathcal{B}}(t) + \frac{e^{T}w_0}{\rho}e^{-(1-\beta)t}e_{\mathcal{B}}.
$$

Therefore, from $\hat{w}_{\mathcal{B}}(t) = w_{\mathcal{B}}(t)e^{(1-\beta)t}$, we obtain

$$
\hat{w}_{\mathcal{B}}(t)x_{\mathcal{B}}^{-1}(t)\dot{x}_{\mathcal{B}}(t) + \hat{w}_{\mathcal{B}}(t) - \frac{e^{T}w_0}{\rho}e_{\mathcal{B}} \in \mathcal{R}(X_{\mathcal{B}}(t)A_{\mathcal{B}}^{T}). \tag{3.4}
$$

From (3.3) it also follows that $\hat{w}_{\mathcal{B}}(t) \in \mathcal{R}(X_{\mathcal{B}}(t)A_{\mathcal{B}}^T)$. Note also that

$$\frac{e^T w_0}{\rho} e_{\mathcal{B}} = \frac{n}{\rho\,(1-e^{-\beta t})}\hat{w}_{\mathcal{B}}(t) - \frac{n\cdot e^{-\beta t}}{\rho\,(1-e^{-\beta t})} w_{\mathcal{B}}(0).$$

Combining these observations with (3.4) we get

$$\hat{w}_{\mathcal{B}}(t)x_{\mathcal{B}}^{-1}(t)\dot{x}_{\mathcal{B}}(t) + \frac{n\cdot e^{-\beta t}}{\rho\,(1-e^{-\beta t})} w_{\mathcal{B}}(0) \in \mathcal{R}(X_{\mathcal{B}}(t)A_{\mathcal{B}}^T). \tag{3.5}$$

Next, observe that

$$A_{\mathcal{B}}\dot{x}_{\mathcal{B}}(t) = -A_{\mathcal{N}}\dot{x}_{\mathcal{N}}(t). \tag{3.6}$$

Using the notation introduced before the statement of the lemma, (3.5) and (3.6) can be rewritten as follows:

$$\tilde{x}_{\mathcal{B}}(t) + \frac{n\cdot e^{-\beta t}}{\rho\,(1-e^{-\beta t})} u_{\mathcal{B}}(t) \in \mathcal{R}(\tilde{A}_{\mathcal{B}}^T), \tag{3.7}$$

$$\tilde{A}_{\mathcal{B}}(t)\tilde{x}_{\mathcal{B}}(t) = -A_{\mathcal{N}}\dot{x}_{\mathcal{N}}(t). \tag{3.8}$$

Let $\tilde{A}_{\mathcal{B}}^+(t)$ denote the pseudoinverse of $\tilde{A}_{\mathcal{B}}(t)$ [3]. For example, if rank$(\tilde{A}_{\mathcal{B}}(t))$=$m$, then $\tilde{A}_{\mathcal{B}}^+(t) = \tilde{A}_{\mathcal{B}}^T(t)(\tilde{A}_{\mathcal{B}}(t)\tilde{A}_{\mathcal{B}}^T(t))^{-1}$. Then, $P_{\mathcal{R}(\tilde{A}_{\mathcal{B}}^T)} := \tilde{A}_{\mathcal{B}}^+(t)\tilde{A}_{\mathcal{B}}(t)$ is the orthogonal projection matrix onto $\mathcal{R}(\tilde{A}_{\mathcal{B}}^T)$ and $P_{\mathcal{N}(\tilde{A}_{\mathcal{B}})} := I - \tilde{A}_{\mathcal{B}}^+(t)\tilde{A}_{\mathcal{B}}(t)$ is the orthogonal projection matrix onto $\mathcal{N}(\tilde{A}_{\mathcal{B}})$ [3]. From (3.8) we obtain

$$P_{\mathcal{R}(\tilde{A}_{\mathcal{B}}^T)}\tilde{x}_{\mathcal{B}}(t) = \tilde{A}_{\mathcal{B}}^+(t)\tilde{A}_{\mathcal{B}}(t)\tilde{x}_{\mathcal{B}}(t) = -\tilde{A}_{\mathcal{B}}^+(t)A_{\mathcal{N}}\dot{x}_{\mathcal{N}}(t),$$

and from (3.7), using the fact that $\mathcal{R}(\tilde{A}_{\mathcal{B}}^T)$ and $\mathcal{N}(\tilde{A}_{\mathcal{B}})$ are orthogonal to each other, we get

$$P_{\mathcal{N}(\tilde{A}_{\mathcal{B}})}\tilde{x}_{\mathcal{B}}(t) = -\frac{n\cdot e^{-\beta t}}{\rho\,(1-e^{-\beta t})}\left(I - \tilde{A}_{\mathcal{B}}^+(t)\tilde{A}_{\mathcal{B}}(t)\right) u_{\mathcal{B}}(t).$$

Combining these results, we have

$$\begin{aligned}\tilde{x}_{\mathcal{B}}(t) &= P_{\mathcal{R}(\tilde{A}_{\mathcal{B}}^T)}\tilde{x}_{\mathcal{B}}(t) + P_{\mathcal{N}(\tilde{A}_{\mathcal{B}})}\tilde{x}_{\mathcal{B}}(t)\\ &= -\tilde{A}_{\mathcal{B}}^+(t)A_{\mathcal{N}}\dot{x}_{\mathcal{N}}(t) - \frac{n\cdot e^{-\beta t}}{\rho\,(1-e^{-\beta t})}\left(I - \tilde{A}_{\mathcal{B}}^+(t)\tilde{A}_{\mathcal{B}}(t)\right) u_{\mathcal{B}}(t),\end{aligned}$$

which gives (3.1). □

To determine the convergence directions of the trajectories, we need to study the relative convergence speeds of $\dot{x}_{\mathcal{B}}(t), \dot{x}_{\mathcal{N}}(t)$, etc. Thus, we compute limits of some of the expressions that appear in (3.1) and (3.2):

$$\lim_{t\to\infty} \hat{w}_{\mathcal{B}}(t) = \frac{e^T w_0}{n} e_{\mathcal{B}}, \qquad \lim_{t\to\infty} \hat{w}_{\mathcal{N}}(t) = \frac{e^T w_0}{n} e_{\mathcal{N}}, \tag{3.9}$$

$$\lim_{t\to\infty} D_{\mathcal{B}}(t) = \sqrt{\frac{e^T w_0}{n}}(X_{\mathcal{B}}^*)^{-1}, \qquad \lim_{t\to\infty} D_{\mathcal{N}}(t) = \sqrt{\frac{e^T w_0}{n}}(S_{\mathcal{N}}^*)^{-1}, \tag{3.10}$$

$$\lim_{t\to\infty} \tilde{A}_{\mathcal{B}}(t) = \sqrt{\frac{n}{e^T w_0}} A_{\mathcal{B}} X_{\mathcal{B}}^*, \qquad \lim_{t\to\infty} \tilde{G}_{\mathcal{N}}(t) = \sqrt{\frac{n}{e^T w_0}} G_{\mathcal{N}} S_{\mathcal{N}}^*, \tag{3.11}$$

$$\lim_{t\to\infty} u_{\mathcal{B}}(t) = \sqrt{\frac{n}{e^T w_0}} w_{\mathcal{B}}(0), \qquad \lim_{t\to\infty} u_{\mathcal{N}}(t) = \sqrt{\frac{n}{e^T w_0}} w_{\mathcal{N}}(0). \tag{3.12}$$

LEMMA 3.2.

$$\lim_{t\to\infty} \tilde{A}^{+}_{\mathcal{B}}(t) = \sqrt{\frac{e^T w_0}{n}}\,(A_{\mathcal{B}} X^*_{\mathcal{B}})^{+}\,, \tag{3.13}$$

$$\lim_{t\to\infty} \tilde{G}^{+}_{\mathcal{N}}(t) = \sqrt{\frac{e^T w_0}{n}}\,(G_{\mathcal{N}} S^*_{\mathcal{N}})^{+}\,. \tag{3.14}$$

Proof. This result about the limiting properties of the pseudoinverses is an immediate consequence of Lemma 2.3 in [4] and (3.11). ∎

Differentiating the identity

$$x(t)s(t) = w(t) = e^{-t}w_0 + h(t)e,$$

we obtain

$$\begin{aligned} x(t)\dot{s}(t) + \dot{x}(t)s(t) &= -e^{-t}w_0 + \dot{h}(t)e \\ &= -e^{-t}w_0 - \frac{e^T w_0}{n}\,(1-\beta)\,e^{-(1-\beta)t}e + \frac{e^T w_0}{n}e^{-t}e. \end{aligned} \tag{3.15}$$

Next, we will establish that $\dot{x}_{\mathcal{N}}(t)$ and $\dot{s}_{\mathcal{B}}(t)$ converge to zero no slower than $\exp\{-(1-\beta)t\}$. For this purpose, we consider the normalized direction vectors $(\hat{x}, \hat{s})$ which are defined as follows:

$$\hat{x}(t) = \exp\{(1-\beta)\,t\}\,\dot{x}(t), \text{ and } \hat{s}(t) = \exp\{(1-\beta)\,t\}\,\dot{s}(t). \tag{3.16}$$

From (3.15) it follows that

$$x(t)\hat{s}(t) + \hat{x}(t)s(t) = -\frac{e^T w_0}{n}\,(1-\beta)\,e + e^{-\beta t}\left(\frac{e^T w_0}{n}e - w_0\right). \tag{3.17}$$

The expression on the right-hand side of (3.17) is clearly bounded. With some more work, we have the following conclusion.

LEMMA 3.3. *Let* $(\hat{x}(t), \hat{s}(t))$ *be as in* (3.16), *and assume that* $\rho > n$. *Then* $(\hat{x}_{\mathcal{N}}(t), \hat{s}_{\mathcal{B}}(t))$ *remains bounded as* t *tends to* ∞.

Proof. We will prove that $x(t)\hat{s}(t)$ and $\hat{x}(t)s(t)$ remain bounded as $t \to \infty$. Then, since $x_{\mathcal{B}}(t)$ and $s_{\mathcal{N}}(t)$ converge to $x^*_{\mathcal{B}} > 0$ and $s^*_{\mathcal{N}} > 0$, respectively, and therefore remain bounded away from zero, we can conclude that $\hat{x}_{\mathcal{N}}(t)$ and $\hat{s}_{\mathcal{B}}(t)$ remain bounded.

Since the right-hand side of (3.17) is bounded as t tends to ∞, it is sufficient to show that $[x(t)\hat{s}(t)]^T\,[\hat{x}(t)s(t)]$ remains bounded below to conclude that both $x(t)\hat{s}(t)$ and $\hat{x}(t)s(t)$ have bounded norms as $t \to \infty$.

Let $v(t) = x^{\frac{1}{2}}(t)s^{\frac{1}{2}}(t) = w^{\frac{1}{2}}(t)$ and $\delta(t) = x^{\frac{1}{2}}(t)s^{-\frac{1}{2}}(t)$. Then

$$x(t)\hat{s}(t) = v(t)\delta(t)\hat{s}(t) \text{ and } \hat{x}(t)s(t) = v(t)\delta^{-1}(t)\hat{x}(t).$$

Note that $[\delta(t)\hat{s}(t)]^T \left[\delta^{-1}(t)\hat{x}(t)\right] = [\delta(t)\dot{s}(t)]^T \left[\delta^{-1}(t)\dot{x}(t)\right] = 0$. Let $V(t) = \mathrm{diag}(v(t))$, $\Delta(t) = \mathrm{diag}(\delta(t))$, and $W_0 = \mathrm{diag}(w_0)$. Then $V^2(t) = X(t)S(t) = \frac{e^T w_0}{n}\left(e^{-(1-\beta)t} - e^{-t}\right) I + e^{-t} W_0$. Now,

$$\begin{aligned}
[x(t)\hat{s}(t)]^T [\hat{x}(t)s(t)] &= [v(t)\delta(t)\hat{s}(t)]^T \left[v(t)\delta^{-1}(t)\hat{x}(t)\right] \\
&= [\delta(t)\hat{s}(t)]^T V^2(t) \left[\delta^{-1}(t)\hat{x}(t)\right] \\
&= e^{2(1-\beta)t} [\delta(t)\dot{s}(t)]^T V^2(t) \left[\delta^{-1}(t)\dot{x}(t)\right] \\
&= \frac{e^T w_0}{n}\left(e^{(1-\beta)t} - e^{(1-2\beta)t}\right) [\delta(t)\dot{s}(t)]^T \left[\delta^{-1}(t)\dot{x}(t)\right] \\
&\quad + e^{(1-2\beta)t} [\delta(t)\dot{s}(t)]^T W_0 \left[\delta^{-1}(t)\dot{x}(t)\right] \\
&= e^{-\beta t} \left[e^{(1-\beta)t/2}\delta(t)\dot{s}(t)\right]^T W_0 \left[e^{(1-\beta)t/2}\delta^{-1}(t)\dot{x}(t)\right]. \qquad (3.18)
\end{aligned}$$

Recall from (3.9) that $\lim_{t\to\infty} \hat{w}_j(t) = \lim_{t\to\infty} e^{(1-\beta)t} w_j(t) = \frac{e^T w_0}{n}$ for all j. Therefore, we have $\lim_{t\to\infty} \sqrt{\hat{w}_j(t)} = \lim_{t\to\infty} e^{(1-\beta)t/2} v_j(t) = \sqrt{\frac{e^T w_0}{n}}$, and defining

$$\tilde{v}(t) = e^{(1-\beta)t/2}\left(v(t) - \frac{v(t)^T v(t)}{\rho} v^{-1}(t)\right),$$

we have $\lim_{t\to\infty} \tilde{v}_j(t) = (1-\beta)\sqrt{\frac{e^T w_0}{n}}$.

Now, recalling (2.8) we observe that the vectors $e^{(1-\beta)t/2}\delta^{-1}(t)\dot{x}(t)$ and $e^{(1-\beta)t/2}\delta(t)\dot{s}(t)$ are orthogonal projections of the vector $-\tilde{v}(t)$ into the null space of $A\Delta(t)$ and range space of $[A\Delta(t)]^T$, respectively. Since we showed that the vector $\tilde{v}(t)$ is convergent as t tends to ∞, both of these projections converge, and therefore the expression in (3.18) converges to zero. Thus, $x(t)\hat{s}(t)$ and $\hat{x}(t)s(t)$ have bounded norms as $t \to \infty$. ∎

It is interesting that the conclusion of the lemma above holds for any $\rho \geq 0$. An alternative proof of Lemma 3.3 can be obtained using the proof technique in [5]. Combining (3.1) and Lemmas 3.2 and 3.3 we obtain the following result.

LEMMA 3.4. *Let* $(\hat{x}(t), \hat{s}(t))$ *be as in* (3.16), *and assume that* $n < \rho \leq 2n$. *Then* $(\hat{x}(t), \hat{s}(t))$ *remains bounded as* t *tends to* ∞.

Proof. From (3.1) we have that

$$D_{\mathcal{B}}(t)\hat{x}_{\mathcal{B}}(t) = -\tilde{A}_{\mathcal{B}}^{+}(t) A_{\mathcal{N}} \hat{x}_{\mathcal{N}}(t) - \frac{n \cdot e^{(1-2\beta)t}}{\rho\left(1 - e^{-\beta t}\right)} \left(I - \tilde{A}_{\mathcal{B}}^{+}(t)\tilde{A}_{\mathcal{B}}(t)\right) u_{\mathcal{B}}(t). \qquad (3.19)$$

When $\rho \leq 2n$, the factor $\frac{n \cdot e^{(1-2\beta)t}}{\rho(1-e^{-\beta t})}$ is convergent as t tends to ∞. Now, using Lemma 3.2 and (3.11)–(3.12), we conclude that the second term in the right-hand side of the equation above remains bounded. Combining this observation with the fact that $\hat{x}_{\mathcal{N}}(t)$ remains bounded as t tends to ∞, we obtain that $D_{\mathcal{B}}(t)\hat{x}_{\mathcal{B}}(t)$ remains bounded. Using (3.10) we conclude that $\hat{x}_{\mathcal{B}}(t)$ is also bounded as t tends to ∞. The fact that $\hat{s}(t)$ is bounded follows similarly. ∎

Now, the following two results are easy to prove.

THEOREM 3.5. *Let* $(\hat{x}(t), \hat{s}(t))$ *be as in* (3.16). *Then we have that* $\lim_{t\to\infty} \hat{x}_{\mathcal{N}}(t)$ *and* $\lim_{t\to\infty} \hat{s}_{\mathcal{B}}(t)$ *exist and satisfy the following equations:*

$$\lim_{t\to\infty} \hat{x}_{\mathcal{N}}(t) = -\frac{e^T w_0}{n} (1-\beta) (s^*_{\mathcal{N}})^{-1}, \tag{3.20}$$

$$\lim_{t\to\infty} \hat{s}_{\mathcal{B}}(t) = -\frac{e^T w_0}{n} (1-\beta) (x^*_{\mathcal{B}})^{-1}. \tag{3.21}$$

Proof. From (3.17) we have that

$$x_{\mathcal{B}}(t)\hat{s}_{\mathcal{B}}(t) + \hat{x}_{\mathcal{B}}(t)s_{\mathcal{B}}(t) = -\frac{e^T w_0}{n} (1-\beta) e_{\mathcal{B}} + e^{-\beta t} \left(\frac{e^T w_0}{n} e_{\mathcal{B}} - w_{\mathcal{B}}(0) \right).$$

Taking the limit on the right-hand side as $t \to \infty$ we obtain $-\frac{e^T w_0}{n}(1-\beta)e_{\mathcal{B}}$. Since $s_{\mathcal{B}}(t) \to 0$ and $\hat{x}_{\mathcal{B}}(t)$ is bounded, we must then have that $x_{\mathcal{B}}(t)\hat{s}_{\mathcal{B}}(t)$ converges to $-\frac{e^T w_0}{n}(1-\beta)e_{\mathcal{B}}$. Since $x_{\mathcal{B}}(t) \to x^*_{\mathcal{B}}$, it follows that $\lim_{t\to\infty} \hat{s}_{\mathcal{B}}(t)$ exists and satisfies (3.21). The corresponding result for $\hat{x}_{\mathcal{N}}(t)$ follows identically. ∎

Let

$$\xi_{\mathcal{B}} = X^*_{\mathcal{B}} (A_{\mathcal{B}} X^*_{\mathcal{B}})^+ A_{\mathcal{N}} (s^*_{\mathcal{N}})^{-1}, \qquad \sigma_{\mathcal{N}} = S^*_{\mathcal{N}} (G_{\mathcal{N}} S^*_{\mathcal{N}})^+ G_{\mathcal{B}} (x^*_{\mathcal{B}})^{-1},$$
$$\pi_{\mathcal{B}} = X^*_{\mathcal{B}} \left(I - (A_{\mathcal{B}} X^*_{\mathcal{B}})^+ A_{\mathcal{B}} X^*_{\mathcal{B}} \right) w_{\mathcal{B}}(0), \qquad \pi_{\mathcal{N}} = S^*_{\mathcal{N}} \left(I - (G_{\mathcal{N}} S^*_{\mathcal{N}})^+ G_{\mathcal{N}} S^*_{\mathcal{N}} \right) w_{\mathcal{N}}(0).$$

Observe that $\pi_{\mathcal{B}} = 0$ if and only if $w_{\mathcal{B}}(0) \in \mathcal{R}(X^*_{\mathcal{B}} A^T_{\mathcal{B}})$, which holds, for example, when (x_0, s_0) is on the central path and $w(0) = \mu e$ for some $\mu > 0$—the observation that $e \in \mathcal{R}(X^*_{\mathcal{B}} A^T_{\mathcal{B}})$ follows easily from the optimality of $x^*_{\mathcal{B}}$ for the first problem in (2.5). Similarly, $\pi_{\mathcal{N}} = 0$ if and only if $w_{\mathcal{N}}(0) \in \mathcal{R}(S^*_{\mathcal{N}} G^T_{\mathcal{N}})$.

THEOREM 3.6. *Let* $(\hat{x}(t), \hat{s}(t))$ *be as in* (3.16), *and assume that* $\rho \le 2n$. *Then we have that* $\lim_{t\to\infty} \hat{x}_{\mathcal{B}}(t)$ *and* $\lim_{t\to\infty} \hat{s}_{\mathcal{N}}(t)$ *exist. When* $\rho < 2n$ *we have the following identities:*

$$\lim_{t\to\infty} \hat{x}_{\mathcal{B}}(t) = \frac{e^T w_0}{n} (1-\beta) \xi_{\mathcal{B}}, \tag{3.22}$$

$$\lim_{t\to\infty} \hat{s}_{\mathcal{N}}(t) = \frac{e^T w_0}{n} (1-\beta) \sigma_{\mathcal{N}}. \tag{3.23}$$

When $\rho = 2n$, *the following equations hold:*

$$\lim_{t\to\infty} \hat{x}_{\mathcal{B}}(t) = \frac{e^T w_0}{2n} \xi_{\mathcal{B}} - \frac{n}{2(e^T w_0)} \pi_{\mathcal{B}}, \tag{3.24}$$

$$\lim_{t\to\infty} \hat{s}_{\mathcal{N}}(t) = \frac{e^T w_0}{2n} \sigma_{\mathcal{N}} - \frac{n}{2(e^T w_0)} \pi_{\mathcal{N}}. \tag{3.25}$$

Proof. Recall (3.19). When $\rho < 2n$, the second term on the right-hand side converges to zero since $e^{(1-2\beta)t}$ tends to zero and everything else is bounded. Thus, using (3.10) and (3.11) we have $\lim_{t\to\infty} \hat{x}_{\mathcal{B}}(t) = -X^*_{\mathcal{B}} (A_{\mathcal{B}} X^*_{\mathcal{B}})^+ A_{\mathcal{N}} \lim_{t\to\infty} \hat{x}_{\mathcal{N}}(t)$, and (3.22) is obtained using Theorem 3.5. Similarly, one obtains (3.23).

When $\rho = 2n$, the factor in front of the second term in (3.19) converges to the positive constant $\beta = \frac{1}{2}$. Therefore, using Theorem 3.5 and (3.9)–(3.12) we get (3.24) and (3.25). ∎

Limits of the normalized vectors $(\frac{\dot{x}(t)}{\|\dot{x}(t)\|}, \frac{\dot{s}(t)}{\|\dot{s}(t)\|})$ are obtained immediately from Theorems 3.5 and 3.6.

COROLLARY 3.7. *Let $(x(t), s(t))$ for $t \geq 0$ denote the solution of the ODE (2.9) with the initial condition $(x(0), s(0)) = (x_0, s_0)$ with $(x^0, s^0) \in \mathcal{F}^0$, and assume that $\rho \leq 2n$. All trajectories of this form satisfy the following equations:*

$$\lim_{t\to\infty} \frac{\dot{x}(t)}{\|\dot{x}(t)\|} = \frac{q_P}{\|q_P\|}, \qquad \lim_{t\to\infty} \frac{\dot{s}(t)}{\|\dot{s}(t)\|} = \frac{q_D}{\|q_D\|}, \tag{3.26}$$

where

$$q_P = \begin{bmatrix} \xi_{\mathcal{B}} \\ -(s^*_{\mathcal{N}})^{-1} \end{bmatrix} \text{ and } \quad q_D = \begin{bmatrix} -(x^*_{\mathcal{B}})^{-1} \\ \sigma_{\mathcal{N}} \end{bmatrix} \qquad \text{if } \rho < 2n, \tag{3.27}$$

$$q_P = \begin{bmatrix} \xi_{\mathcal{B}} - \left(\frac{n}{e^T w_0}\right)^2 \pi_{\mathcal{B}} \\ -(s^*_{\mathcal{N}})^{-1} \end{bmatrix} \text{ and } \quad q_D = \begin{bmatrix} -(x^*_{\mathcal{B}})^{-1} \\ \sigma_{\mathcal{N}} - \left(\frac{n}{e^T w_0}\right)^2 \pi_{\mathcal{N}} \end{bmatrix} \qquad \text{if } \rho = 2n.$$

When $\rho = 2n$ the TTY potential-function $\Phi_\rho(x, y)$ is a homogeneous function and $\exp\{\Phi_\rho(x, y)\}$ is a convex function for all $\rho \geq 2n$ [17]. The value $2n$ also represents a threshold value for the convergence behavior of the KMY trajectories. When $\rho = 2n$ the direction of convergence depends on the initial point $(x^0, s^0) \in \mathcal{F}^0$, as indicated by the appearance of the $w_0 = x^0 s^0$ terms in the formulas. We note that when $\rho < 2n$ the asymptotic direction of convergence does not depend on the initial point and is identical to that of the central path. Therefore, when $\rho < 2n$ all trajectories of the vector field given by the search direction of the KMY primal-dual potential-reduction algorithm converge to the analytic center of the optimal face tangentially to the central path. We show below that the asymptotic behavior of the trajectories is significantly different when $\rho > 2n$.

THEOREM 3.8. *Let $(x(t), s(t))$ for $t \geq 0$ denote the solution of the ODE (2.9) with the initial condition $(x(0), s(0)) = (x_0, s_0)$, and assume that $\rho > 2n$. Define*

$$\bar{x}(t) = e^{\beta t}\dot{x}(t) \text{ and } \quad \bar{s}(t) = e^{\beta t}\dot{s}(t). \tag{3.28}$$

If $\pi_{\mathcal{B}} \neq 0$ and $\pi_{\mathcal{N}} \neq 0$, then we have that $\lim_{t\to\infty} \bar{x}(t)$ and $\lim_{t\to\infty} \bar{s}(t)$ exist and satisfy the following equations:

$$\lim_{t\to\infty} \frac{\bar{x}(t)}{\|\bar{x}(t)\|} = \frac{q_P}{\|q_P\|} \text{ and } \quad \lim_{t\to\infty} \frac{\bar{s}(t)}{\|\bar{s}(t)\|} = \frac{q_D}{\|q_D\|}, \tag{3.29}$$

where

$$q_P = \begin{bmatrix} -\pi_{\mathcal{B}} \\ 0_{\mathcal{N}} \end{bmatrix} \text{ and } \quad q_D = \begin{bmatrix} 0_{\mathcal{B}} \\ -\pi_{\mathcal{N}} \end{bmatrix}.$$

Proof. From (3.1) we have that

$$D_{\mathcal{B}}(t)\bar{x}_{\mathcal{B}}(t) = -\tilde{A}^+_{\mathcal{B}}(t)A_{\mathcal{N}}\bar{x}_{\mathcal{N}}(t) - \frac{n}{\rho\,(1 - e^{-\beta t})}\left(I - \tilde{A}^+_{\mathcal{B}}(t)\tilde{A}_{\mathcal{B}}(t)\right)u_{\mathcal{B}}(t). \tag{3.30}$$

Note that $\bar{x}_N(t) = e^{-(1-2\beta)t}\hat{x}_{\mathcal{N}}(t)$. Since $\hat{x}_{\mathcal{N}}(t)$ is bounded and $-(1 - 2\beta) < 0$, we conclude that $\bar{x}_N(t) \to 0$. Therefore, using (3.30) and (3.9)–(3.12) we observe that $\bar{x}_B(t)$ converges to a positive multiple of $-\pi_{\mathcal{B}} \neq 0$ and immediately obtain the first equation in (3.29). The second identity in (3.29) is obtained similarly. ∎

This final theorem indicates that when $\rho > 2n$, most trajectories associated with the KMY algorithm converge to the analytic center of the optimal face tangentially to the optimal face, and their direction of convergence depends on the initial point. In the exceptional case of $\pi_{\mathcal{B}} = 0$ or $\pi_{\mathcal{N}} = 0$ (for example, when (x_0, s_0) is on the central path), the last term in (3.30) no longer dominates the right-hand side, and in such cases we conjecture that the trajectory converges tangentially to the central path.

We conclude by noting the similarity of our asymptotic results to those of Monteiro [12]. In his analysis of the trajectories based on primal-only potential-reduction algorithms, Monteiro also finds that there is a threshold value of the potential function parameter that leads to different asymptotic behavior. In his case, this threshold value is $2\,|\mathcal{N}|$ rather than $2n$, where $|\mathcal{N}|$ denotes the cardinality of the set $\mathcal{N}$ from the optimal partition of $\{1, \ldots, n\}$. Just like in our case, when the potential function parameter is above this value, his trajectories converge tangentially to the central path, and below the threshold the convergence is tangential to the optimal face.

Acknowledgments. The conclusion of Lemma 3.3 was stated without a proof in an earlier draft of this paper. The author would like to thank Margaréta Halická, whose careful reading led to the correction of this omission. She also provided an alternative argument for the proof. The author would also like to thank an anonymous referee for bringing references [7, 9, 10] to his attention and making several constructive comments. Results in these references helped make this a more concise paper.

REFERENCES

[1] I. Adler and R. D. C. Monteiro, *Limiting behavior of the affine scaling continuous trajectories for linear programming problems*, Math. Programming, 50 (1991), pp. 29–51.

[2] K. M. Anstreicher, *Potential reduction algorithms*, in Interior Point Methods of Mathematical Programming, T. Terlaky, ed., Kluwer Academic Publishers, Dordrecht, The Netherlands, 1996, pp. 125–158.

[3] G. H. Golub and C. F. Van Loan, *Matrix Computations*, 2nd ed., The Johns Hopkins University Press, Baltimore, 1989.

[4] O. Güler, *Limiting behavior of weighted central paths in linear programming*, Math. Programming, 65 (1994), pp. 347–363.

[5] M. Halická, *Two simple proofs for analyticity of the central path*, Oper. Res. Lett., 28 (2001), pp. 9–19.

[6] M. W. Hirsch and S. Smale, *Differential Equations, Dynamical Systems, and Linear Algebra*, Academic Press, New York, 1974.

[7] M. Kojima, N. Megiddo, T. Noma, and A. Yoshise, *A Unified Approach to Interior Point Algorithms for Linear Complementarity Problems,* Lecture Notes in Comput. Sci. 538, Springer-Verlag, Berlin, 1991.

[8] M. Kojima, S. Mizuno, and A. Yoshise, *An $O(\sqrt{n}L)$ iteration potential reduction algorithm for linear complementarity problems*, Math. Programming, 50 (1991), pp. 331–342.

[9] L. McLinden, *An analogue of Moreau's proximation theorem, with applications to the nonlinear complementarity problem*, Pacific J. Math., 88 (1980), pp. 101–161.

[10] L. McLinden, *The complementarity problem for maximal monotone multifunctions*, in Variational Inequalities and Complementarity Problems, R. W. Cottle, F. Giannessi, and J.-L. Lions, eds., Wiley, New York, 1980, pp. 251–270.

[11] R. D. C. Monteiro, *Convergence and boundary behavior of the projective scaling trajectories for linear programming*, Math. Oper. Res., 16 (1991), pp. 842–858.

[12] R. D. C. Monteiro, *On the continuous trajectories for potential reduction algorithms for linear programming*, Math. Oper. Res., 17 (1992), pp. 225–253.

[13] Yu. Nesterov, *Long-step strategies in interior-point primal-dual methods*, Math. Programming, 76 (1997), pp. 47–94.

[14] K. Tanabe, *Centered Newton method for mathematical programming*, in System Modelling and Optimization, Lecture Notes in Control and Inform. Sci. 113, M. Iri and K. Yajima, eds., Springer-Verlag, Berlin, 1988, pp. 197–206.
[15] M. J. Todd, *Potential-reduction methods in mathematical programming*, Math. Programming, 76 (1997), pp. 3–45.
[16] M. J. Todd and Y. Ye, *A centered projective algorithm for linear programming*, Math. Oper. Res., 15 (1990), pp. 508–529.
[17] R. H. Tütüncü, *A primal-dual variant of the Iri-Imai algorithm for linear programming*, Math. Oper. Res., 25 (2000), pp. 195–213.
[18] S. J. Wright, *Primal-Dual Interior-Point Methods*, SIAM, Philadelphia, 1997.

SIAM J. OPTIM.
Vol. 14, No. 2, pp. 415–438

FRAMES AND GRIDS IN UNCONSTRAINED AND LINEARLY CONSTRAINED OPTIMIZATION: A NONSMOOTH APPROACH*

C. J. PRICE† AND I. D. COOPE†

Abstract. This paper describes a class of frame-based direct search methods for unconstrained and linearly constrained optimization. A template is described and analyzed using Clarke's nonsmooth calculus. This provides a unified and simple approach to earlier results for grid- and frame-based methods, and also provides partial convergence results when the objective function is not smooth, undefined in some places, or both. The template also covers many new methods which combine elements of previous ideas using frames and grids. These new methods include grid-based simple descent algorithms which allow moving to points off the grid at every iteration and can automatically control the grid size, provided function values are available. The concept of a grid is also generalized to that of an admissible set, which allows sets, for example, with circular symmetries. The method is applied to linearly constrained problems using a simple barrier approach.

Key words. derivative-free optimization, positive basis methods, nonsmooth convergence analysis, frame-based methods

AMS subject classifications. 49M30, 65K05

DOI. 10.1137/S1052623402407084

1. Introduction.

This paper discusses the use of frames and grids in derivative-free optimization. The unconstrained optimization problem is examined first and analyzed using Clarke's nonsmooth calculus [3]. This is extended to linearly constrained problems by aligning the frames and grids with appropriate subsets of the linear constraints. Herein a grid is the set of points in R^n which contains a given origin point and all points which differ from this origin by an integer combination of members of a basis for R^n.

In 1997, Torczon [17] showed that many existing direct search methods conform to a common structure called generalized pattern search (GPS), which restricts attention to a sequence of interrelated meshes. A mesh is defined in the same way as a grid, except that only nonnegative integer combinations are used, and the basis is replaced by a set of vectors whose nonnegative combinations (with real coefficients) contain R^n. In [17] GPS was shown to converge under mild conditions, including continuous differentiability of the objective function. Since then a number of generalizations and modifications of GPS have been proposed. Amongst them is the work of Lewis and Torczon [12, 13] in extending GPS to bound and linearly constrained problems. A simple barrier approach is used where the objective function is declared to be infinite at any point which violates one or more constraints. The barrier approach aligns the set of interrelated meshes with the constraints. This allows the set of search steps to adequately reflect each possible cone of feasible directions. More recently, Audet and Dennis [1] have simplified the analysis in [13, 17] and extended it to nonsmooth functions by using Clarke's generalized derivatives [3].

In GPS the meshes are related because each mesh is a subset of some member of a sequence of nested grids. Coope and Price [6] have shown that for unconstrained

*Received by the editors May 7, 2002; accepted for publication (in revised form) February 25, 2003; published electronically October 2, 2003.
http://www.siam.org/journals/siopt/14-2/40708.html
†Department of Mathematics and Statistics, University of Canterbury, Private Bag 4800, Christchurch, New Zealand (c.price@math.canterbury.ac.nz, i.coope@math.canterbury.ac.nz).

optimization problems, the grids do not have to be related to one another. In [6] a step to any lower point (not necessarily a grid point) is permitted each time a new grid is selected. The orientation and shape of the new grid can be chosen independently from those of previous grids. This permits grids to be chosen to reflect information gathered during previous iterations. In contrast to GPS, the algorithm in [6] must force the grids to become arbitrarily fine. As shall be shown later there are convenient methods for doing this.

A disadvantage of the grid-based template in [6] is that steps to arbitrary lower points can occur only when there is a change of grid, and such changes can be infrequent. In [4] a frame-based method which allows steps to arbitrary lower points is described, and convergence is shown under mild conditions. This method uses a sufficient descent condition to enforce convergence. A similar set of methods is described by García-Palomares and Rodríguez [10]. These methods [10] are not explicitly formulated in terms of frames and restrict themselves to a fixed set of search directions for all iterations. Indeed, it can be shown that the implemented algorithms SDSA and NSDSA in [10] are special cases of the framework presented in [4] and also of the template presented herein. It is shown in section 4 that the prototype sequential algorithms presented in [10] also conform to a simple extension of the template presented herein. Without the extension, our work uses a single sufficient descent condition for all directions, whereas [10] uses a different condition for each direction. An explicit sufficient descent condition is used in [4], which is the only reason why the prototype sequential algorithms in [10] do not conform to the framework in [4]. The convergence results in [10] are similar to the ones in this paper and exceed those in [4] as the latter restricts attention to C^1 functions. Unlike [10], we do not consider the case when f is locally convex.

This paper looks at the use of grids and frames in unconstrained and linearly constrained derivative-free optimization. The optimization problem may be concisely expressed as

$$\min_{x\in\Omega} f(x), \quad \text{where} \quad \Omega \subseteq R^n \tag{1.1}$$

and where a local minimum is sought. The objective function f maps R^n into $R \cup \{+\infty\}$, with the convention that f is assigned the value $+\infty$ in regions where it is undefined. We also focus attention on the cases where f is locally Lipschitz, strictly differentiable [3], or C^1. The lack of second derivatives means that stationary points will be accepted as solutions in practice. The case when linear constraints are present is also examined. These constraints are used to define the feasible region Ω. We look at how grids and frames can be chosen to take into account the geometry of Ω.

This paper shows that grid-based methods can be expressed and analyzed in terms of frames, thereby unifying the treatment of grid- and frame-based methods. This unification allows many hybrid methods to be formed, including those which permit arbitrary simple descent steps at every iteration. This is achieved by formulating a frame-based template which can temporarily restrict attention to a subset of R^n called an admissible set. The concept of an admissible set is a generalization of the idea of a mesh in GPS or of a grid in [6]. A sequence of admissible sets may be used, where these sets eventually become progressively finer. For grid-based methods the grids are the admissible sets. For frame-based methods the admissible sets are equal to R^n. The introduction of admissible sets allows methods which are analogous to grid-based methods but do not use rectangular grids. The template is described in terms of sufficient descent. *For appropriate choices of admissible set the phrase*

sufficient descent means simple descent; for other choices of admissible set sufficient descent is stricter than simple descent. The template is strongly connected with GPS methods when simple descent is always used and only points in the admissible sets are considered, where the admissible sets are nested grids (or subsets thereof). This is discussed in detail in [6].

The basic strategy is to generate a sequence of iterates in Ω whose cluster points are solutions of (1.1) under appropriate conditions. The function values at these iterates form a decreasing sequence. For convenience, one point is said to be lower (or better) than another if it has a lower function value. At each iteration a search is conducted for a point which is sufficiently lower than the current iterate. When the search is unsuccessful, the current iterate is called quasi-minimal. The search for a sufficiently lower point is required to satisfy a number of conditions. Included are conditions which ensure it is a finite process and conditions which ensure the search is not declared unsuccessful until it has adequately explored the region around the current iterate. This exploration takes into account the local geometry of Ω and evaluates f at a set of points called a frame.

Frames are defined precisely in section 2, but, loosely speaking, a frame is a group of points which surround a central point called the frame center. If none of these surrounding points is significantly lower than the frame center, then the frame and the frame center are called quasi-minimal. A quasi-minimal frame center (or quasi-minimal iterate) is, in some sense, a discrete approximation to a local minimum. The frame center itself is not part of the frame.

The basic approach of a frame-based algorithm is to generate an infinite sequence of quasi-minimal frames such that the distances between points in these frames shrink to zero in the limit. The nature of the cluster points of the sequence of quasi-minimal iterates is examined using Clarke's nonsmooth analysis. In this part our approach is similar to Audet and Dennis's analysis of GPS [1].

We first examine an arbitrary unspecified algorithm that generates a sequence of iterates $\{x^{(k)}\} \in \Omega$, where this sequence of iterates contains an infinite subsequence $\{z^{(m)}\}$ of quasi-minimal frame centers. The two indices k and m count the number of iterations and quasi-minimal frames, respectively. The function $k = k(m)$ gives the number of the iteration in which the mth quasi-minimal frame occurs. At each iteration a new iterate $x^{(k+1)}$ is chosen which satisfies one of two conditions: either $x^{(k+1)}$ is an admissible point which is sufficiently lower than $x^{(k)}$ or the algorithm finds a quasi-minimal frame centered on $x^{(k+1)}$, and $x^{(k+1)}$ is not higher than $x^{(k)}$. In the latter case, this new frame center $x^{(k+1)}$ may be anywhere in Ω, but the quasi-minimal frame may be required to consist of points which lie in the current admissible set. Various strategies are used to ensure a quasi-minimal frame is located in a finite time. It is shown that this guarantees the sequence $\{z^{(m)}\}$ is infinite. It is then shown that the Clarke generalized derivative at each cluster point of $\{z^{(m)}\}$ in each limiting direction is nonnegative. In the case when the objective function is C^1 and $\Omega = R^n$, it is shown that all such cluster points are stationary points of f. These results are extended to linearly constrained optimization problems by using a barrier approach [1, 13] and choosing each frame to span the relevant tangent cone.

In section 3 the algorithm template is described, and its behavior is analyzed in sections 4 and 5. Section 4 develops the main convergence results and applies them to the unconstrained optimization problem. Section 5 addresses the linearly constrained optimization problem. It describes how frames can be constructed which take into account the linear constraints and presents the convergence results for methods using

such frames. Section 6 looks at how the frames' sizes may be chosen, and concluding remarks are made in section 7.

The template (Template D) described herein is opportunistic, as are framework A in [6] and the framework presented in [4]. This means it can abandon a partially completed frame immediately after discovering a point of sufficient descent. The price paid for this opportunism is that the convergence theory applies only to the subsequence of quasi-minimal iterates $\{z^{(m)}\}$. A nonopportunistic approach is presented in framework B of [6] and template C of [14]. These templates require each frame to be completed and to search along the ray from the frame's center through a point not higher than the lowest frame point. The advantage of this is that the convergence theory applies to the *whole sequence of iterates* $\{x^{(k)}\}$, not merely $\{z^{(m)}\}$. The restriction that each frame must be completed is not serious for certain types of algorithms. For instance, methods using finite differences [7] or polytopes [15] must come within one point of completing a frame in order to construct the gradient estimate or polytope.

2. Positive bases and frames. A frame is a finite set of points which strictly contains another point (the frame's center) in its convex hull. The directions from the frame's center to each point in the frame form a positive basis [9], which is a set of vectors $\mathcal{V}_+ = \{v_i\}$ such that

B1: every vector in R^n can be written as a nonnegative combination of the vectors in $\mathcal{V}_+$ and

B2: no proper subset of $\mathcal{V}_+$ satisfies B1.

The term "nonnegative combination" means a (finite) linear combination without negative coefficients. Sets of vectors which satisfy property B1 only are called positive spanning sets. Any positive spanning set not satisfying B2 must contain a positive basis as a proper subset. It is also shown in [9] that any positive basis for R^n must satisfy $n+1 \le |\mathcal{V}_+| \le 2n$. The members of each positive basis $\mathcal{V}_+$ which is constructed are assigned a specific order, and from now on each positive basis is assumed to be ordered unless stated otherwise.

A frame Φ is the set of points

$$\Phi = \Phi(z, h, \mathcal{V}_+) = \{z + hv : v \in \mathcal{V}_+\}, \tag{2.1}$$

where z is the *frame center* and the positive scalar h is the *frame size*.

A frame Φ is called minimal if and only if

$$f(z) \le f(x) \quad \forall x \in \Phi(z, h, \mathcal{V}_+).$$

It is useful to work with frames which are only "nearly" minimal. Such frames are called quasi-minimal and are easier to generate than minimal frames. The generation of quasi-minimal (or minimal) frames is important for two reasons: the convergence theory applies to the sequence of centers of quasi-minimal frames, and some algorithm parameters can be altered only after a quasi-minimal frame has been found. A frame Φ is called ϵ_z–quasi-minimal if and only if

$$f(z) \le f(x) + \epsilon_z \quad \forall x \in \Phi(z, h, \mathcal{V}_+) \tag{2.2}$$

for a preselected nonnegative ϵ_z. The notation $\Phi^{(m)} = \Phi(z^{(m)}, h_z^{(m)}, \mathcal{V}_+^{(m)})$ is used to denote the mth quasi-minimal frame.

Each quasi-minimal frame may have a different value $\epsilon_z^{(m)}$ for ϵ_z. Hence, when a frame $\Phi^{(m)}$ is called "quasi-minimal" this is understood to mean $\epsilon_z^{(m)}$–quasi-minimal.

It is necessary to have values for ϵ and h at every iteration, and so new sequences $\{\epsilon^{(k)}\}$ and $\{h^{(k)}\}$ are introduced. The sequences $\{h^{(k)}\}$ and $\{h_z^{(m)}\}$ are linked by the relation $h_z^{(m)} = h^{(k(m))}$. In other words, $h_z^{(m)}$ is the value $h^{(k)}$ takes in the iteration $k(m)$ in which the mth frame is located. The quantities $\epsilon_z^{(m)}$ and $\epsilon^{(k)}$ are similarly related.

The sequence of ϵ values is required to satisfy the following condition:

$$\lim_{k\to\infty} \frac{\epsilon^{(k)}}{h^{(k)}} = 0. \tag{2.3}$$

A requirement of the convergence theory is that $h^{(k)} \to 0$ as $k \to \infty$, and so one simple choice that satisfies (2.3) is $\epsilon = Nh^\nu$, with $\nu > 1$ and $N \geq 0$. In any case (2.3) requires that $\{\epsilon^{(k)}\}$ goes to zero faster than $\{h^{(k)}\}$.

One could easily define frames using positive spanning sets rather than positive bases. However, there are a number of advantages to the latter (see, e.g., [4, 14]). For convergence purposes a number of restrictions must be imposed on the set $\mathcal{V}_+$ used to define a frame, and this is more easily done if $\mathcal{V}_+$ is a positive basis rather than a positive spanning set. Second, frame-based templates permit a finite number of arbitrary points (not included in $\mathcal{V}_+$) to be examined during each iteration. Including such points in $\mathcal{V}_+$ subjects them to unnecessary restrictions. In practice these extra points may be used in a similar way to the members of $\mathcal{V}_+$, but for theoretical purposes they are best kept separate.

An upper bound K is imposed on the length of each member of each $\mathcal{V}_+^{(m)}$,

$$\|v\| \leq K \quad \forall m \quad \text{and} \quad \forall v \in \mathcal{V}_+^{(m)}, \tag{2.4}$$

where K is independent of m and k.

A set $\mathcal{V}_+^{(\infty)} = \{v_1^{(\infty)}, \ldots, v_p^{(\infty)}\}$ is a limit of the sequence of ordered positive bases $\{\mathcal{V}_+^{(m)}\}_{m=1}^{\infty}$ if and only if an infinite subsequence of $\{\mathcal{V}_+^{(m)}\}$ exists such that each positive basis belonging to this subsequence has cardinality p, and

$$\lim_{m\to\infty} v_i^{(m)} = v_i^{(\infty)} \quad \forall i = 1, \ldots, p, \tag{2.5}$$

where the limit is understood to be taken over this subsequence. Condition (2.4) ensures that such limits exist. The following assumption is needed.

Assumption 2.1. All members of the sequence $\{\mathcal{V}_+^{(m)}\}$ satisfy (2.4), and each limit $\mathcal{V}_+^{(\infty)}$ of the sequence $\{\mathcal{V}_+^{(m)}\}$ is an ordered positive basis.

This assumption may be enforced in a variety of ways, some of which are discussed in [4, 14].

3. The algorithm template. The template consists of two nested loops. The outer loop (steps 2–6, indexed by m) generates a sequence of quasi-minimal frames with the desired properties. The purpose of the inner loop (steps 3–5, indexed by k) is to generate a quasi-minimal frame. Iterations of the inner loop are performed until a quasi-minimal frame is found, where quasi minimality is defined in (2.2) by $\epsilon_z^{(m)}$. Each iteration of the inner loop which does not find a quasi-minimal frame obtains a point of sufficient descent instead. Fixing certain quantities during each iteration of the outer loop (and hence each execution of the inner loop) ensures that a quasi-minimal frame must be located in a finite number of inner loop iterations under standard assumptions. In particular, it is assumed that the sequences of function

values $\{f^{(k)}\}$ and iterates $\{x^{(k)}\}$ remain bounded. Here the notation $f^{(k)} = f(x^{(k)})$ has been used.

The purpose of the outer loop is to generate a sequence of quasi-minimal frames with the desired properties. In particular, this sequence of quasi-minimal frames must be infinite. In other words, each iteration of the outer loop must be a finite process. Termination of the mth iteration of the outer loop can be guaranteed either by choosing ϵ to be bounded away from zero or by restricting points of sufficient descent to an admissible set $\mathcal{G}^{(m)}$. In the former case, ϵ is given a strictly positive lower bound $E^{(m)}$, and $E^{(m)}$ is kept constant between quasi-minimal frames. Sufficient descent means that $f(x^{(k)})$ is reduced by more than $E^{(m)}$ at each iteration of the inner loop. The mth iteration of the outer loop can fail to terminate only if sufficient descent is always obtained. This means that $f^{(k)} \to -\infty$ as k goes to infinity. In the latter case, $\epsilon = 0$ is permitted, but $\mathcal{G}^{(m)}$ must contain only a finite number of points in any bounded subset of R^n, amongst other things. Hence the inner loop cannot generate a bounded infinite sequence of iterates with strictly decreasing function values. A new $\mathcal{G}^{(m)}$ can be chosen after each quasi-minimal frame, and so $\mathcal{G}^{(m)}$ denotes the admissible set used during the search for the mth quasi-minimal frame. When $E^{(m)} > 0$ we define $\mathcal{G}^{(m)} = R^n$ for completeness.

At each iteration of the inner loop f is calculated at a finite number of points. An iteration is completed when either sufficient descent is obtained or a quasi-minimal frame is located. Here sufficient descent means reducing f by more than ϵ, where ϵ is the same constant used to define quasi minimality. At each iteration the algorithm may calculate f at a finite number of points. If neither sufficient descent nor a quasi-minimal frame has been obtained, then the algorithm begins forming a frame in $\mathcal{G}^{(m)}$ about a frame center x, where x is not higher than the previous iterate $x^{(k-1)}$. The frame either is quasi-minimal or contains a point in $\mathcal{G}^{(m)}$ more than ϵ lower than $x^{(k-1)}$. This completes an iteration of the inner loop. If sufficient descent was obtained, then the algorithm increments k and starts a new iteration of the inner loop. Otherwise, the inner loop terminates.

During each iteration of the outer loop a positive bound $H^{(m)}$ on $h^{(k)}$ is imposed. Theoretically this bound is superfluous, but its presence highlights the existence of a lower bound on $h^{(k)}$ implicit in the choice of $\mathcal{G}^{(m)}$ (if $E^{(m)} = 0$) or $E^{(m)}$ (otherwise). Further remarks on this are made later in this section and section 6, respectively.

Algorithm Template D.

1. Initialize: set $k = 1$, $m = 1$, and choose the initial point $x^{(0)} \in \Omega$.
2. Choose $H^{(m)} > 0$, $E^{(m)} \geq 0$, and $\mathcal{G}^{(m)}$.
3. Choose $h^{(k)} \geq H^{(m)}$ and $\epsilon^{(k)} \geq E^{(m)}$.
4. Execute any finite process which satisfies one of these conditions:
   (a) generates an iterate $x^{(k)} \in \Omega \cap \mathcal{G}^{(m)}$ satisfying $f\left(x^{(k)}\right) < f^{(k-1)} - \epsilon^{(k)}$; or
   (b) generates a quasi-minimal frame $\Phi^{(m)} = \Phi(z^{(m)}, h_{\mathrm{z}}^{(m)}, \mathcal{V}_{+}^{(m)})$, where $x^{(k)} \in \Omega$ and $f^{(k)} \leq f^{(k-1)}$. Here $z^{(m)} = x^{(k)}$, $h_{\mathrm{z}}^{(m)} = h^{(k)}$, and $\epsilon_{\mathrm{z}}^{(m)} = \epsilon^{(k)}$; or
   (c) case (b) of this step with the added restriction $\Phi^{(m)} \subset \mathcal{G}^{(m)}$.
5. If $x^{(k)}$ is not quasi-minimal, increment k and go to step 3.
6. Increment m and k. If stopping conditions are not satisfied, go to step 2.

Condition (c) is included in step 4 to highlight the fact that an attempt to satisfy condition (c) by forming a frame in $\mathcal{G}^{(m)}$ guarantees the satisfaction of either (a) or (c). In contrast, attempts to satisfy either (a) or (b) may end in failure without

satisfying any of (a)–(c). An example of how condition (c) is used is presented later in Figure 3.1.

The arbitrary process in step 4 allows f to be evaluated at points anywhere in Ω. These points can be used, for example, to include a quasi-Newton step, points chosen by an heuristic, or even randomly selected points. This arbitrary process also permits the lowest point from a previous quasi-minimal frame to be included in the current iteration. This is useful because a quasi-minimal frame which is not minimal contains at least one point which is lower than the frame's center. Inclusion of such points allows movement away from a strictly concave maximum. For example, if $f = -x^2$ in R^1, with $x^{(0)} = 0$, $h^{(0)} = 1$, $\mathcal{V}_+ = \{1, -1\}$, and $\epsilon = 2h^2$, then $x^{(0)}$ is a quasi-minimal frame center for all positive h. However, if the frame points from the first iteration are included in the arbitrary finite process of the next iteration, then on the second iteration an algorithm will step to a point $x^{(2)}$ satisfying $f^{(2)} \leq -1$ and escape the local maximum. Otherwise, an algorithm might generate an infinite sequence of quasi-minimal frames centered on the origin.

The points examined in step 4's arbitrary process are useful in the analysis of the template, and so we define $\mathcal{S}_+^{(m)}$ as the set containing all nonzero vectors v satisfying (2.4) such that $f(z^{(m)} + h_{\mathrm{z}}^{(m)} v) \geq f(z^{(m)}) - \epsilon_{\mathrm{z}}^{(m)}$ is established by the arbitrary process in step 4. That is to say,

$$\mathcal{S}_+^{(m)} = \left\{ v : 0 < \|v\| \leq K \text{ and } f\left(z^{(m)} + h_{\mathrm{z}}^{(m)} v\right) \geq f\left(z^{(m)}\right) - \epsilon_{\mathrm{z}}^{(m)} \text{ is shown in step 4} \right\}. \tag{3.1}$$

The set $\mathcal{G}^{(m)}$ may be a grid [6], R^n, or otherwise. For example [6], $\mathcal{G}^{(m)}$ may be a grid centered on $z^{(m-1)}$ and containing all points differing from $z^{(m-1)}$ by a sum of integer multiples of the vectors $v_1^{(m)}, \ldots, v_n^{(m)}$, where $v_1^{(m)}, \ldots, v_n^{(m)}$ form a basis for R^n. Many other possibilities also exist. The choice of $\mathcal{G}^{(m)}$ is subject to a number of restrictions when $E^{(m)}$ is zero. When $E^{(m)} \neq 0$ we use $\mathcal{G}^{(m)} = R^n$ without loss of generality.

Assumption 3.1. If $E^{(m)} = 0$, then the following two conditions hold:

G1: $\mathcal{G}^{(m)}$ contains only a finite number of points in any bounded subset of Ω; and

G2: for all $z \in \Omega$ there exists at least one frame $\Phi(z, h_{\mathrm{z}}, \mathcal{V}_+)$ in $\mathcal{G}^{(m)}$ for which h_{z} and $\mathcal{V}_+$ satisfy all restrictions required by the template (including Assumption 2.1 or, in the constrained case, Assumption 5.3).

Condition G2 in Assumption 3.1 is used to ensure that an algorithm can always find a frame centered on any point it chooses. Condition G2 excludes such sets as the set of positive integers in R^1 because this set does not contain a frame about $x = 0$. Condition G2 conspires with inequality (2.4) to impose a lower limit on h. For example, if $\mathcal{G}^{(m)}$ is the grid of all integer points in R^n, and $K = 5$ in (2.4), then h must be at least $1/5$ in order for condition G2 to be satisfied.

When $E^{(m)} = 0$ points of sufficient descent must be chosen from $\mathcal{G}^{(m)}$, but quasi-minimal iterates are not required to belong to $\mathcal{G}^{(m)}$. This permits an algorithm to consider points not in $\mathcal{G}^{(m)}$ *at every iteration* via the following process. Let step 4 start with an iterate $x^{(k-1)}$ in $\mathcal{G}^{(m)}$. The arbitrary finite process in this step selects a point $x \in \Omega$ which is not higher than $x^{(k-1)}$. Note that $x \in \mathcal{G}^{(m)}$ is not required. If the algorithm subsequently locates a quasi-minimal frame around x, then condition (b) has been achieved, which completes step 4. Otherwise, termination of step 4 is *forced* by trying to achieve condition (c): that is to say, the algorithm forms a frame in $\mathcal{G}^{(m)}$ with x as the frame's center. Either this frame is quasi-minimal (condition (c)

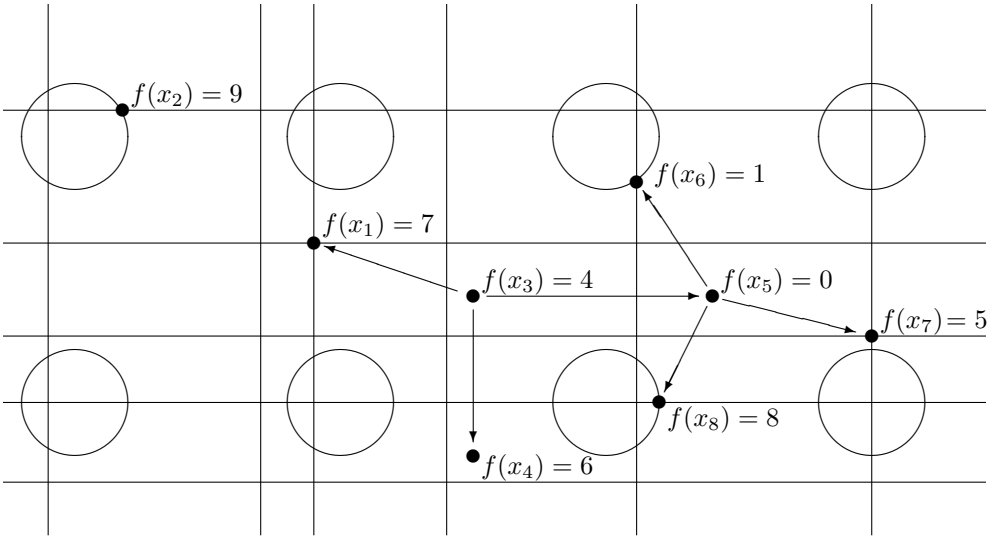


FIG. 3.1. *An illustration of how step* 4 *works. All intersections of two lines, or of a line and a circle, are admissible points. Black dots are points considered by step* 4*, and arrows point from each frame center to the corresponding frame points.*

is satisfied) or a point in $\mathcal{G}^{(m)}$ which is sufficiently lower than $x^{(k-1)}$ is located (which satisfies condition (a)). Either way step 4 then terminates. Condition (c) is actually superfluous; if condition (c) holds, then condition (b) is automatically satisfied.

An illustration of step 4 is given in Figure 3.1. Here $E^{(m)} = \epsilon^{(k)} = 0$, and $\mathcal{G}^{(m)}$ is the set of all points which are intersections of either two lines or of a line and a circle. The lines form two irregularly spaced parallel sets. The circles are centered on integer points in R^2, and all have the same radius. Clearly $\mathcal{G}^{(m)}$ satisfies Assumption 3.1, provided $h^{(k)}$ is not too small. Points considered by step 4 are marked with dots, and also the legend of the form $f(x_i) = F$. Here points are used in the order given by the index i, and F is the function value of the ith such point. The index i is not an iteration number; inside step 4 both k and m are fixed. The notations x_i and $f_i = f(x_i)$ are restricted to this paragraph and Figure 3.1. Arrows point from each frame center to the points in the corresponding frame. Step 4 begins with the current iterate x_1. It calculates f_2. Now x_2 is an admissible point, so if x_2 were also lower than x_1, then step 4 would terminate under condition (a) and return x_2 as a point of sufficient descent. However, x_2 is higher than x_1 and is thus rejected. Step 4 then calculates f at x_3, which is lower than x_1. Now x_3 is not admissible, so step 4 cannot return x_3 as a point of sufficient descent. Instead step 4 forms a frame around x_3, consisting of x_1, x_4, and x_5. If this frame were quasi-minimal (which is the same as minimal since $\epsilon^{(k)} = 0$), then step 4 would terminate under condition (b) and return x_3 as a quasi-minimal iterate. However, the frame is not quasi-minimal because x_5 is lower than x_3. Unfortunately, x_5 is not admissible, and so it cannot be returned as a point of sufficient descent. Step 4 then forms a frame around x_5 consisting only of admissible points: x_6, x_7, and x_8. This *forces* the termination of step 4: either at least one of x_6, x_7, and x_8 is lower than x_5 (and hence lower than x_1) or all three are at least as high as x_5. In the former case step 4 would terminate under condition (a) and return the lowest of x_6, x_7, and x_8. In the latter case step 4 would terminate under condition (c) and return x_5 as a quasi-minimal center (which is what happens).

Hence x_5 becomes both the new $x^{(k)}$ and the new $z^{(m)}$. The set $\mathcal{S}_+^{(m)}$ consists of the vectors $(x_i - x_5)/h_{\mathrm{z}}^{(m)}$, $i \neq 5$, which satisfy the inequalities in (3.1).

Template D uses the same set of admissible points $\mathcal{G}^{(m)}$ for each iteration between quasi-minimal frames. At each iteration attention could be restricted to a subset of $\mathcal{G}^{(m)}$ which satisfies conditions G1 and G2, and this would not affect the convergence results. This has not been done for two reasons. First, the current form of Template D and the resulting analysis is clearer. Second, by judicious choice of $h_{\mathrm{z}}^{(m)}$ and $\mathcal{V}_+^{(m)}$ restricting attention to a subset of $\mathcal{G}^{(m)}$ can be achieved implicitly.

The convergence analysis examines the asymptotic properties of the sequences of iterates when the stopping conditions are never invoked. Practical considerations make stopping conditions essential, which is why they are featured in Template D. The current placement of stopping conditions ensures that the algorithm always terminates with a quasi-minimal frame. Stopping conditions could also be checked in the inner loop, for example at step 5.

4. The main convergence results. First it is shown that the subsequence of quasi-minimal frames is infinite under appropriate conditions.

THEOREM 4.1. *Assume that for each m either $E^{(m)} > 0$ or $\mathcal{G}^{(m)}$ satisfies conditions* G1 *and* G2 *in Assumption* 3.1. *Then at least one of the three following possibilities holds:*

(i) *the subsequence of quasi-minimal iterates is infinite; or*
(ii) *the sequence of iterates is unbounded; or*
(iii) $f^{(k)} \to -\infty$ *as* $k \to \infty$.

Proof. We assume case (i) does not occur and that J is the final value of m. In the case when $E^{(J)} = 0$ it is then shown that (ii) must occur. Similarly, when $E^{(J)}$ is strictly positive it is shown that (iii) must occur.

If $E^{(J)} = 0$, then step 4 generates a sequence of points in $\mathcal{G}^{(J)} \cap \Omega$ with strictly decreasing function values. This sequence must contain an infinite number of distinct points in $\mathcal{G}^{(J)} \cap \Omega$. However, $\mathcal{G}^{(J)} \cap \Omega$ can contain only a finite number of points inside any bounded subset of Ω, by condition G1. Hence the sequence of iterates must be unbounded.

Let $E^{(J)}$ be strictly positive. Once $m = J$ occurs, step 4 is executed endlessly, and it reduces the best known function value by more than $E^{(J)}$ each time it is executed. Hence $f^{(k)} \to -\infty$ in the limit $k \to \infty$, as required. ∎

In addition to conditions on the sequences of ordered positive bases and admissible sets, the following assumption is needed to establish convergence.

Assumption 4.2. The following conditions hold:

(a) the points at which f is calculated lie in a compact subset of R^n;
(b) the sequence of function values $\{f^{(k)}\}$ is bounded below;
(c) $h^{(k)} \to 0$ as $k \to \infty$; and
(d) $\epsilon^{(k)}/h^{(k)} \to 0$ as $k \to \infty$.

The first two parts of this assumption eliminate possibilities (ii) and (iii) of Theorem 4.1, which guarantees that the sequence $\{z^{(m)}\}$ has cluster points. Parts (c) and (d) ensure that these cluster points have interesting properties. Satisfaction of these latter two parts can be ensured by an appropriate implementation of the template. Collectively parts (c) and (d) ensure $\epsilon^{(k)} \to 0$ as $k \to \infty$.

The next theorem establishes the basic convergence result using Clarke's generalized derivative [3], which is

$$f^\circ(x;v) = \limsup_{h\downarrow 0\ y\to x} \frac{f(y+hv)-f(y)}{h}.$$

Provided f is locally Lipschitz at x it can be shown [3] that $f^\circ(x;v)$ is subadditive and positively homogeneous in v. Moreover, if M is a Lipschitz constant for f at x, then $|f^\circ(x;v)| \le M\|v\|$.

THEOREM 4.3. *Let f be locally Lipschitz at $z^{(\infty)}$. Let v be any vector such that there exists a sequence $\{(z^{(m)}, v^{(m)})\}$ with $v^{(m)} \in \mathcal{S}_+^{(m)}$ for all m and such that $(z^{(\infty)}, v)$ is a cluster point of this sequence, where $\mathcal{S}_+^{(m)}$ is defined in* (3.1). *Then*

$$f^\circ\left(z^{(\infty)}; v\right) \ge 0.$$

Proof. We restrict our attention to a subsequence of $\{z^{(m)}\}$ for which the corresponding subsequence $\{(z^{(m)}, v^{(m)})\}$ converges uniquely to $(z^{(\infty)}, v)$. The definition (3.1) implies that

$$f\left(z^{(m)} + h_{\mathrm{z}}^{(m)} v^{(m)}\right) - f\left(z^{(m)}\right) + \epsilon_{\mathrm{z}}^{(m)} \ge 0.$$

Hence

$$\limsup_{m\to\infty} \frac{f\left(z^{(m)}+h_{\mathrm{z}}^{(m)}\left(w^{(m)}+v\right)\right) - f\left(z^{(m)}+h_{\mathrm{z}}^{(m)}w^{(m)}\right) + f\left(z^{(m)}+h_{\mathrm{z}}^{(m)}w^{(m)}\right) - f(z^{(m)})}{h_{\mathrm{z}}^{(m)}} \ge 0,$$

where $w^{(m)} = v^{(m)} - v$. Now $w^{(m)} \to 0$ as $m \to \infty$, and so the first two terms provide a lower bound on f°, which yields

$$f^\circ\left(z^{(\infty)}; v\right) + \limsup_{m\to\infty} \frac{f\left(z^{(m)} + h_{\mathrm{z}}^{(m)} w^{(m)}\right) - f(z^{(m)})}{h_{\mathrm{z}}^{(m)}} \ge 0.$$

The last term vanishes because f is locally Lipschitz and because $w^{(m)} \to 0$, which yields the required result. ∎

An alternative way of looking at Theorem 4.3 is as follows.

COROLLARY 4.4. *There does not exist an open halfspace on which $f^\circ(z^{(\infty)}; v)$ is negative for all v in this halfspace.*

Proof. The template guarantees $\mathcal{V}_+^{(m)} \subseteq \mathcal{S}_+^{(m)}$ for all m. Assumption 2.1 and Theorem 4.3 imply there exists a positive basis $\mathcal{V}_+^{(\infty)}$ such that

$$f^\circ\left(z^{(\infty)}; v\right) \ge 0 \quad \forall v \in \mathcal{V}_+^{(\infty)}.$$

Now every open halfspace contains a member of $\mathcal{V}_+^{(\infty)}$. Hence no open halfspace exists on which the generalized derivative of f at $z^{(\infty)}$ is negative. ∎

4.1. The differentiable case. Corollary 4.4 is useful because all C^1 functions have open halfspaces of descent directions at all nonstationary points. We now look at the case when f is strictly differentiable [3] at $z^{(\infty)}$, i.e.,

$$\exists w \in R^n \quad \text{such that} \quad f^\circ(z^{(\infty)}; v) = w^T v \quad \forall\, v \in R^n.$$

This yields the following important corollary.

COROLLARY 4.5. *If f is strictly differentiable at $z^{(\infty)}$, then $z^{(\infty)}$ is a stationary point of f.*

Proof. Strict differentiability implies

$$\exists w \in R^n \text{ such that } f^\circ(z^{(\infty)}; v) = w^T v \quad \forall\, v \in R^n.$$

If w is nonzero, then $f^\circ(z^{(\infty)}; v)$ is negative on the open halfspace $\{v : w^T v < 0\}$, contradicting Corollary 4.4. The only remaining possibility is that $w = 0$. Hence $z^{(\infty)}$ must be a stationary point of f. □

The difference between these two corollaries is that Corollary 4.4 can eliminate many points of nondifferentiability from the set of possible cluster points of $\{z^{(m)}\}$, whereas Corollary 4.5 cannot. For example, let $f = \min\{\|x\|, \|x\|^2\}$, where the 2-norm has been used. Corollary 4.5 has $x = 0$ and $\{x : \|x\| = 1\}$ as possible cluster points, whereas Corollary 4.4 shows that $x = 0$ is the only possible cluster point.

Clearly if f is continuously differentiable at z, then it is also strictly differentiable there, and so Corollary 4.5 establishes the convergence results of [4] and framework A of [6]. It also establishes convergence for methods which do not conform to either [4] or [6]. An example of such a method is any algorithm which uses $E = \epsilon = 0$ and also uses frame centers which are not necessarily members of the current admissible set $\mathcal{G}^{(m)}$. Further examples include any method using $E = \epsilon = 0$ and a grid with hexagonal, triangular, or circular symmetries in some dimensions. An example of an admissible set with both circular and rectangular symmetries is the set of all points in R^2 which have either integer Cartesian coordinates (x_1 and x_2) or have integer values for r and $r\theta/\pi$, where r and θ are the standard polar coordinates. An admissible set like this could be used with functions that may have both straight grooves and circular grooves centered on the origin. Many other possibilities for the admissible set exist, including those which incorporate random elements. For example, $\mathcal{G}^{(m)}$ could be the set of all points $x + v(x) \in R^n$, where all components of x are integer and where $v(x)$ is a random vector function of x over the set of vectors satisfying $\|v(x)\| \le 1$.

4.2. The Lipschitz condition. In this subsection the case when f is locally Lipschitz but not differentiable is discussed. Let f be locally Lipschitz with Lipschitz constant M at $z^{(\infty)}$, and let v be a direction satisfying $f^\circ(z^{(\infty)}; v) < 0$ at $z^{(\infty)}$. Let u be a unit vector, and let $\eta \in R$ be positive. Then

$$f^\circ\left(z^{(\infty)}; v + \eta u\right) \le f^\circ\left(z^{(\infty)}; v\right) + \eta f^\circ\left(z^{(\infty)}; u\right) \le f^\circ\left(z^{(\infty)}; v\right) + \eta M.$$

This shows that $f^\circ(z^{(\infty)}; .)$ is negative for all directions in a cone containing v in its interior. Hence an algorithm conforming to the template will eventually find a descent direction if it looks along a sequence of directions converging to v as z goes to $z^{(\infty)}$.

It should be noted that the existence of descent directions at a point does not guarantee that f° is negative along these directions. A very simple example is the function $f = -|x|$ in one dimension at the origin. A more interesting example in two dimensions is

$$f = \begin{cases} r, & |\theta| \ge \theta_0, \\ r\,(2|\theta| - \theta_0)\,/\theta_0, & |\theta| < \theta_0, \end{cases}$$

again at the origin. For clarity, this example is described using polar coordinates r and θ, with $r \ge 0$ and $-\pi < \theta \le \pi$. The function is well defined for all θ_0 values, but we are primarily interested in $0 < \theta_0 < \pi$. For these values f looks like an

upward pointing cone with a notch slanting downwards along $\theta = 0$. The example would be presented to an algorithm as an unconstrained problem in the rectangular coordinates $x_1 = r\cos(\theta)$ and $x_2 = r\sin(\theta)$. Simple calculations show that f° is positive for every direction whenever $\theta_0 < \pi/2$. However, directions with $2|\theta| < \theta_0$ are descent directions at the origin.

The necessity of the Lipschitz condition can be seen by considering, for example, the function $f = -x_2 + 5\sqrt{|x_1|}$. Elementary calculations show the directional derivative $f'(0; e_2) = -1$, where e_i is the ith unit vector. However, if the direction e_2 is replaced by the parabolic arc $tv(t)$, where $v(t) = e_2 + te_1$, then

$$f'_{\text{arc}}(0; v(\cdot)) = \lim_{t\to 0} \frac{f(0 + tv(t)) - f(0)}{t} = 4.$$

Here the fact that the direction of $v(t)$ alters as t goes to zero means that a descent step is not located even though $v(t)$ becomes parallel to the descent direction e_2 as t tends to zero. There is nothing special about keeping the direction constant. Similar calculations with the function $f = -x_2 + 5\sqrt{|x_1 - x_2^2|}$ give $f'_{\text{arc}}(0; v(\cdot)) = -1$ and $f'(0; e_2) = 4$. This time the fixed direction fails, and by curving $v(t)$ into the limiting direction e_2, a descent step is found. So if f is not locally Lipschitz and lacks any other special properties, then little can be said.

There is one computationally expensive way to attack such problems using the arbitrary finite process in step 4 of the template. The idea is eventually to look everywhere in some neighborhood of each cluster point of the sequence of iterates. Let $x^{(k)} + y$, $y \in Y^{(k)}$, be the set of points at which f is calculated in the arbitrary finite process in step 4 during iteration k.

THEOREM 4.6. *If f is continuous and the sequence of sets $\{Y^{(k)}\}$ satisfies the following two properties,*

- Y1: *the sequence is eventually nested, i.e., $Y^{(k)} \subseteq Y^{(k+1)}$ for all k sufficiently large; and*
- Y2: *there exists a positive constant μ such that $\cup_{k=1}^{\infty} Y^{(k)}$ is dense in the open ball of radius μ centered on the origin,*

then all cluster points of the sequence of iterates are local minimizers of f.

Proof. The proof is by contradiction. Let $x^{(\infty)}$ be a cluster point of the sequence of iterates which is not a local minimizer. Replace the sequence of iterates $\{x^{(k)}\}$ with an infinite subsequence of itself such that all members of this subsequence are within $\mu/3$ of $x^{(\infty)}$, and replace $\{Y^{(k)}\}$ with the corresponding subsequence of itself. We note that this subsequence of $\{Y^{(k)}\}$ satisfies both Y1 and Y2. Property Y1 ensures that property Y2 is not lost when moving to this subsequence. Now there exists a point x_μ within $\mu/3$ of $x^{(\infty)}$ which is strictly lower than $x^{(\infty)}$. Continuity of f means that there is a ball of strictly positive radius $\xi < \mu/3$ around x_μ on which f is strictly less than $f(x^{(\infty)})$. Property Y2 means that for some finite k step 4 will evaluate f at a point in the ball of radius ξ about x_μ. This contradicts the fact that the sequence of function values $\{f(x^{(k)})\}$ is monotonically decreasing. ∎

This is not a particularly practical way of ensuring convergence except on very small problems. However, it is one way of gaining some confidence in a solution when f is not smooth.

4.3. Generalizing sufficient descent. The sequential algorithms of García-Palomares and Rodríguez [10] conform to Template D except on one point: the choice of sufficient descent condition. Herein the same measure of sufficient descent (i.e., ϵ) is used for all search steps, whereas the prototype sequential algorithms in [10] use

a different value for each search direction. Template D is easily adapted to include these prototype algorithms. This is done by replacing the sequence of constants $\{\epsilon^{(k)}\}$ with a sequence of functions $\{\epsilon^{(k)}(v)\}$. The sufficient descent condition for a step $h^{(k)}v$ from the iterate $x^{(k)}$ becomes

$$f\left(x^{(k)} + h^{(k)}v\right) < f\left(x^{(k)}\right) + \epsilon^{(k)}(v).$$

A frame Φ which contains no point of sufficient descent is quasi-minimal. The sequence $\{\epsilon^{(k)}(v)\}$ is required to have the following properties:

$$\lim_{k\to\infty} \left(\sup_{v\in R^n} \epsilon^{(k)}(v)\right) \Big/ h^{(k)} = 0 \tag{4.1}$$

and

$$\epsilon^{(k)}(v) \geq E^{(m(k))} \quad \forall v \in R^n \quad \text{and} \quad \forall k.$$

Here $m(k)$ is the value of m at step 3 of iteration k, which is the index of the quasi-minimal frame the template is searching for at iteration k. A corresponding sequence of functions $\{\epsilon_z^{(m)}(v)\}$ is also defined, with $\epsilon_z^{(m)}(v) = \epsilon^{(k(m))}(v)$ for all v, as before. Equation (4.1) ensures that the proof of Theorem 4.3 is still valid, and the rest of the convergence theory depends only on the lower bounds $E^{(m)}$, not on ϵ itself.

5. The linearly constrained case. Following [1, 13, 18] we develop a theory for the linearly constrained optimization problem (LCOP)

$$\min_{x\in\Omega} f(x), \text{ where } \Omega = \{x : a_i^T x + b_i = 0 \ \ \forall i = 1, \ldots, q \ \text{ and } \ a_i^T x + b_i \leq 0 \ \ \forall i = q+1, \ldots, L\}. \tag{5.1}$$

We regard any point $x \in \Omega$ as a solution of (5.1) if and only if no feasible direction exists at x along which the directional derivative of f is negative. The constraints defining Ω are imposed via a barrier function. A new objective function $f_c(x) = f(x) + \psi(x)$ is defined, where $\psi(x)$ is the indicator function for the set Ω. Hence $f_c(x) = f(x)$ if $x \in \Omega$, and $f_c(x) = \infty$ otherwise. Algorithms conforming to Template D may be applied to f_c; however, the discontinuous nature of f_c means that Theorem 4.3 does not guarantee convergence to one or more solutions of (5.1). To ensure that an algorithm locates solution(s) of the LCOP we consider a specialization of Template D which requires that each positive basis $\mathcal{V}_+^{(m)}$ conforms to the shape of the feasible region Ω near $z^{(m)}$. This specialization is presented later as Template E.

For the unconstrained case the crucial feature of a positive basis is that at any point x it positively spans the set of feasible directions at x and also at any point near x. For constrained problems we need finite sets of directions with the same property, although, in general, the set of feasible directions is now a closed polyhedral cone rather than R^n. The set of feasible directions can also vary from point to point, in contrast to the unconstrained case.

Template E generates a sequence of feasible iterates which contains an infinite subsequence $\{z^{(m)}\}$ of quasi-minimal frame centers. At each frame center the constraints which could be active (i.e., hold with equality) at or near this frame center are identified. The directions in that frame's positive basis are aligned with the identified set of constraints. More precisely, for any cone of feasible directions defined by a subset of those constraints, there is a subset of the frame's positive basis which

positively spans that cone of feasible directions. These *aligned* positive bases can be used to extend the convergence theory in section 4 to the linearly constrained case.

For each frame a subset of the constraints is selected which includes those constraints which are active at or near the quasi-minimal center $z^{(m)}$. This is done by choosing a positive constant δ and selecting all constraints with residuals not more than δ. These constraints are indexed by the working set $W^{(m)}$ which must satisfy

$$\left|a_i^T z^{(m)} + b_i\right| \leq \delta \Longrightarrow i \in W^{(m)}, \tag{5.2}$$

where $\delta > 0$ is independent of m. The feasibility of each $z^{(m)}$ means that every constraint which is active (which includes all equality constraints) at some point near $z^{(m)}$ appears in $W^{(m)}$. Hence, for any $x \in \Omega$ near $z^{(m)}$, the set of active constraints at x is contained in $W^{(m)}$. The positive basis $\mathcal{V}_+^{(m)}$ is then constructed so that some subset of it positively spans the cone of feasible directions at x. In practice $W^{(m)}$ would often contain constraints with residuals much greater than δ. This would assist an algorithm in traversing the boundary of the feasible region more quickly.

The constraints in $W^{(m)}$ define a polyhedral cone

$$\mathcal{K}^{(m)} = \left\{v : a_i^T v = 0 \quad \forall i = 1, \ldots, q \quad \text{and} \quad a_i^T v \leq 0 \quad \forall i \in W^{(m)} \cap \{i : i > q\}\right\} \tag{5.3}$$

which is the cone of feasible directions at any point in Ω for which $W^{(m)}$ is the active set of constraints. A positive basis for the null space of the equality constraints is constructed which contains a positive basis for any cone (see section 5.1) defined by any subset of $W^{(m)}$ containing all equality constraints. A positive basis which satisfies these conditions is said to be *aligned* with the set of constraints $W^{(m)}$ at $z^{(m)}$ or, more simply, aligned. Occasionally the phrase "aligned with a cone" is used; it means aligned with the set of constraints defining that cone. A frame is constructed by the same process used in (2.1). Any such frame is also called aligned. In the next section the formation of aligned frames is discussed, followed by the barrier approach to linearly constrained problems.

5.1. Generating aligned positive bases and frames. A polyhedral cone $\mathcal{K}$ may be defined as the intersection of a finite number of halfspaces and hyperplanes:

$$\mathcal{K} = \left\{v : a_i^T v = 0, \quad i = 1, \ldots, q \quad \text{and} \quad a_i^T v \leq 0, \quad i = q+1, \ldots, \ell\right\}. \tag{5.4}$$

For convenience we have omitted the (m) superscripts and have assumed that the first $\ell - q$ inequality constraints are those in the current working set. *In this subsection only*, the constraints under discussion are those defining the cones of feasible directions. These constraints are of the form

$$a_i^T v = 0, \quad i = 1, \ldots, q \quad \text{and} \quad a_i^T v \leq 0, \quad i = q+1, \ldots, \ell.$$

That is to say, the constants b_i have been omitted from the constraints which define Ω. Any such cone can be rewritten as a finitely generated cone

$$\exists v_1, \ldots, v_p \quad \text{such that} \quad \mathcal{K} = \left\{\sum_{i=1}^{p} \eta_i v_i : \eta_i \geq 0 \ \ \forall i = 1, \ldots, p\right\} \tag{5.5}$$

as is shown by Theorem 4.18 of [16]. The vectors $v_1, \ldots, v_p$ are often referred to as a set of *generators* of the cone $\mathcal{K}$. A minimal set of generators $\mathcal{V}_+$ for a closed polyhedral cone $\mathcal{K}$ is a set of vectors $\{v_1, \ldots, v_p\}$ such that

K1: $\{v_1, \ldots, v_p\}$ satisfies (5.5) and
K2: no proper subset of $\mathcal{V}_+$ satisfies (5.5).

Initially we consider the special case where the a_i, $i \leq \ell$, are linearly independent. A positive basis aligned with $\mathcal{K}$ is constructed in two parts: one each for the subspace containing these a_i and for the subspace orthogonal to these a_i. For illustrative purposes, choose any basis for R^n which satisfies $a_i = e_i$ for $i = 1, \ldots, \ell$ but is otherwise arbitrary. Here e_i is the ith unit vector. If $\mathcal{U}_+$ is any positive basis for the subspace spanned by $e_{\ell+1}, \ldots, e_n$, then

$$\{-e_i : i = q+1, \ldots, \ell\} \cup \mathcal{U}_+$$

is a set of generators for $\mathcal{K}$. Interestingly, this is a subset of the following positive basis for the null space of the equality constraints

$$\mathcal{V}_+ = \{\pm e_i : i = q+1, \ldots, \ell\} \cup \mathcal{U}_+.$$

This positive basis for the null space of the equality constraints contains a set of generators for every polyhedral cone defined by the equality constraints and any subset of the constraints $v^T e_i \geq 0$, $v^T e_i \leq 0$, and $v^T e_i = 0$ for $i = q+1, \ldots, \ell$. This property is crucial: it means that $\mathcal{V}_+$ contains a set of generators for every possible cone of feasible directions at $z^{(m)}$ and at all points near $z^{(m)}$.

We now revert back to the original basis for R^n and work with a_i. The assumption that the set $\{a_i : i \in W^{(m)}\}$ is linearly independent is retained. For notational simplicity we continue to assume that $W^{(m)} = \{1, \ldots, \ell\}$. Let $A = [a_1, \ldots, a_\ell]$ and select an invertible matrix $S = [s_1, \ldots, s_n]$ satisfying $S^T A = [e_1, \ldots, e_\ell]$. This allows the following Theorem to be stated.

THEOREM 5.1. *If W and S are as defined above, then the set*

$$\{-s_i : i = q+1, \ldots, \ell\} \cup \{Su : u \in \mathcal{U}_+\} \tag{5.6}$$

is an ordered minimal set of generators for the cone $\mathcal{K}$ defined in (5.4). *Here $\mathcal{U}_+$ is an ordered positive basis for the subspace spanned by $e_{\ell+1}, \ldots, e_n$.*

Proof. First, it is clear that all members of (5.6) lie in $\mathcal{K}$. We now must show that an arbitrary $w_1 \in \mathcal{K}$ can be expressed as a nonnegative linear combination of the members of (5.6). Since $w_1 \in \mathcal{K}$, it follows that $A^T w_1 \leq 0$. Moreover, the first q elements of $A^T w_1$ must be zero. For convenience let $A^T w_1 = y$. Now, for appropriate nonnegative choices of η_i, $i = q+1, \ldots, \ell$, the vector

$$w_2 = \sum_{i=q+1}^{\ell} \eta_i(-s_i) \quad \text{solves} \quad A^T w_2 = \sum_{i=q+1}^{\ell} -\eta_i e_i = y \leq 0.$$

Hence $w_2 - w_1$ is a member of the null space of A^T (hereafter $N(A^T)$). Clearly w_2 is a nonnegative linear combination of the members of (5.6). Moreover, $\{Su : u \in \mathcal{U}_+\}$ is an ordered positive basis for the null space $N(A^T)$, and so w_1 can be written as a nonnegative linear combination of the members of (5.6).

Minimality can be seen as follows. For a specific $j \in q+1, \ldots, \ell$ one has $A^T(-s_j) = -e_j$, whereas $e_j^T A^T v = 0$ for all other v in (5.6). Hence $-s_j$ cannot be expressed as a nonnegative linear combination of the remaining members of (5.6). Finally, assume some Su_j, $u_j \in \mathcal{U}_+$ is redundant, i.e.,

$$Su_j = \sum_{i=q+1}^{\ell} \sigma_i(-s_i) + \sum_{i \neq j} \theta_i Su_i \tag{5.7}$$

for some $\theta_i, \sigma_i \geq 0$. Now $A^T S u = 0$ for all $u \in \mathcal{U}_+$, which implies $\sigma_i = 0$ for all i. Multiplying (5.7) by S^{-1} yields a contradiction with the fact that $\mathcal{U}_+$ is a positive basis. ∎

COROLLARY 5.2. *The set*

$$\mathcal{V}_+ = \{s_i : i = q+1, \ldots, \ell\} \cup \{-s_i : i = q+1, \ldots, \ell\} \cup \{Su : u \in \mathcal{U}_+\} \tag{5.8}$$

contains a set of generators for any cone defined by the equality constraints and any subset of the inequality constraints in (5.4).

Proof. Without loss of generality let the selected subset of inequality constraints be indexed by $i = q+1, \ldots, r$, where $r \leq \ell$. Using $\mathcal{U}_+$ as a positive basis for the subspace spanned by $e_{\ell+1}, \ldots, e_n$ as above, the set

$$\{\pm s_i : i = r+1, \ldots, \ell\} \cup \{Su : u \in \mathcal{U}_+\}$$

is a positive basis for the null space $N([a_1 \ldots a_r]^T)$. The corollary then follows from Theorem 5.1. ∎

As an illustration, consider the constraints $x_1 \leq 0$ and $x_2 \leq 0$ in R^2. Equation (5.8) gives $\mathcal{V}_+ = \{\pm e_1, \pm e_2\}$. There are four possible sets of active constraints: none; $x_1 \leq 0$ only; $x_2 \leq 0$ only; and both. The sets of generators for the corresponding tangent cones are $\mathcal{V}_+$, $\{-e_1, \pm e_2\}$, $\{\pm e_1, -e_2\}$, and $\{-e_1, -e_2\}$. Equation (5.8) is used to define each $\mathcal{V}_+^{(m)}$. Corollary 5.2 means that every $\mathcal{V}_+^{(m)}$ is a positive basis for the null space of the set of equality constraints.

When degeneracy is present in a set of active constraints the above approach must be modified. (Readers not interested in the degenerate case may wish to proceed directly to Assumption 5.3.) The existence of an aligned positive spanning set $\mathcal{V}_+^{\mathrm{ld}}$ is guaranteed by Theorem 4.18 of [16], but its construction can be computationally expensive [13]. The superscript "ld" is used to highlight the fact that $\mathcal{V}_+^{\mathrm{ld}}$ is not necessarily a positive basis and is no longer defined by (5.8). The set $\mathcal{V}_+^{\mathrm{ld}}$ must contain a set of generators for the cone $\mathcal{K}$ in (5.4) and *also* for every cone defined by any subset of the constraints in W which includes all equality constraints. If the constraints are linearly dependent, then $\mathcal{V}_+^{\mathrm{ld}}$ is a positive spanning set for the subspace defined by the equality constraints, but it is no longer a positive basis. For convenience, in the following discussion we assume any linear dependence in the subset of equality constraints has been removed by deleting redundant equality constraints.

The construction of $\mathcal{V}_+^{\mathrm{ld}}$ is in two parts. The first part is a positive basis for the null space of the normals of the constraints indexed by W. The second part is for the subspace T spanned by $a_{q+1}, \ldots, a_\ell$, where T is of dimension $r - q$.

For the first part of $\mathcal{V}_+^{\mathrm{ld}}$, order the constraints so that $a_1, \ldots, a_{q+r}$ are linearly independent. Let the invertible matrix $S = [s_1, \ldots, s_n]$ satisfy $S^T[a_1, \ldots, a_r] = [e_1, \ldots, e_r]$. The first part of $\mathcal{V}_+^{\mathrm{ld}}$ is

$$S\mathcal{U}_+ = \{Su : u \in \mathcal{U}_+\},$$

where $\mathcal{U}_+$ is a positive basis for the subspace spanned by $e_{r+1}, \ldots, e_n$. Clearly $S\mathcal{U}_+$ positively spans the null space $N([a_1, \ldots, a_r]^T)$.

The second part of $\mathcal{V}_+^{\mathrm{ld}}$ contains a positive scalar multiple of each vector $\pm v$ which lies in T and satisfies with equality any $r-1$ linearly independent constraints indexed by the set W, including all equality constraints. Note that $\pm v$ may violate

any constraint it is not required to satisfy with equality. Clearly $\mathcal{V}_+^{\rm ld}$ must contain the $\mathcal{V}_+$ defined by (5.8) for each subset of W which includes all equality constraints and has r linearly independent constraint normals. The set of all such vectors for all such subsets of W includes many pairs of vectors which are positive scalar multiples of one another. Eliminating such pairs gives $\mathcal{V}_+^{\rm ld}$. In the particular case when all constraint normals indexed by W are linearly independent, these vectors $\pm v$ are positive scalar multiples of $\pm s_{q+1}, \ldots, \pm s_\ell$ in (5.8), and the definition of $\mathcal{V}_+^{\rm ld}$ reverts back to that in (5.8).

It is now shown that $\mathcal{V}_+^{\rm ld}$ contains a set of generators for every cone defined by any subset of the constraints in W which includes all equality constraints. Consider a cone $\mathcal{K}_{\rm s}$ defined by a subset $W_{\rm s}$ of W which contains all equality constraints but is otherwise arbitrary. Let the dimension of $\mathcal{K}_{\rm s}$ be ρ. For convenience reorder the inequality constraints so that $a_1, \ldots, a_\rho$ are linearly independent, and $1, \ldots, \rho \in W_{\rm s}$. Define $T_{\rm s}$ to be the subspace spanned by $a_{q+1}, \ldots, a_\rho$. We now add further constraint normals a_i, $i \in W$, to $a_1, \ldots, a_\rho$ to obtain a maximal linearly independent set $a_1, \ldots, a_r$. The construction of $\mathcal{V}_+^{\rm ld}$ ensures that it contains an ordered positive basis (given by (5.8)) defined by the working set $\{1, \ldots, r\}$. Hence $\mathcal{V}_+^{\rm ld}$ must contain a positive basis for the null space of $a_1, \ldots, a_\rho$ by Corollary 5.2.

It remains to show that $\mathcal{V}_+^{\rm ld}$ contains a set of generators for the cone $\mathcal{K}_{\rm s} \cap T_{\rm s}$. Define the hyperplane $\mathcal{H} = \{v \in T_{\rm s} : (a_{q+1} + \cdots + a_\rho)^T v = -1\}$. Clearly $\mathcal{K}_{\rm s} \cap T_{\rm s}$ is contained in the cone $\{v \in T_{\rm s} : a_i^T v \le 0 \text{ for all } i = q+1, \ldots, \rho\}$. Also, because $a_{q+1}, \ldots, a_\rho$ is a basis for $T_{\rm s}$, it is clear that $\mathcal{K}_{\rm s} \cap T_{\rm s} \cap \mathcal{H}$ is bounded, and hence is a polytope P. It can be shown [16] that a set of generators for $\mathcal{K}_{\rm s} \cap T_{\rm s}$ is precisely the set of vectors from the origin to the vertices of P. Each of these vectors v satisfies $a_i^T v = 0$ for all but one of $i = 1, \ldots, \rho$. By adding an appropriate member of $N([a_1, \ldots, a_\rho]^T)$ to v one can obtain a vector v_+ which satisfies $a_i^T v_+ = 0$ for all but one $i \in 1, \ldots, r$. Hence each such v_+ is a positive scalar multiple of a member of $\mathcal{V}_+^{\rm ld}$. Thus $\mathcal{V}_+^{\rm ld}$ contains an ordered set of generators for every cone $\mathcal{K}_{\rm s}$ defined by any subset of constraints in W which includes all equality constraints.

The following assumption is needed to ensure the limits of the sequence of positive bases have the required properties.

Assumption 5.3.

(a) All limits of the sequence of ordered positive bases $\{\mathcal{U}_+^{(m)}\}$ are ordered positive bases.

(b) The methods used to generate S and $\mathcal{V}_+ - S\mathcal{U}_+$ are repeatable. That is to say, they will always return the same S and $\mathcal{V}_+ - S\mathcal{U}_+$ when given the same working set W.

(c) Each $\mathcal{V}_+^{(m)}$ satisfies (2.4).

This assumption is the equivalent of Assumption 2.1 for the LCOP (5.1). In the case when constraints are absent, $\mathcal{U}_+^{(m)} \equiv \mathcal{V}_+^{(m)}$ for all m, and Assumption 5.3 reduces to Assumption 2.1. For the case when the normals of the constraints in W are linearly independent Assumption 5.3(b) amounts to returning the same S when given the same W. It is possible that some members of $\mathcal{V}_+^{(m)}$ violate the bound in (2.4). This bound is imposed retrospectively by replacing any $v \in \mathcal{V}_+^{(m)}$ violating (2.4) with $Kv/\|v\|$.

5.2. The template for linearly constrained problems. The following template lists the specialized form of Template D required for linearly constrained problems.

Algorithm Template E.

1. Initialize: set $k = 1$, $m = 1$, and choose the initial point $x^{(0)} \in \Omega$. Choose $\delta > 0$.
2. Choose $H^{(m)} > 0$, $E^{(m)} \geq 0$, and $\mathcal{G}^{(m)}$.
3. Choose $h^{(k)} \geq H^{(m)}$ and $\epsilon^{(k)} \geq E^{(m)}$.
4. Execute any finite process which satisfies one of these conditions:
   (a) generates an iterate $x^{(k)} \in \Omega \cap \mathcal{G}^{(m)}$ satisfying $f\left(x^{(k)}\right) < f^{(k-1)} - \epsilon^{(k)}$; or
   (b) generates a quasi-minimal frame $\Phi^{(m)} = \Phi(z^{(m)}, h_{\mathrm{z}}^{(m)}, \mathcal{V}_{+}^{(m)})$, where $x^{(k)} \in \Omega$ and $f^{(k)} \leq f^{(k-1)}$. Here $z^{(m)} = x^{(k)}$, $h_{\mathrm{z}}^{(m)} = h^{(k)}$, and $\epsilon_{\mathrm{z}}^{(m)} = \epsilon^{(k)}$. The frame $\Phi^{(m)}$ must be aligned with an identified working set $W^{(m)}$ satisfying (5.2); or
   (c) case (b) of this step with the added restriction $\Phi^{(m)} \subset \mathcal{G}^{(m)}$.
5. If $x^{(k)}$ is not quasi-minimal, increment k and go to step 3.
6. Increment m and k. If stopping conditions are not satisfied, go to step 2.

If the active constraint normals are linearly independent at every point on the boundary of Ω, then the constant δ in Template E can be defined implicitly. The case when all constraint normals are linearly independent is trivial. For the remaining case, the equality constraints are indexed by $i = 1, \ldots, q$ as above, and the inequality constraints are ordered so that $|a_i^T z^{(m)} + b_i|$ is an increasing function of i. The working set W is chosen as the largest set $\{1, \ldots, r\}$ for which the corresponding constraint normals are linearly independent. The residual of the $r + 1$st constraint in this list must have a uniform positive lower bound for all feasible z in an arbitrary compact set Ξ, and δ can be chosen as this bound. If this were not the case there would be a linearly dependent set of constraints indexed by W, say, and also a sequence of points $\{z_j\} \subset \Omega$ for which

$$\lim_{j \to \infty} \left(\max_{i \in W} \left| a_i^T z_j + b_i \right| \right) = 0.$$

Continuity of the constraint functions then implies a degenerate point exists on the boundary of Ω, contradicting the initial assumption. Assumption 4.2(a) ensures a compact set Ξ exists which contains all points of interest.

5.3. Convergence results for the linearly constrained case. Algorithms conforming to Template D may be applied to the LCOP by applying such methods to the barrier function f_c. The non-Lipschitz nature of f_c means that Theorem 4.3 is not directly applicable, and convergence to solution points of the LCOP (5.1) must be established some other way. In order to guarantee convergence to solution(s) of the LCOP under standard conditions a further restriction must be imposed. Specifically, each frame $\Phi^{(m)}$ generated by such an algorithm must be aligned with the working set of constraints $W^{(m)}$. This working set includes all constraints with small residuals ($\leq \delta$) at $z^{(m)}$. With this restriction, Template D becomes Template E.

First we note that Theorem 4.1 is directly applicable, and parts (a) and (b) of Assumption 4.2 ensure the sequence of quasi-minimal iterates $\{z^{(m)}\}$ is infinite and has cluster points. Next we establish a constrained version of Theorem 4.3.

Definition 5.4. *Let $z^{(\infty)}$ be a cluster point of the sequence of quasi-minimal iterates. Define $\mathcal{S}_{+}^{(\infty)}(W)$ as the set of all vectors v for which there exists an infinite subsequence $\{(z^{(m)}, v^{(m)})\}_{m \in \mathcal{M}}$ with the following properties:*

(i) *this subsequence converges uniquely to $(z^{(\infty)}, v)$;*

(ii) $z^{(m)} + h_z^{(m)} v^{(m)} \in \Omega$ *for all* $m \in \mathcal{M}$;
(iii) $v^{(m)} \in \mathcal{S}_+^{(m)}$ *for all* $m \in \mathcal{M}$, *where* $\mathcal{S}_+^{(m)}$ *is as defined in* (3.1);
(iv) $W^{(m)} = W$ *for all* $m \in \mathcal{M}$;
(v) $\|z^{(m)} - z^{(\infty)}\| < \gamma$ *for all* $m \in \mathcal{M}$, *where* $\gamma > 0$;
(vi) *no point in the closed ball of radius* 2γ *about* $z^{(\infty)}$ *violates any constraint not in the active set for the point* $z^{(\infty)}$; *and*
(vii) $h_z^{(m)} < \gamma / K$, *where* K *is the constant used in the upper bound* (2.4) *on each* $\|v^{(m)}\|$.

The notation $\mathcal{S}_+^{(\infty)}$ *is used to denote the union of all* $\mathcal{S}_+^{(\infty)}(W)$ *when* W *ranges over all possible subsets of the set of constraint indices* $\{1, \ldots, L\}$.

Condition (ii) in Definition 5.4 requires that v be a feasible direction at $z^{(\infty)}$. Note that the finiteness of the number of different working sets W means that any $\mathcal{M}$ satisfying all conditions except (iv) will have an infinite subset which satisfies all seven conditions. Hence condition (iv) does not exclude any v from $\mathcal{S}_+^{(\infty)}$. For any $(z^{(\infty)}, v)$, conditions (iv)–(vii) can always be satisfied for an appropriate choice of $\mathcal{M}$, provided the first three conditions can. This follows directly from condition (i) and the facts that the number of constraints is finite and $h_z^{(m)} \to 0$ as $m \to \infty$. Conditions (v)–(vii) mean that constraints not in W are automatically satisfied by all $z^{(m)} + h_z^{(m)} v$ when $v \in \mathcal{S}_+^{(m)}$ and $m \in \mathcal{M}$. In particular, this includes all points in the frames $\Phi^{(m)}$, $m \in \mathcal{M}$. Conditions (i)–(iii) are needed for the proof of Theorem 5.5. The last four conditions are superfluous to the proof of Theorem 5.5 but are needed in the proof of Theorem 5.6.

THEOREM 5.5. *Let* $z^{(\infty)}$ *be a cluster point of the sequence of quasi-minimal iterates. Then* $f^\circ\left(z^{(\infty)}, v\right) \geq 0$ *for all* v *in* $\mathcal{S}_+^{(\infty)}$.

Proof. Since every $z^{(m)}$ is feasible, condition (ii) of Definition 5.4 allows us to use the fact that $f \equiv f_c$ on Ω. Conditions (i) and (iii) of Definition 5.4 allow Theorem 4.3 to be invoked, yielding the required result. ∎

Next it is shown that Theorem 5.5 applies to a set of directions rich enough to include a set of generators for the cone of feasible directions at $z^{(\infty)}$.

THEOREM 5.6. *Let* $z^{(\infty)}$ *be a limit point of the sequence of quasi-minimal iterates. The set* $\mathcal{S}_+^{(\infty)}$, *as defined in Definition* 5.4, *contains a set of generators for the cone of feasible directions* $K^{(\infty)}$ *at the limit point* $z^{(\infty)}$.

Proof. Let $W^{(\infty)}$ be the set of active constraints at $z^{(\infty)}$. Consider an infinite increasing sequence of positive integers $\mathcal{M}$ with the following properties:

(a) $z^{(m)} \to z^{(\infty)}$ as $m \to \infty$, $m \in \mathcal{M}$;
(b) $\mathcal{V}_+^{(m)} \to \mathcal{V}_+^{(\infty)}$ as $m \to \infty$, $m \in \mathcal{M}$;
(c) $W^{(m)}$ is the same for all $m \in \mathcal{M}$;
(d) $\|z^{(m)} - z^{(\infty)}\| < \gamma$ for all $m \in \mathcal{M}$, where γ is a positive constant;
(e) no point in the closed ball of radius 2γ centered on $z^{(\infty)}$ violates any constraint not in $W^{(\infty)}$; and
(f) $h_z^{(m)} < \gamma / K$ for all $m \in \mathcal{M}$, where K is the constant in (2.4).

The existence of $\mathcal{M}$ is guaranteed by the following facts: (a) holds because $z^{(\infty)}$ is a limit point of $\{z^{(m)}\}$; (b) holds by Assumption 5.3; (c) and (e) hold because the number of different possible working sets is finite; (d) follows from (a); and (f) follows from Assumption 4.2(c).

If attention is restricted to the sequence $\{v_i^{(m)}\}_{m \in \mathcal{M}}$ for a fixed value of i, then (a) and (b) together, and (c), (d), (e), and (f), respectively, yield items (i), (iv), (v),

(vi), and (vii) of Definition 5.4. The fact that $\mathcal{V}_+^{(m)} \subseteq \mathcal{S}_+^{(m)}$ for all m yields item (iii) of Definition 5.4, which leaves just condition (ii).

Now (5.2) implies $W^{(\infty)} \subseteq W^{(m)}$ for all m such that $z^{(m)}$ is sufficiently near $z^{(\infty)}$. Hence (a) and (c) imply $W^{(\infty)} \subseteq W^{(m)}$ for all m in $\mathcal{M}$. Therefore each $\mathcal{V}_+^{(m)}$, $m \in \mathcal{M}$, contains a set of generators for the cone $\mathcal{K}^{(\infty)}$ of feasible directions at $z^{(\infty)}$. Each such set of generators for $\mathcal{K}^{(\infty)}$ consists of two parts. The first part (hereafter $\mathcal{W}_+$) is the part contained in the span of $\{a_i : i \in W^{(m)}\}$ and the second part is $S\mathcal{U}_+^{(m)} = \{Su : u \in \mathcal{U}_+^{(m)}\}$. The construction of the first part depends only on $W^{(m)}$, and so $\mathcal{W}_+$ is the same for all $m \in \mathcal{M}$ by Assumption 5.3(b). The matrix S is also independent of m, again by Assumption 5.3(b). Item (b) means that $\{\mathcal{U}_+^{(m)}\}_{m\in\mathcal{M}}$ has a unique limit $\mathcal{U}_+^{(\infty)}$, which is an ordered positive basis, by Assumption 5.3(a). Clearly $\mathcal{W}_+$ and $S\mathcal{U}_+^{(\infty)} = \{Su : u \in \mathcal{U}_+^{(\infty)}\}$ are both subsets of $\mathcal{S}_+^{(\infty)}(W^{(m)})$. The set $\mathcal{W}_+ \cup S\mathcal{U}_+^{(\infty)}$ is also a set of generators for $\mathcal{K}^{(\infty)}$.

Each $z^{(m)}$ lies in Ω, and each member of $\mathcal{W}_+ \cup S\mathcal{U}_+^{(m)}$ lies in $\mathcal{K}^{(\infty)}$. Hence $z^{(m)} + h_z^{(m)} v$ does not violate any constraint indexed by the set $W^{(\infty)}$ for all $m \in \mathcal{M}$ and for all $v \in \mathcal{W}_+ \cup S\mathcal{U}_+^{(m)}$. Items (d)–(f) and the bound on $\|v\|$ in (2.4) imply no member of $\mathcal{W}_+ \cup S\mathcal{U}_+^{(m)}$, $m \in \mathcal{M}$, can violate any constraint not in $W^{(\infty)}$. Thus all conditions of Definition 5.4 hold for all $v^{(m)} \in \mathcal{W}_+ \cup S\mathcal{U}_+^{(m)}$ for every m in $\mathcal{M}$. Their limits $\mathcal{W}_+ \cup S\mathcal{U}_+^{(\infty)}$ are a set of generators for $\mathcal{K}^{(\infty)}$ contained in $\mathcal{S}_+^{(\infty)}$, as required. ∎

Theorems 5.5 and 5.6 show that a set of generators for the cone $\mathcal{K}^{(\infty)}$ exists such that f° is nonnegative at $z^{(\infty)}$ along each of these generators, where $\mathcal{K}^{(\infty)}$ is the cone of feasible directions at $z^{(\infty)}$. The following result extends this to all feasible directions in the case when f is strictly differentiable at $z^{(\infty)}$.

THEOREM 5.7. *If f is strictly differentiable at $z^{(\infty)}$, then no feasible direction exists at $z^{(\infty)}$ along which f has a negative directional derivative.*

Proof. Now, for a general $v \in \mathcal{K}^{(\infty)}$, we can write

$$v = \sum_{i=1}^{p} \eta_i v_i^{(\infty)}, \quad \text{where} \quad \eta_i \geq 0 \quad \forall i$$

and where $\{v_1^{(\infty)}, \ldots, v_p^{(\infty)}\} \subseteq \mathcal{S}_+^{(\infty)}$ is a set of generators for $\mathcal{K}^{(\infty)}$. The strict differentiability of f at $z^{(\infty)}$ yields

$$\nabla f^T v = \sum_{i=1}^{p} \eta_i \nabla f^T v_i^{(\infty)} = \sum_{i=1}^{p} \eta_i f^\circ \left(z^{(\infty)}; v_i^{(\infty)}\right) \geq 0,$$

as required. Hence no feasible direction exists at $z^{(\infty)}$ along which f has a negative directional derivative, and $z^{(\infty)}$ is a solution of the LCOP (5.1). ∎

For the moment we continue to consider a subsequence of quasi-minimal iterates as defined in the proof of Theorems 5.5–5.7. It is shown in the proof of Theorem 5.6 that $z^{(m)} + h_z^{(m)} v \in \Omega$ for every $v \in \mathcal{W}_+ \cup S\mathcal{U}_+^{(m)}$ when $m \in \mathcal{M}$. In contrast, in early iterations $z + hv$ can easily violate constraints not in $W^{(\infty)}$. When this occurs $f_c = +\infty$ at $z + hv$, and the direction v is effectively ignored. Rather than do this, one could evaluate f at $z + \alpha v$, where $\alpha \in R$ is the largest value such that $z + \alpha v \in \Omega$. Any such function evaluations can be included in the arbitrary finite process of step 4

of the template, which means that Theorem 5.6 still applies. The advantage of such function evaluations is that an algorithm can look along the direction v immediately, rather than having to wait until h is small enough to make $z + hv$ feasible.

6. Selecting the frame size. Template D imposes a number of restrictions on h. In addition to the explicit requirement that h tend to zero, there is also a sequence of lower bounds $\{H^{(m)}\}$. These lower bounds are not required for convergence purposes, but other lower bounds on h are implicit in Assumption 4.2(d) and condition G2. The presence of the explicit lower bounds on h in Template D is to reinforce the fact that the implicit lower bounds exist. These implicit lower bounds are discussed first, and the cases $E = 0$ and $E > 0$ are treated separately.

If $E = 0$, then condition G1 and the bound (2.4) mean that condition G2 can not be satisfied if h is too small. In practice one could define $\mathcal{G}^{(m)}$ using a length $H^{(m)}$ which would become a lower bound for h until the next quasi-minimal frame is located. For example, in [6]

$$\mathcal{G}^{(m)} = \left\{ x_0 + h \sum_{i=1}^{n} \eta_i v_i : \eta_i \text{ is integer } \forall i = 1, \ldots, n \right\}$$

is used, where x_0 is the origin of the grid and where $v_1, \ldots, v_n$ are a basis for R^n. In [6] h is used to define both the admissible set $\mathcal{G}^{(m)}$ and also the quasi-minimal frame Φ contained in that set. Therefore in [6], h is kept constant between quasi-minimal frames. Under Template D the h value used to define the grid would become the lower bound $H^{(m)}$, and h values in excess of this would be permitted.

If $E > 0$, then the requirement that $\epsilon/h \to 0$ means that h must approach zero more slowly than E. The simplest method of ensuring this is to connect ϵ and h via a relation like $\epsilon = Nh^\nu$, where $N > 0$ and $\nu > 1$. The bound $\epsilon \geq E$ is then equivalent to a positive lower bound on h. In fact, in [4] the lower bound on ϵ is imposed indirectly via this relation and a specific positive lower bound H on h.

The convergence theory requires that $h \to 0$ as $k \to \infty$ but does not state how this is to be done. Simple approaches such as using $h^{(k)} = 2^{-k}$ have obvious drawbacks. Indeed, h permanently falls below machine precision after a fixed number of iterations. Such an approach takes no account of how quickly or slowly the sequence of iterates is converging. When a solution is located quickly h should become small quickly in order to verify that it is indeed a solution. In contrast, if good reductions in f occur with h large, then h should remain large until such reductions cease. Similarly, if the sequence of iterates moves from a region where small steps are necessary into a region where large steps are better, then h should increase. This suggests that h should vary in sympathy with the lengths of recent steps and also with the recent reductions in f. One possibility is to impose an upper bound on h of the form

$$h^{(k)} \leq \max \left\{ \gamma H^{(m(k))}, \Upsilon_{\mathrm{x}}^{(k)} \Upsilon_{\mathrm{f}}^{(k)} \right\}. \tag{6.1}$$

Here Υ_{f} and Υ_{x} are moving averages of the past decreases in function values and step lengths, respectively, and γ is a constant satisfying $\gamma \geq 1$. The value $m(k)$ is the value of m at step 3 of iteration k. The two moving averages are defined in terms of two sequences $\{\omega_i\}_{i=1}^{\infty}$ and $\{\beta_i\}_{i=1}^{\infty}$ of nonnegative weights as follows:

$$\Upsilon_{\mathrm{f}}^{(k)} = \sum_{i=1}^{k-1} \omega_{k-i} \left| f^{(i-1)} - f^{(i)} \right| \quad \text{and} \quad \Upsilon_{\mathrm{x}}^{(k)} = \sum_{i=1}^{k-1} \beta_{k-i} \left\| x^{(i)} - x^{(i-1)} \right\|,$$

with the convention that $\Upsilon_{\rm f}^{(1)} = \Upsilon_{\rm x}^{(1)} = 0$. The two sequences of weights chosen are so that

$$\sum_{i=1}^{\infty} \omega_i \quad \text{and} \quad \sum_{i=1}^{\infty} \beta_i$$

are both finite. If the sequence of iterates is bounded, then the sequence of $\Upsilon_{\rm x}^{(k)}$ values is bounded. If f is bounded below on any bounded set, then the sequence of $\Upsilon_{\rm f}^{(k)}$ values must converge to zero. Given the quadratic nature of smooth functions near local minima, one could replace $\Upsilon_{\rm f}^{(k)}$ with its square root in (6.1).

An alternative for the case when E is always positive is presented in [4]. There $\epsilon = Nh^{\nu}$ is used, with $\nu > 1$, and $N > 0$. The bound $\epsilon \geq E$ is imposed indirectly by imposing a strictly positive lower bound $H^{(m)}$ on h. When sufficient descent is obtained h may be increased by up to a fixed multiple of itself. If a quasi-minimal frame is located, then h is reduced in such a way that if h is reduced repeatedly, then $h \to 0$. In essence, if the sequence $\{h\}$ has a strictly positive lower bound ($h_{\min}$ say), then f is reduced by at least $Nh_{\min}^{\nu}$ an infinite number of times. Hence either $h \to 0$ or $f \to -\infty$. More details are presented in [4].

Template D and related algorithm frameworks gain much flexibility by not making $h \to 0$ a direct consequence of conforming to the template. In contrast, GPS [17] guarantees $h \to 0$ when f is C^1 and the sequence of iterates is bounded. If Template D satisfies (6.1) or uses the approach in [4], then it also guarantees $h \to 0$ under the same conditions.

7. Concluding remarks. Template D contains algorithms which bear a striking resemblance to the implicit filtering algorithms in [2, 11]. These implicit filtering algorithms use the positive basis $\{\pm e_1, \ldots, \pm e_n\}$ to form frames (or, in the language of [2, 11], stencils) about the current iterate x. Using the frame an estimate g of the gradient at x is formed, and a search direction is generated. A finite line search is conducted along this direction, where satisfaction of a sufficient descent condition is sought. If sufficient descent is not obtained, if the gradient at x is small, or (in [2]) if the frame is minimal, then h is reduced. Given that convergence occurs, [11] shows that convergence rate is linear for bound constrained problems when the line search direction is $-g$, except that any infeasible point in the line search is replaced with the closest feasible point to it. In the absence of bounds, [2] shows that a superlinear rate can be obtained when a quasi-Newton search direction is used.

Implicit filtering differs from Template D on a number of points, including the type of sufficient descent condition in the line search and that frame points are not considered as candidates for the next iterate. This last difference has enormous theoretical implications. The possibility of a frame point becoming the next iterate rather than a line search point means that the rate theorems of implicit filtering are not guaranteed to apply to any algorithm conforming to Template D. The absence of steps to frame points in implicit filtering means that the convergence theory behind Template D is inapplicable to implicit filtering. Steps to frame points are *crucial* to the convergence theory, and so implicit filtering in its current form [2, 11] falls outside the scope of Template D.

Nevertheless, from a practical perspective, implicit filtering is very similar to some algorithms conforming to Template D. Minor modifications to implicit filtering would make it conform to Template D, and hence provably convergent. In the case when steps to frame points occur only in early iterations (numerical experiments in [8]

suggest that this is common) the rate theorems of implicit filtering would apply. The numerical results for implicit filtering [2, 11] and the work on the global aspects of finite differences [19] show that the use of frames can enable algorithms to "step over" many local minima to find a much lower minimum.

A frame-based template for unconstrained and linearly constrained optimization has been developed. Applicability to linearly constrained problems is achieved by aligning frames with active and nearly active constraints. The use of frames means that clearly inactive constraints can be ignored, and the linear constraints can involve irrational numbers, in contrast to [1, 13]. It has been shown that algorithms conforming to the template generate sequences of quasi-minimal iterates whose cluster points are stationary points of the optimization problem under mild conditions. The cluster points of the sequence of quasi-minimal iterates retain interesting properties even when the objective function is not differentiable.

The approach taken unifies methods using sufficient descent and simple descent. The former use the sufficient descent condition to ensure quasi-minimal frames are generated. The latter do so when necessary by restricting the frame points (but not the frame centers) to admissible sets. The frame centers are not restricted, which allows these simple descent methods to select quasi-minimal iterates which lie outside of the admissible set *every* iteration. The facts that the admissible sets can be unrelated to one another, can incorporate random elements, and can sometimes yield quasi-minimal frames outside the admissible set means that *for some algorithms conforming to Template D* there is no "pattern" restricting the locations of iterates. All that can be said is that the admissible sets get finer as h approaches zero. This is a level of flexibility not present in previous simple descent methods such as GPS [13, 17] or [6]. These earlier simple descent methods also specifically use rectangular grids or subsets of them. The greater choice of admissible sets allows these sets to possess, for example, circular or spherical symmetries in some dimensions. This could be very useful when, for example, minimizing a quadratic penalty function involving nonlinear constraints with known symmetries.

Template D encompasses a wide class of algorithms including existing frame-based and grid-based methods. Numerical results for existing methods in this class [5, 8, 15] show that there are effective methods conforming to Template D. There is much scope for future work in developing algorithms which exploit the great flexibility afforded by the template.

Acknowledgments. The authors wish to thank Professor John Dennis for some very useful suggestions, including those concerning the use of Clarke's nonsmooth analysis. The authors also wish to thank Professor Tim Kelley for drawing implicit filtering to their attention and three anonymous referees for their many helpful comments and suggestions.

REFERENCES

[1] C. Audet and J. E. Dennis, Jr., *Analysis of generalized pattern searches*, SIAM J. Optim., 13 (2003), pp. 889–903.

[2] T. D. Choi and C. T. Kelley, *Superlinear convergence and implicit filtering*, SIAM J. Optim., 10 (2000), pp. 1149–1162.

[3] F. H. Clarke, *Optimization and Nonsmooth Analysis*, Classics Appl. Math. 5, SIAM, Philadelphia, 1990.

[4] I. D. Coope and C. J. Price, *Frame based methods for unconstrained optimization*, J. Optim. Theory Appl., 107 (2000), pp. 261–274.

[5] I. D. Coope and C. J. Price, *A direct search conjugate directions algorithm for unconstrained minimization*, ANZIAM J., 42 (2000), pp. C478–C498.
[6] I. D. Coope and C. J. Price, *On the convergence of grid-based methods for unconstrained optimization*, SIAM J. Optim., 11 (2001), pp. 859–869.
[7] I. D. Coope and C. J. Price, *Positive bases in numerical optimization*, Comput. Optim. Appl., 21 (2002), pp. 169–175.
[8] I. D. Coope and C. J. Price, *A Derivative-Free Frame-Based Conjugate Gradients Method*, Report UCDMS2002-07, Department of Mathematics and Statistics, University of Canterbury, New Zealand, 2002.
[9] C. Davis, *Theory of positive linear dependence*, Amer. J. Math., 76 (1954), pp. 733–746.
[10] U. M. García-Palomares and J. F. Rodríguez, *New sequential and parallel derivative-free algorithms for unconstrained minimization*, SIAM J. Optim., 13 (2002), pp. 79–96.
[11] P. Gilmore and C. T. Kelley, *An implicit filtering algorithm for optimization of functions with many local minima*, SIAM J. Optim., 5 (1995), pp. 269–285.
[12] R. M. Lewis and V. Torczon, *Pattern search algorithms for bound constrained minimization*, SIAM J. Optim., 9 (1999), pp. 1082–1099.
[13] R. M. Lewis and V. Torczon, *Pattern search methods for linearly constrained minimization*, SIAM J. Optim., 10 (2000), pp. 917–941.
[14] C. J. Price and I. D. Coope, *Frame based ray search algorithms in unconstrained optimization*, J. Optim. Theory Appl., 116 (2003), pp. 359–377.
[15] C. J. Price, I. D. Coope, and D. Byatt, *A convergent variant of the Nelder–Mead algorithm*, J. Optim. Theory Appl., 113 (2002), pp. 5–19.
[16] J. van Tiel, *Convex Analysis*, John Wiley and Sons, 1984.
[17] V. Torczon, *On the convergence of pattern search algorithms*, SIAM J. Optim., 7 (1997), pp. 1–25.
[18] W.-C. Yu and Y.-X. Li, *A direct search method by the local positive basis for linearly constrained optimization*, Chinese Ann. Math., 2 (1981), pp. 139–145.
[19] S. K. Zavriev, *On the global optimization properties of finite difference local descent algorithms*, J. Global Optim., 3 (1993), pp. 67–78.

SIAM J. OPTIM.
Vol. 14, No. 2, pp. 439–463

SOLVING KARUSH–KUHN–TUCKER SYSTEMS VIA THE TRUST REGION AND THE CONJUGATE GRADIENT METHODS*

HOUDUO QI†, LIQUN QI‡, AND DEFENG SUN§

Abstract. A popular approach to solving the Karush–Kuhn–Tucker (KKT) system, mainly arising from the variational inequality problem, is to reformulate it as a constrained minimization problem with simple bounds. In this paper, we propose a trust region method for solving the reformulation problem with the trust region subproblems being solved by the truncated conjugate gradient (CG) method, which is cost effective. Other advantages of the proposed method over existing ones include the fact that a good approximated solution to the trust region subproblem can be found by the truncated CG method and is judged in a simple way; also, the working matrix in each iteration is H, instead of the condensed H^TH, where H is a matrix element of the generalized Jacobian of the function used in the reformulation. As a matter of fact, the matrix used is of reduced dimension. We pay extra attention to ensure the success of the truncated CG method as well as the feasibility of the iterates with respect to the simple constraints. Another feature of the proposed method is that we allow the merit function value to be increased at some iterations to speed up the convergence. Global and superlinear/quadratic convergence is shown under standard assumptions. Numerical results are reported on a subset of problems from the MCPLIB collection [S. P. Dirkse and M. C. Ferris, *Optim. Methods Softw.*, 5 (1995), pp. 319–345].

Key words. variational inequality problem, constrained optimization, semismooth equation, trust region method, truncated conjugate gradient method, global and superlinear convergence

AMS subject classifications. 65H10, 90C30, 90C33

DOI. 10.1137/S105262340038256X

1. Introduction.

Given a continuously differentiable function $F : \mathbb{R}^n \to \mathbb{R}^n$ and twice continuously differentiable functions $h : \mathbb{R}^n \to \mathbb{R}^p$ and $g : \mathbb{R}^n \to \mathbb{R}^m$, we consider the following Karush–Kuhn–Tucker (KKT) system in (x, y, z):

$$\left.\begin{aligned} L(x,y,z) &= 0 \\ h(x) &= 0 \\ g(x) \geq 0, z \geq 0, z^T g(x) &= 0 \end{aligned}\right\}, \tag{1}$$

where L is called the Lagrangian of the functions F, g, and h and is defined by

$$L(x,y,z) := F(x) + \nabla h(x)y - \nabla g(x)z.$$

Due to its close relationship with the variational inequality problem (VIP) and the nonlinear constrained optimization problem (NLP) (in both cases, the functions h and g define the corresponding equality and inequality constraints, respectively), there is a growing interest in constructing efficient algorithms for (1); for the latest references, see [30, 11, 19]. In particular, Qi and Jiang [30] reformulate (1) to various

*Received by the editors December 15, 2000; accepted for publication (in revised form) April 30, 2003; published electronically October 14, 2003.
http://www.siam.org/journals/siopt/14-2/38256.html

†School of Mathematics, The University of New South Wales, Sydney 2052, Australia (hdqi@maths.unsw.edu.au). The research of this author was supported by the Australian Research Council.

‡Department of Applied Mathematics, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong (maqilq@polyu.edu.hk). The research of this author was supported by the Hong Kong Research Grant Council and the Australian Research Council.

§Department of Mathematics, National University of Singapore, 2 Science Drive 2, Singapore 117543, Republic of Singapore (matsundf@nus.edu.sg). The research of this author was supported by grant R146-000-035-101 of the National University of Singapore.

semismooth equations, and local semismooth Newton methods are studied for these semismooth equations. One of these semismooth equations is based on the Fischer–Burmeister function [12]: $\varphi : \mathbb{R}^2 \to \mathbb{R}$ defined by $\varphi(a,b) := (a+b) - \sqrt{a^2+b^2}$. An interesting property of φ_α is that $\varphi(a,b) = 0$ if and only if $a, b \geq 0, ab = 0$. Define

$$\phi(g(x), z) := (\varphi(g_1(x), z_1), \ldots, \varphi(g_m(x), z_m))^T \in \mathbb{R}^m,$$

and let $\Phi : \mathbb{R}^n \times \mathbb{R}^p \times \mathbb{R}^m \to \mathbb{R}^n \times \mathbb{R}^p \times \mathbb{R}^m$ be the equation operator

$$\Phi(w) := \Phi(x,y,z) := \begin{pmatrix} L(x,y,z) \\ h(x) \\ \phi(g(x),z) \end{pmatrix}.$$

Then $w^* = (x^*, y^*, z^*) \in \mathbb{R}^n \times \mathbb{R}^p \times \mathbb{R}^m$ is a solution of (1) if and only if it solves the system of nonlinear equations $\Phi(w) = 0$. In other words, solving (1) is equivalent to finding a global solution of the problem

$$\min \ \Psi(w), \tag{2}$$

where

$$\Psi(w) := \frac{1}{2}\Phi(w)^T\Phi(w) = \frac{1}{2}\|\Phi(w)\|^2$$

denotes the natural merit function of the equation operator Φ. This unconstrained optimization approach has been used in [8, 9, 30] to develop some Newton-type methods for the solution of (1). Despite their strong theoretical and numerical properties, these methods may fail to find the unique solution of (1) arising from strongly monotone variational inequalities because the variable z is not forced to be nonnegative in [8, 9, 30]. For such an example, see [26, 11]. This, together with the fact that the variable z has to be nonnegative at a solution of (1), motivates Facchinei et al. [11] to investigate a quadratic programming (QP) based method for the solution of the constrained minimization problem

$$\min \ \Psi(w) \quad \text{subject to (s.t.)} \quad z \geq 0. \tag{3}$$

The subproblem in $d = (d_x, d_y, d_z) \in \mathbb{R}^n \times \mathbb{R}^p \times \mathbb{R}^m$ solved at the current iteration $w^k = (x^k, y^k, z^k)$ (given $z^k \geq 0$) is of the type

$$\begin{array}{ll} \min & \nabla\Psi(w^k)^T d + \dfrac{1}{2} d^T (H_k^T H_k + \rho_k I) d \\ \text{s.t.} & z^k + d_z \geq 0, \end{array} \tag{4}$$

where $\rho_k > 0$, $H_k \in \partial\Phi(w^k)$, and $\partial\Phi(w^k)$ is the set of the generalized Jacobian of Φ at w^k in the sense of Clarke [4]. An inexact version of this QP-based method was provided by Kanzow [17] for the nonlinear complementarity problem (NCP) with an inexact solution of the QP subproblem being calculated by an interior-point method and the inexactness being measured in a similar way as described by Gabriel and Pang [14]. We emphasize that it is the interior-point method that guarantees the constraints to be nonviolated. Using an active-set strategy, Kanzow and Qi [19] proposed a QP-free method, which requires solving one system of linear equations rather than a QP problem per iteration and enjoys the favorable property that all iterates remain feasible with respect to (3). These two properties are also shared in a feasible equation-based method recently proposed by Kanzow [18]. We note that all of these methods are of the line-search type, and the superlinear/quadratic convergence

of these methods when applied to (3) requires that all the elements in $\partial\Phi(w^*)$ be nonsingular, where w^* is a solution of (1). Such a solution is usually called a strongly regular solution of (1).

There are several semismooth equation-based trust region methods which can be used to solve (1). The unconstrained trust region methods in [16, 20] are applicable to (2), while the box-constrained trust region method in [37] can be adapted to solve (3). Each of these methods requires at each iteration either an exact solution of the trust region subproblem [16]; a solution restricted to a (very small) subspace [20]; or an inexact solution which should satisfy some prescribed accuracy [37]. To be more precise, for global convergence only, the subspace in [20] can be as small as one-dimensional (i.e., spanned by the gradient direction), while for Ulbrich's method, any inexact solution satisfying the fraction of Cauchy decrease condition is enough (i.e., the affinely scaled gradient is a candidate). For local convergence, [20] requires that the subspace contain the generalized Newton direction, whereas [37] requires one to use the inexact generalized Newton direction. The use of iterative methods such as the conjugate gradient (CG) method is attractive because, on the one hand, first direction used in the CG method is the gradient direction, and quite often (e.g., when the trust region radius is larger than the length of the generalized Newton direction) the CG method yields an inexact generalized Newton direction and hence often speeds up the convergence process; on the other hand, when the number of variables is large, it is cost effective by the CG method to solve the trust region subproblem approximately. The key issue of efficiently implementing the CG method is the *preconditioning*. Although it is understood that no single preconditioning is "best" for all conceivable types of matrices, we will use the symmetric successive overrelaxation (SSOR) preconditioner in our numerical experiments. We will discuss it more in our numerical implementations.

In this paper we study how to apply the truncated CG method to a trust region subproblem of (3) so as to keep the computational cost at a reasonable level, and we show how to merge the truncated CG method with the semismooth Newton method as the iterates of our trust region method approach a minimizer, so that the use of the truncated CG method does not slow down the fast convergence of the proposed trust region method. Another favorable consequence of using the truncated CG method is that, although the quadratic term in the subproblem is constructed with $d^T H_k^T H_k d$, the calculation process of an approximation to the solution of the subproblem works directly on the matrix H_k, not on the usually condensed matrix $H_k^T H_k$. In addition, all iterates of our method remain feasible with respect to the simple bounds in (3), and the trust region subproblem is in a reduced form. This latter property is essential for the success of the truncated CG method and is guaranteed by incorporating an active-set strategy into the proposed trust region method. Finally, the superlinear/quadratic convergence of the proposed method is established under the assumption of nonsingularity.

The truncated CG method was first used by Toint [36] and Steihaug [33] to solve the trust region subproblem for unconstrained optimization problems and is shown to be efficient, especially in large scale optimization. Let $f : \mathbb{R}^n \to \mathbb{R}$ be continuously differentiable. Then the usual trust region subproblem for the unconstrained optimization problem $\min_{x\in\mathbb{R}^n} f(x)$ is

$$\begin{aligned} \min \quad & \phi(d) = g^T d + \frac{1}{2} d^T B d \\ \text{s.t.} \quad & \|d\| \le \Delta, \end{aligned} \tag{5}$$

where $\Delta > 0$ is a trust region bound, $g \in \mathbb{R}^n$ is the gradient of the objective function f at the current iterate, and $B \in \mathbb{R}^{n\times n}$ is symmetric and is an approximation to the Hessian of $f(x)$. Although the truncated CG method is widely used in practice, it was only recently proved that it indeed provides a sufficient decrease in the objective function for the case of strict convexity. In fact, when B is positive definite, Yuan[1] proved in [38] that the reduction in the objective function by the truncated CG method is at least half of the reduction by the global minimizer in the trust region. However, this result may be invalid when the bound constraint $x + d \geq 0$ is preserved. This can be shown by the following example. Consider the strictly convex problem in $\mathbb{R}^2$: $\min_{x\geq 0}\ x_1 + .5x_1^2 - x_2 + .5x_2^2$. Obviously, $(0, 1)$ is the unique solution. The trust region subproblem at $x = (\varepsilon, \varepsilon)$ is

$$\begin{array}{ll} \min & (1+\varepsilon)d_1 + (-1+\varepsilon)d_2 + \dfrac{1}{2}d_1^2 + \dfrac{1}{2}d_2^2 \\ \text{s.t.} & \|d\| \leq \Delta,\ x + d \geq 0. \end{array}$$

The truncated CG method for solving this subproblem first generates a direction by ignoring the bound constraint and then takes a small enough step along this direction to ensure feasibility. Hence, the next iterate is $x^1 = (0, 2\varepsilon/(1+\varepsilon))$ ($\Delta = 1$ and $\varepsilon \in (0, 1)$). We continue to build the trust region subproblem around x^1; this time, the truncated CG method results in a direction which immediately goes infeasible, leading to the zero steplength. In other words, the truncated CG method fails to solve the strictly convex problem. The reason is that very small components of x may result in a very small (even zero) steplength. One way to avoid the collapse of the steplength is to build the trust region subproblem only around those components of the current iterate which are relatively large enough, while paying a special attention to the smaller ones.

The paper realizes the above ideas with a trust region method, which is solved by a truncated CG method. The paper is organized as follows. Some background is summarized in the next section. The subproblem is derived in section 3. A truncated CG method for this subproblem is introduced in section 4. Our algorithm is presented in section 5. Global and local convergence results are established in sections 6 and 7, respectively. Numerical results on a subset of problems from the MCPLIB collection [7] are presented in section 8. Finally, some conclusions are drawn in section 9.

2. Mathematical background.

2.1. Notation. A function $G : \mathbb{R}^t \to \mathbb{R}^t$ is called a C^k function if it is k times continuously differentiable, and an LC^k function if it is a C^k function and its kth derivative is locally Lipschitz continuous everywhere. The Jacobian of a C^1 function G at a point $w \in \mathbb{R}^t$ is denoted by $G'(w)$, whereas $\nabla G(w)$ is the transposed Jacobian. This notation is consistent with our notation of a gradient vector $\nabla g(w)$ for a real-valued function $g : \mathbb{R}^t \to \mathbb{R}$ since we view $\nabla g(w)$ as a column vector.

If $M \in \mathbb{R}^{t\times t}$, $M = (m_{ij})$, is any given matrix and $I, J \subseteq \{1, \ldots, t\}$ are two subsets, then M_{IJ} denotes the $|I| \times |J|$ submatrix with elements $m_{ij}, i \in I, j \in J$. Similarly, $M_{.J}$ indicates the submatrix with elements $m_{ij}, i \in \{1, \ldots, t\}, j \in J$; i.e., we obtain $M_{.J}$ from M by removing all columns with indices $j \notin J$. Similar notation is used for subvectors. If $w = (x^T, y^T, z^T)^T \in \mathbb{R}^n \times \mathbb{R}^p \times \mathbb{R}^m$, we often simplify our

[1] Yuan attributes to P. Tseng a slightly weaker result that the reduction of the objective function by the truncated CG method is at least 1/3 (instead of 1/2) of the reduction by the global minimizer. And Yuan's result for the positive definite case is generalized to the positive semidefinite case in [6].

notation and write $w = (x, y, z)$. All vector norms used in this paper are Euclidean norms, and matrix norms are the 2-norms of the matrices. For a given symmetric positive definite matrix C, a norm induced by C is defined by $\|x\|_C := \sqrt{x^T C x}$.

2.2. Properties of the reformulation. Our analysis will make frequent use of some properties on the generalized Jacobian $\partial\Phi(w)$ (in the sense of Clarke [4]). An explicit formula of calculating an element from $\partial\Phi(w)$ is given in [9]. We will discuss more about it in the section of numerical experiments. Although Φ itself is not continuously differentiable in general, its square norm Ψ is continuously differentiable and

$$\nabla\Psi(w) = H^T\Phi(w) \qquad \forall\, H \in \partial\Phi(w). \tag{6}$$

Another favorable property of the equation operator Φ is that the nonsingularity of its generalized Jacobian is ensured by Robinson's strong regularity condition. We formally state this result as the following. For the precise definition, some further characterizations, and sufficient conditions for the strong regularity, we refer the reader to Robinson [32] as well as to Liu [22].

PROPOSITION 2.1 (see [9]). *A point $w^* = (x^*, y^*, z^*) \in \mathbb{R}^n \times \mathbb{R}^p \times \mathbb{R}^m$ is a strongly regular solution of* (1) *if and only if all elements in the generalized Jacobian $\partial\Phi(w^*)$ are nonsingular.*

The next property follows from the fact that Φ is a (strongly) semismooth operator under certain smoothness assumptions for F, h, and g; see, e.g., [29, 31, 25, 13].

PROPOSITION 2.2. *For any $w = (x, y, z) \in \mathbb{R}^n \times \mathbb{R}^p \times \mathbb{R}^m$, we have*

$$\|\Phi(w + d) - \Phi(w) - Hd\| = o(\|d\|) \qquad \text{for } d \to 0 \text{ and } H \in \partial\Phi(w + d).$$

If F is an LC^1 mapping, and h and g are LC^2 mappings, then

$$\|\Phi(w + d) - \Phi(w) - Hd\| = O(\|d\|^2) \qquad \text{for } d \to 0 \text{ and } H \in \partial\Phi(w + d).$$

An immediate consequence of the strong regularity of w^* and the semismoothness of Φ is that the function value $\|\Phi(w)\|$ provides a local error bound near w^*; see, e.g., [29, 25].

PROPOSITION 2.3. *Assume that w^* is a strongly regular solution of* (1). *Then there are constants $c_1 > 0$ and $\delta_1 > 0$ such that*

$$\|\Phi(w)\| \geq c_1\|w - w^*\|$$

for all w with $\|w - w^*\| \leq \delta_1$.

3. Subproblem. Given a current iterate $w^k = (x^k, y^k, z^k) \in \mathbb{R}^n \times \mathbb{R}^p \times \mathbb{R}^m$ with $z^k \geq 0$, a traditional choice of the trust region subproblem for (3) is

$$\begin{aligned} \min \quad & \nabla\Psi(w^k)^T d + \frac{1}{2} d^T H_k^T H_k d \\ \text{s.t.} \quad & \|d\| \leq \Delta_k,\ z^k + d_z \geq 0, \end{aligned}$$

where $H_k \in \partial\Phi(w^k)$ and Δ_k is the current trust region radius. In order to make the truncated CG method successful with this subproblem, small components, as we observed in the introduction, should be detected and not involved in this subproblem, and the matrix $H_k^T H_k$ should be regularized to be positive definite in order for our algorithm to be well defined.

To make the idea precise, we introduce three index sets,

$$\begin{aligned}\mathcal{I} &:= \{1, \ldots, n\},\\ \mathcal{P} &:= \{n+1, \ldots, n+p\},\\ \mathcal{J} &:= \{n+p+1, \ldots, n+p+m\},\end{aligned}$$

where $\mathcal{I}$ denotes the index set for the variables x, $\mathcal{P}$ is the index set for the equality constraints and the variables y, and $\mathcal{J}$ is the index set for the inequality constraints and the variables z. For example, if $w = (x, y, z) \in \mathbb{R}^n \times \mathbb{R}^p \times \mathbb{R}^m$ is any given vector, then $w_{\mathcal{I}} = x$, $w_{\mathcal{P}} = y$, and $w_{\mathcal{J}} = z$. We also stress that if $j \in \mathcal{J}$ or $J \subseteq \mathcal{J}$, then w_j is a component of the z-part of the vector w and w_J is a subvector of the z-part of w.

To detect the small components of z^k, we introduce at each iteration an indicator $\delta_k > 0$ and the index set

$$J_k := \{j \in \mathcal{J} |\ w_j^k \le \delta_k\}, \tag{7}$$

which contains all indices whose corresponding components in z^k are thought to be small. We shall give it special attention in our algorithm since the truncated CG method may fail depending on it. Let

$$\bar{J}_k := \mathcal{I} \cup \mathcal{P} \cup (\mathcal{J} \setminus J_k). \tag{8}$$

In order to make the matrix $H_k^T H_k$ positive definite, we need a continuous function $\rho : \mathbb{R} \to \mathbb{R}$ with the following properties: (i) $\rho(s) \ge 0$ for all $s \in \mathbb{R}$, and (ii) $\rho(s) = 0$ if and only if $s = 0$. Such a function is usually called a forcing function.

From now on, we will often abbreviate the gradient vector $\nabla\Psi(w^k)$ by g^k (in contrast to $g(x^k)$, which denotes the function value of the inequality constraints at the current point x^k, so there should be no ambiguity). Partition $H_k = (H^k_{\cdot J_k}, H^k_{\cdot \bar{J}_k})$. We build our trust region subproblem only around the components of $w^k_{\bar{J}_k}$; that is,

$$\begin{array}{ll}\min & (g^k_{\bar{J}_k})^T d_{\bar{J}_k} + \dfrac{1}{2} d^T_{\bar{J}_k} \left((H^k_{\cdot \bar{J}_k})^T H^k_{\cdot \bar{J}_k} + \rho(\Psi(w^k)) I \right) d_{\bar{J}_k}\\ \text{s.t.} & \|d_{\bar{J}_k}\| \le \Delta_k,\ w^k_{\mathcal{J} \setminus J_k} + d_{\mathcal{J} \setminus J_k} \ge 0.\end{array} \tag{9}$$

It is clear that the above subproblem is a strictly convex quadratic problem if $\Psi(w^k) \neq 0$.

4. Truncated CG method. In this section, we adapt the truncated preconditioned conjugate gradient (PCG) method described in [33] to our subproblem, which is a $q := (n + p + m - |J_k|)$-dimensional convex problem.

Suppose a symmetric positive definite matrix $C \in \mathbb{R}^{q \times q}$ is given with decomposition property $C = P^T P$, where P is nonsingular. For simplicity we denote the variable $d_{\bar{J}_k}$ by s. Consider the preconditioned version of (9):

$$\begin{array}{ll}\min_{s \in \mathbb{R}^q} & m_k(s) := s^T b + \dfrac{1}{2} s^T \mathcal{B}_k s\\ \text{s.t.} & \|s\|_C \le \Delta_k,\ w^k_{\mathcal{J} \setminus J_k} + s_{\mathcal{J} \setminus J_k} \ge 0,\end{array} \tag{10}$$

where $b := g^k_{\bar{J}_k}$, $A := H^k_{\cdot \bar{J}_k}$, $\sigma := \rho(\Psi(w^k))$, and $\mathcal{B}_k := A^T A + \sigma I$. Taking into consideration the special structure of $\mathcal{B}_k$ and the decomposition property of the preconditioner C, we arrive at the following truncated PCG method for problem (10).

ALGORITHM 4.1 (truncated PCG method).

(S.0) *Let* $s^0 = 0$, $r^0 = b$, $\tilde{r}^0 = P^{-T}r^0$, $p^0 = -\tilde{r}^0$, $i := 0$.

(S.1) *If* $\|\nabla m_k(s^i)\| = 0$, *then set* $s^* = s^i$ *and go to* (S.4). *Otherwise calculate*

$$t^i = P^{-1}p^i,\ q^i = At^i,\ \text{and } \alpha_i = \|\tilde{r}^i\|^2/(\|q^i\|^2 + \sigma\|t^i\|^2).$$

(S.2) *If* $\|s^i + \alpha_i t^i\|_C \geq \Delta_k$, *then go to* (S.3). *Otherwise set*

$$s^{i+1} := s^i + \alpha_i t^i,\ r^{i+1} := r^i + \alpha_i(\sigma t^i + A^T q^i),$$
$$\tilde{r}^{i+1} := P^{-T}r^{i+1},\ \beta_i := \|\tilde{r}^{i+1}\|^2/\|\tilde{r}^i\|^2,\ p^{i+1} := -\tilde{r}^{i+1} + \beta_i p^i.$$

Set $i := i + 1$, *and go to* (S.1).

(S.3) *Calculate* $\alpha_i^* \geq 0$ *satisfying* $\|s^i + \alpha_i^* t^i\|_C = \Delta_k$; *set* $s^* := s^i + \alpha_i^* t^i$.

(S.4) *Compute the largest* $\tau_k \geq 0$ *satisfying* $w^k_{\mathcal{J}\setminus J_k} + \tau s^*_{\mathcal{J}\setminus J_k} \geq 0$ *for all* $\tau \in (0, \tau_k]$.

(S.5) *Output the approximate solution to* (10): $d^k_{\bar{J}_k} = \min\{1, \tau_k\}s^*$.

Remarks. First, we note that the computation of s^* above is exactly Steihaug's algorithm [33]. Second, in each iteration of the PCG method, there are two matrix-vector multiplications involving A (to get At^i and $A^T(At^i)$) and two matrix-inverse multiplications involving P (to get $P^{-T}r^i$ and $P^{-1}(P^{-T}r^i)$). If we choose C to be the SSOR preconditioner, we shall see in section 8 by referring to several specific references that the usually condensed matrix $\mathcal{B}_k$ is not involved in the calculation. In fact, only nonzero elements in A come to be used. Third, the termination rule in (S.1) is only for theoretical purposes. We shall use a more practical criterion in our implementation.

The vector s^* generated above has a close relation to the exact solution of the following problem:

$$\min\ m_k(d_{\bar{J}_k}) \qquad \text{s.t. } \|d_{\bar{J}_k}\|_C \leq \Delta_k. \tag{11}$$

This relation follows from a recent result of Yuan [38, Thm. 2].

PROPOSITION 4.2. *Suppose that* $\mathcal{B}_k$ *is positive definite. Let* $d^*_{\bar{J}_k}$ *be the exact solution of* (11) *and* s^* *be generated by Algorithm* 4.1. *Then we have*

$$m_k(s^*) \leq \frac{1}{2}m_k(d^*_{\bar{J}_k}).$$

Moreover, if $\|d^*_{\bar{J}_k}\|_C < \Delta_k$, *then* $s^* = d^*_{\bar{J}_k}$.

Proof. As remarked above, the calculation of s^* in Algorithm 4.1 is actually Steihaug's algorithm applied to the subproblem (11). Let $y = Ps$ and $y^* = Ps^*$. Then y^* is the point obtained by applying the truncated CG method to the following trust region problem:

$$\begin{array}{ll} \min_{y\in\mathbb{R}^q} & \tilde{m}_k(y) := y^T(P^{-T}b) + \dfrac{1}{2}y^T(P^{-T}\mathcal{B}_k P^{-1})y \\ \text{s.t.} & \|y\| \leq \Delta_k. \end{array} \tag{12}$$

Let $\tilde{y}^*$ be the unique solution of (12). According to a result of Yuan [38, Thm. 2], it holds that $\tilde{m}_k(y^*) \leq \frac{1}{2}\tilde{m}_k(\tilde{y}^*)$ and $y^* = \tilde{y}^*$ if $\|\tilde{y}^*\| < \Delta_k$. We note that $\tilde{m}_k(y^*) = m_k(s^*)$ and $\tilde{m}_k(\tilde{y}^*) = m_k(d^*_{\bar{J}_k})$. Then the inequality relation in the proposition follows. Moreover, $s^* = d^*_{\bar{J}_k}$ if $\|d^*_{\bar{J}_k}\|_C < \Delta_k$. ∎

In the following as well as in our convergence analysis, we assume that $C_k = I$ for simplicity, where C_k is the preconditioner in (10) at each iteration. However, to keep

all convergence results in sections 6 and 7 valid, we need more assumptions on the preconditioner sequence $\{C_k\}$; namely, there exist two constants $\underline{\kappa}$ and $\bar{\kappa}$ such that, for all k,

$$\|C_k^{-1}\| \le \underline{\kappa} \ \text{ and } \ \|C_k\| \le \bar{\kappa}. \tag{13}$$

This condition implies that C_k has a condition number that is bounded independent of the iterate. The latter condition has been singled out in [21, p. 1120] to emphasize its theoretical importance in a trust region method. We note that the two conditions are equivalent under boundedness of the whole iterate sequence $\{w^k\}$. Standard convergence proofs involving condition (13) (in the special case that C_k is diagonal) can be found in [5] (proofs leading up to Theorem 11 there). We also note that we use the Hessian matrix H_k in building up $\mathcal{B}_k$, implying that the standard condition restricted on $\mathcal{B}_k$ in trust region methods is fulfilled automatically in our setting. Proposition 4.2 allows us to put a bound on the predicted decrease $m_k(d^k_{\bar{J}_k})$. Two bounds are given in the following result. The first corresponds to the case where $\tau_k \ge 1$ so that $d^k_{\bar{J}_k} = s^*$; the second corresponds to the case where $\tau_k < 1$ so that $d^k_{\bar{J}_k} \ne s^*$.

PROPOSITION 4.3. *Let $d^k_{\bar{J}_k}$ and s^* be generated by Algorithm* 4.1 *applied to* (9), *and define*

$$\Omega_k := \{j \in \mathcal{J} \setminus J_k | - s^*_j > w^k_j\}.$$

Then the following statements hold:

(i) *If Ω_k is empty (in particular if $\delta_k \ge \Delta_k$), then we have*

$$m_k(d^k_{\bar{J}_k}) \le -\frac{1}{4}\|g^k_{\bar{J}_k}\| \min\left\{\Delta_k, \frac{\|g^k_{\bar{J}_k}\|}{\|\mathcal{B}_k\|}\right\}.$$

(ii) *If $\Omega_k \ne \emptyset$, then*

$$m_k(d^k_{\bar{J}_k}) \le -\frac{\delta_k}{4\Delta_k}\|g^k_{\bar{J}_k}\| \min\left\{\Delta_k, \frac{\|g^k_{\bar{J}_k}\|}{\|\mathcal{B}_k\|}\right\}.$$

Proof. (i) Let d^* be the unique solution to (9). Then it follows from [27, Thm. 4] that

$$m_k(d^*) \le -\frac{1}{2}\|g^k_{\bar{J}_k}\| \min\left\{\Delta_k, \frac{\|g^k_{\bar{J}_k}\|}{\|\mathcal{B}_k\|}\right\}. \tag{14}$$

Also by simple calculation we have $\tau_k \ge 1$ if $\Omega_k = \emptyset$. This is also true in particular if $\delta_k \ge \Delta_k$. Hence $d^k_{\bar{J}_k} = s^*$ is also generated by Yuan's truncated CG method [38] applied to (9) with the simple constraints not being violated. Therefore, Proposition 4.2 implies

$$m_k(d^k_{\bar{J}_k}) = m_k(s^*) \le \frac{1}{2} m_k(d^*). \tag{15}$$

The combination of (15) and (14) gives the result in (i).

(ii) Let $j \in \Omega_k$. Then $\tau_k < 1$ and

$$\Delta_k \ge -s^*_j > w^k_j > \delta_k.$$

Hence

$$d^k_{\bar{J}_k} = \tau_k s^* \quad \text{and} \quad \tau_k \geq \delta_k/\Delta_k.$$

Now we consider the one-dimensional function

$$\mathcal{M}(t) := m_k(ts^*) = t(g^k_{\bar{J}_k})^T s^* + \frac{1}{2}t^2 (s^*)^T \mathcal{B}_k s^*.$$

Let

$$\tau_* = \frac{-(g^k_{\bar{J}_k})^T s^*}{(s^*)^T \mathcal{B}_k s^*}.$$

It follows from the fact that $\mathcal{M}(1) = m_k(s^*) \leq 0$ that $2\tau_* \geq 1$. Now we consider two cases. First if $\tau_k \leq \tau_*$, the convexity of $\mathcal{M}$ implies that

$$\begin{aligned} m_k(d^k_{\bar{J}_k}) &= \mathcal{M}(\tau_k) \leq \mathcal{M}(\delta_k/\Delta_k) \\ &= \frac{\delta_k}{\Delta_k}(g^k_{\bar{J}_k})^T s^* + \frac{1}{2}\left(\frac{\delta_k}{\Delta_k}\right)^2 (s^*)^T \mathcal{B}_k s^* \\ &\leq \frac{\delta_k}{\Delta_k}\mathcal{M}(1) = \frac{\delta_k}{\Delta_k} m_k(s^*). \end{aligned} \tag{16}$$

If $\tau_k \geq \tau_*$, then it is easy to see from the convexity of $\mathcal{M}$ again that

$$m_k(d^k_{\bar{J}_k}) = \mathcal{M}(\tau_k) \leq \mathcal{M}(1) = m_k(s^*). \tag{17}$$

Now the result in (ii) follows from (14), (16), (17), and Proposition 4.2. □

5. Algorithm. Suppose $\Delta_k, w^k, J_k, \bar{J}_k$, and the function $m_k(\cdot)$ are given as in the last section, and a search direction $\tilde{d}^k$ is partitioned as

$$\tilde{d}^k = \begin{pmatrix} \tilde{d}^k_{J_k} \\ \tilde{d}^k_{\bar{J}_k} \end{pmatrix}.$$

Let the ratio between the actual decrease and the predicted decrease associated with the direction $\tilde{d}^k$ be calculated by

$$r_k := \left(\Psi(w^k) - \Psi(w^k + \tilde{d}^k)\right)/\mathrm{Pred}_k, \tag{18}$$

where

$$\mathrm{Pred}_k := -(g^k_{J_k})^T \tilde{d}^k_{J_k} - m_k(\tilde{d}^k_{\bar{J}_k}).$$

The update rule used in our trust region algorithm is as follows:

$$w^{k+1} := \begin{cases} w^k & \text{if } r_k < \rho_1, \\ w^k + \tilde{d}^k & \text{if } r_k \geq \rho_1, \end{cases} \qquad \Delta_{k+1} := \begin{cases} \sigma_1 \Delta_k & \text{if } r_k < \rho_1, \\ \max\{\Delta_{\min}, \Delta_k\} & \text{if } r_k \in [\rho_1, \rho_2), \\ \max\{\Delta_{\min}, \sigma_2 \Delta_k\} & \text{if } r_k \geq \rho_2, \end{cases} \tag{19}$$

where $\Delta_{\min} > 0$ is a prescribed constant. The proposed trust region algorithm is then formally stated below.

ALGORITHM 5.1 (trust region algorithm).

(S.0) *Choose* $w^0 = (x^0, y^0, z^0) \in \mathbb{R}^n \times \mathbb{R}^p \times \mathbb{R}^m$ *with* $z^0 \geq 0$, $\Delta_0 > 0$, $0 < \rho_1 < \rho_2 < 1$, $0 < \sigma_1 < 1 < \sigma_2$, $\Delta_{\min} > 0$, $\gamma \in (0,1), c > 0$, $\delta > 0$, $\epsilon > 0$, $\mathrm{ind}_0 := 0$, $\beta_0 = 0$, *and set* $k := 0$.

(S.1) *Let*

$$\delta_k := \min\left\{\delta, c\sqrt{\|\Phi(w^k)\|}\right\}$$

and define the set J_k *and* $\bar{J}_k$ *by* (7) *and* (8), *respectively.*

(S.2) *Let*

$$v^k := \begin{pmatrix} v^k_{J_k} \\ v^k_{\bar{J}_k} \end{pmatrix},$$

where

$$v^k_{J_k} = \min\{w^k_{J_k}, g^k_{J_k}\} \quad \textit{and} \quad v^k_{\bar{J}_k} = g^k_{\bar{J}_k}.$$

If $\|v^k\| \leq \epsilon$, *stop.*

(S.3) *Choose preconditioner* C_k *and let* $d^k_{\bar{J}_k}$ *be the final iterate of Algorithm* 4.1 *for the trust region subproblem* (10).

(S.4) *Compute the search directions*

$$d^k := \begin{pmatrix} -w^k_{J_k} \\ d^k_{\bar{J}_k} \end{pmatrix} \quad \textit{and} \quad \tilde{d}^k := \begin{pmatrix} -\min\{1, \Delta_k\} v^k_{J_k} \\ d^k_{\bar{J}_k} \end{pmatrix}.$$

(S.5) (i) *If* $\mathrm{ind}_k = 0$, *check if the following rule holds:*

$$\Psi(w^k + d^k) \leq \gamma\sqrt{\|\Phi(w^k)\|}. \tag{20}$$

If test (20) *is successful, then let*

$$w^{k+1} := w^k + d^k, \quad \Delta_{k+1} := \max\{\Delta_{\min}, \sigma_2\Delta_k\}, \quad \gamma_{k+1} := \frac{\Psi(w^{k+1})}{\Psi(w^k)}$$

and

$$\bar{\gamma} := \gamma_{k+1} \textit{ if } \gamma_{k+1} \geq \gamma, \quad \beta_{k+1} := \begin{cases} \Psi(w^{k+1}) & \textit{if } \gamma_{k+1} \geq \gamma, \\ \beta_k & \textit{if } \gamma_{k+1} < \gamma, \end{cases}$$

$$\mathrm{ind}_{k+1} := \begin{cases} 1 & \textit{if } \gamma_{k+1} \geq \gamma, \\ 0 & \textit{if } \gamma_{k+1} < \gamma. \end{cases}$$

If test (20) *is not successful, then calculate* r_k *by* (18), *update* w^{k+1} *and* Δ_{k+1} *according to rule* (19), *and let* $\beta_{k+1} := \beta_k$, $\mathrm{ind}_{k+1} := 0$.

(ii) *If* $\mathrm{ind}_k = 1$, *check if the following holds:*

$$\Psi(w^k + d^k) \leq \frac{\gamma}{\bar{\gamma}}\beta_k. \tag{21}$$

If test (21) *is successful, let*

$$w^{k+1} := w^k + d^k, \ \Delta_{k+1} := \max\{\Delta_{\min}, \sigma_2\Delta_k\}, \ \beta_{k+1} := \beta_k, \ \mathrm{ind}_{k+1} := 0.$$

If test (21) *is not successful, then calculate* r_k *by* (18), *update* w^{k+1} *and* Δ_{k+1} *according to the rule* (19), *and let* $\beta_{k+1} := \beta_k$, $\mathrm{ind}_{k+1} := 1$.

(S.6) *Set* $k := k + 1$ *and go to* (S.1).

More explanation on Algorithm 5.1 is as follows. The indicator δ_k defined in (S.1) was also used in [19, 18] to identify the actual active set under the nonsingularity assumption. A result due to Kanzow and Qi [19] justifies the termination rule in (S.2). We will present this result in a lemma below. The direction $d^k_{\bar{J}_k}$ in (S.3) has been extensively discussed in the last section. The crucial parts of Algorithm 5.1 are (S.4) and (S.5). In (S.4) two directions are defined. The direction $\tilde{d}^k$ (which we call safe step below) is always a descent direction of the function $\Psi(\cdot)$ at w^k if Δ_k is sufficiently small; while direction d^k (fast step) will yield a superlinear decrease in the function value of $\Psi(\cdot)$ when w^k is sufficiently close to a strongly regular solution w^*. The task in (S.5) is then to decide which step we should take. Fast steps are accepted in a nonmonotone fashion, so that it can happen that an accepted fast step increases the value of Ψ. To keep control over those possible increases, a flag `ind` is used to distinguish between two states of the algorithm: If $\texttt{ind} = 0$ (which is the case at the very beginning), the fast step d^k is accepted if test (20) is successful. Now, if $\Psi(w^k + d^k) \geq \gamma\Psi(w^k)$, the flag `ind` is set to 1 (and remains raised until it is cleared again) to signal that d^k did not achieve sufficient decrease. Now consider any iteration k that is entered with $\texttt{ind} = 1$, indicating that the most recent accepted fast step d^l $(l < k)$ did not achieve sufficient decrease. In this situation, the fast step d^k is accepted only if test (21) is successful. If this occurs, `ind` is set to 0 again. In all iterations where the fast step is not accepted, the safe step is used as the trial step of the trust region method with a standard reduction-ratio-based acceptance test. We stress that both $\bar{\gamma}$ and β_k are used to record the cases where (20) is successful and $\gamma_{k+1} \geq \gamma$, i.e., to record the cases where the function values are possibly increased.

From now on we assume that $\epsilon = 0$. The following result from [19, Lem. 1] justifies the termination criterion used in our trust region algorithm. We recall that a point $w^* = (x^*, y^*, z^*) \in \mathbb{R}^n \times \mathbb{R}^p \times \mathbb{R}^m$ with $z^* \geq 0$ is a stationary point of problem (3) if $\nabla_x \Psi(w^*) = 0$, $\nabla_y \Psi(w^*) = 0$, and

$$z_i^* > 0 \Longrightarrow \frac{\partial \Psi(w^*)}{\partial z_i} = 0, \quad z_i^* = 0 \Longrightarrow \frac{\partial \Psi(w^*)}{\partial z_i} \geq 0.$$

LEMMA 5.2. *Let $w^k = (x^k, y^k, z^k) \in \mathbb{R}^n \times \mathbb{R}^p \times \mathbb{R}^m$ be any given point with $z^k \geq 0$. Then the following holds:*

$$w^k \text{ is a stationary point of (3)} \iff v^k = 0 \iff (g^k)^T v^k = 0.$$

The following result shows that the iterates $\{w^k\}$ generated by Algorithm 5.1 stay feasible with respect to the simple bounds in (3).

LEMMA 5.3. *Let $w^k = (x^k, y^k, z^k) \in \mathbb{R}^n \times \mathbb{R}^p \times \mathbb{R}^m$ be any given point with $z^k \geq 0$, and assume that w^k is not a stationary point of (3). Then the next iterate w^{k+1} can be computed by Algorithm 5.1 and it holds that $z^{k+1} \geq 0$.*

Proof. Let w^k be given as in Lemma 5.3. We have no problem running steps (S.1)–(S.4) of Algorithm 5.1. Hence, for Algorithm 5.1 to be well defined, we need to show that (S.5) is well defined

Since w^k is not a stationary point of (3), $\Psi(w^k) > 0$ so that $\bar{\gamma}$ in (S.5)(i) is well defined if (20) is successful. If (20) does not hold, we need only to show $\text{Pred}_k \neq 0$ so that r_k is well defined. It follows from Proposition 4.3 that

$$m_k(d^k_{\bar{J}_k}) \leq 0 \quad \text{and} \quad m_k(d^k_{\bar{J}_k}) = 0 \iff g^k_{\bar{J}_k} = 0.$$

On the other hand we have, for $j \in J_k$,

$$g_j^k v_j^k = \begin{cases} (g_j^k)^2 & \text{if } \ g_j^k \le w_j^k, \\ g_j^k w_j^k & \text{if } \ g_j^k > w_j^k. \end{cases}$$

Noting that $w_j^k \ge 0$ for $j \in J_k$, we obtain

$$(g_{J_k}^k)^T v_{J_k}^k \ge 0.$$

Hence $\mathrm{Pred}_k = \min\{1, \Delta_k\}(g_{J_k}^k)^T v_{J_k}^k - m_k(d_{\bar{J}_k}^k) \ge 0$, and if $\mathrm{Pred}_k = 0$, we must have both $(g_{J_k}^k)^T v_{J_k}^k = 0$ and $g_{\bar{J}_k}^k = 0$, which means that $(g^k)^T v^k = (g_{J_k}^k)^T v_{J_k}^k + \|g_{\bar{J}_k}^k\|^2 = 0$. By Lemma 5.2, w^k must be a stationary point of (3), contradicting the assumption of this lemma. Hence $\mathrm{Pred}_k > 0$, and consequently r_k is well defined and w^{k+1} is obtained.

Now we prove $z^{k+1} \ge 0$. There are three possible ways to determine w^{k+1}, namely, $w^{k+1} = w^k$, $w^{k+1} = w^k + d^k$, or $w^{k+1} = w^k + \tilde{d}^k$. The first case is trivial, so we consider the remaining two cases. We note that the components w_j^{k+1}, $j \in \mathcal{J} \setminus J_k$, are updated by

$$w_j^{k+1} = w_j^k + (d_{\bar{J}_k}^k)_j = w_j^k + \min\{1, \tau_k\} s_j^* \ge \begin{cases} w_j^k & \text{if } \ s_j^* \ge 0, \\ w_j^k + \tau_k s_j^* & \text{if } \ s_j^* < 0, \end{cases}$$

where s^* is computed by Algorithm 4.1. It follows from (S.4) of Algorithm 4.1 that $w_j^{k+1} \ge 0$. For the components j belonging to J_k, if $w^{k+1} = w^k + d^k$, then $w_j^{k+1} = w_j^k - w_j^k = 0$; if $w^{k+1} = w^k + \tilde{d}^k$, then

$$\begin{aligned} w_j^{k+1} &= w_j^k - \min\{1, \Delta_k\} \min\{w_j^k, g_j^k\} \\ &\ge \begin{cases} w_j^k \ge 0 & \text{if } \ \min\{w_j^k, g_j^k\} \le 0, \\ w_j^k - w_j^k = 0 & \text{if } \ \min\{w_j^k, g_j^k\} > 0. \end{cases} \end{aligned}$$

This proves that $w_j^{k+1} \ge 0$ for all $j \in \mathcal{J}$. ∎

We note that as long as w^k is not a global minimizer of (3), Algorithm 4.1 is always successful for subproblem (9), and the estimation in Proposition 4.3 always holds since $\mathcal{B}_k$ is always positive definite. So we can apply an induction argument by invoking Lemma 5.3 and then obtain the following result.

THEOREM 5.4. *Algorithm* 5.1 *is well defined and generates a sequence* $\{w^k\} = \{(x^k, y^k, z^k)\}$ *with* $z^k \ge 0$ *for all* k.

6. Global convergence. From now on we assume that Algorithm 5.1 generates an infinite sequence $\{w^k\}$. Let K contain all the indices at which the function value is possibly increased; that is,

$$K := \left\{ k \in \{0, 1, 2, \ldots\} \mid \mathrm{ind}_k = 0, \ \Psi(w^k + d^k) \le \gamma \sqrt{\|\Phi(w^k)\|} \text{ and } \gamma_{k+1} \ge \gamma \right\}. \tag{22}$$

Then we have the following convergence result.

LEMMA 6.1. *Suppose that* K *contains infinitely many iterations. Then*

$$\lim_{k \to \infty} \Psi(w^k) = 0.$$

Hence, every limit point of $\{w^k\}$ *is a solution of* (1) *and therefore a stationary point of* (3).

Proof. Let us denote K by

$$K = \{k_0, k_1, k_2, \ldots\}.$$

In the following we want to prove

(a) the sequence $\{\Psi(w^k)\}_{k\in K}$ converges to zero, i.e., $\lim_{l\to\infty} \Psi(w^{k_l}) = 0$;

(b) the sequence $\{\Psi(w^k)\}_{(k-1)\in K}$ converges to zero, i.e., $\lim_{l\to\infty} \Psi(w^{k_l+1}) = 0$; and

(c) for any k such that $k_l + 1 < k \le k_{l+1}$ for some $l \in \{0, 1, \ldots\}$, we have

$$\Psi(w^k) \le \Psi(w^{k_l+1}). \tag{23}$$

It is easy to see that (b) follows directly from (a) since $\Psi(w^{k_l+1}) \le \gamma\sqrt{\|\Phi(w^{k_l})\|}$. (b) and (c) together yield

$$\lim_{\substack{k\in(k_l+1,k_{l+1}]\\ k\to\infty}} \Psi(w^k) = 0.$$

We observe that every iterate w^k must belong to $k \in K$, or $(k-1) \in K$, or $k_l + 1 < k \le k_{l+1}$ for some $l \in \{0, 1, 2, \ldots\}$. Hence we must have $\lim_{k\to\infty} \Psi(w^k) = 0$ if (a), (b), and (c) are true. Consequently, every limit of $\{w^k\}$ is a solution of (1) and therefore a stationary point of (3).

Now we prove (a) and (c) together. First, for any k between 0 and k_0, i.e., $0 \le k < k_0$, we have $\text{ind}_k = 0$. This means Algorithm 5.1 uses (S.5)(i) to find the next iterate. If (20) is successful at k, then we must have $\gamma_{k+1} < \gamma$ (otherwise k would belong to K, resulting in $k_0 \le k$, a contradiction). Hence it follows from the definition of γ_{k+1} that

$$\Psi(w^{k+1}) = \gamma_{k+1}\Psi(w^k) < \gamma\Psi(w^k) < \Psi(w^k). \tag{24}$$

If (20) is not successful at k, then w^{k+1} is obtained by rule (19). In this case, it is obvious that

$$\Psi(w^{k+1}) \le \Psi(w^k). \tag{25}$$

By the induction argument on k between 0 and k_0, relations (24) and (25) give us that

$$\Psi(w^{k_0}) \le \Psi(w^{k_0-1}) \le \cdots \le \Psi(w^1) \le \Psi(w^0). \tag{26}$$

Now we take a look at how Algorithm 5.1 runs at iterations between k_l and k_{l+1}. Our first observation is that

$$\beta_{k_{l+1}} = \beta_{k_{l+1}-1} = \cdots = \beta_{k_l+1} = \Psi(w^{k_l+1}). \tag{27}$$

For any $k_l \in K$, according to (S.5)(i)

$$\Psi(w^{k_l+1}) = \bar{\gamma}\Psi(w^{k_l}), \quad \text{ind}_{k_l+1} = 1 \text{ (since } \gamma_{k_l+1} \ge \gamma). \tag{28}$$

Then Algorithm 5.1 uses (S.5)(ii) (since $\text{ind}_{k_l+1} = 1$) to generate the next iterate w^{k_l+2}. The algorithm will repeat (S.5)(ii) until (21) is successful at some iterate, say

$w^{\bar{k}}$, putting $\text{ind}_{\bar{k}+1}$ back to *zero* so that the algorithm uses (S.5)(i) to find the next iterate until k reaches k_{l+1}. Hence, there is exactly one such $\bar{k}$ satisfying $k_l + 1 \leq \bar{k} < k_{l+1}$. At iteration k, where $k_l + 1 \leq k < \bar{k}$, the algorithm always uses rule (19) to generate the next iterate, which means that

$$\Psi(w^{k_l+1}) \geq \Psi(w^{k_l+2}) \geq \cdots \geq \Psi(w^{\bar{k}}). \tag{29}$$

At this stage, $\bar{\gamma}$ remains unchanged; i.e., $\bar{\gamma} = \Psi(w^{k_l+1})/\Psi(w^{k_l})$ for all $k \in [k_l + 1, \bar{k}]$. Since (21) is successful at $\bar{k}$, we have

$$\begin{aligned}
\Psi(w^{\bar{k}+1}) &\leq \frac{\gamma}{\bar{\gamma}}\beta_{\bar{k}} = \frac{\gamma}{\bar{\gamma}}\Psi(w^{k_l+1}) \quad \text{(using (27))} \\
&\leq \gamma \frac{\Psi(w^{k_l})}{\Psi(w^{k_l+1})}\Psi(w^{k_l+1}) \quad (\text{using } \bar{\gamma} = \Psi(w^{k_l+1})/\Psi(w^{k_l})) \\
&= \gamma\Psi(w^{k_l}). \qquad (30)
\end{aligned}$$

On the other hand, at iteration k, where $\bar{k} + 1 \leq k < k_{l+1}$, the algorithm uses either rule (19) or (20) to generate the next iterate. If the algorithm uses (19), then it is obvious that $\Psi(w^{k+1}) \leq \Psi(w^k)$. If the algorithm uses (20), then we must have $\gamma_{k+1} < \gamma$, which also yields $\Psi(w^{k+1}) < \Psi(w^k)$. Hence, we have

$$\Psi(w^{\bar{k}+1}) \geq \Psi(w^{\bar{k}+2}) \geq \cdots \geq \Psi(w^{k_{l+1}}). \tag{31}$$

Putting (30) and (31) together, we obtain by an induction argument

$$\Psi(w^{k_{l+1}}) \leq \gamma\Psi(w^{k_l}) \leq \gamma^2\Psi(w^{k_{l-1}}) \leq \cdots \leq \gamma^{l+1}\Psi(w^0).$$

The last inequality uses (26). Taking the limit in the above inequalities gives (a). Finally, it follows from (30) that

$$\Psi(w^{\bar{k}+1}) \leq \gamma\Psi(w^{k_l}) \leq \gamma_{k_l+1}\Psi(w^{k_l}) = \Psi(w^{k_l+1}).$$

This together with (29) and (31) implies (23). ∎

We now consider the case that K contains only finitely many elements, say

$$K = \{k_0, k_1, \ldots, k_l\}.$$

LEMMA 6.2. *Suppose that K contains finitely many elements. Then the following hold:*

(i) *The sequence $\{\Psi(w^k)\}_{k \geq k_l+1}$ is monotonically decreasing.*

(ii) *If test* (20) *holds infinitely many times, then*

$$\lim_{k\to\infty} \Psi(w^k) = 0.$$

In this case, every limit of $\{w^k\}$ is a solution of (1).

Proof. Since k_l is the last element in K, we have by the definition of K that $\text{ind}_{k_l+1} = 1$, $\bar{\gamma} = \Psi(w^{k_l+1})/\Psi(w^{k_l})$, and $\bar{\gamma}$ remains unchanged from k_l+1 and onward.

(i) Since $\text{ind}_{k_l+1} = 1$, the algorithm uses (S.5)(ii) to generate the iterate w^{k_l+2}. We note that test (21) could possibly hold only once after the iteration $k_l + 1$ since once (21) holds the algorithm puts ind_k back to *zero* and will never use (S.5)(ii) thereafter. Suppose that (21) holds at iteration $\bar{k}$ ($\bar{k} \geq k_l + 1$). For iterations k

satisfying $k_l + 1 \le k < \bar{k}$, $\mathrm{ind}_k = 1$ and hence the algorithm uses rule (19) to update w^k. Therefore, we have

$$\Psi(w^{k+1}) \le \Psi(w^k) \qquad \forall\, k \in [k_l + 1, \bar{k}).$$

At iteration $\bar{k}$, we have, from $\bar{\gamma} \ge \gamma$,

$$\Psi(w^{\bar{k}+1}) \le \frac{\gamma}{\bar{\gamma}} \Psi(w^{\bar{k}}) \le \Psi(w^{\bar{k}})$$

and $\mathrm{ind}_{\bar{k}+1} = 0$. So from iteration $\bar{k}+1$ onward, the algorithm uses (S.5)(i) to update w^k. Let $k \ge \bar{k} + 1$ be given. If (20) is successful at k, then we must have $\gamma_{k+1} < \gamma$ (otherwise $k \in K$, resulting in $k \le k_l$, a contradiction of $k \ge \bar{k} + 1 \ge k_l + 2$). Then

$$\Psi(w^{k+1}) = \gamma_{k+1} \Psi(w^k) < \gamma \Psi(w^k). \tag{32}$$

If (20) is not successful at k, the algorithm uses rule (19) to update w^k, giving $\Psi(w^{k+1}) \le \Psi(w^k)$. If (21) never holds from $k_l + 1$ onward, then the algorithm uses rule (19) to generate w^{k+1} for all $k \ge k_l + 1$. We then have $\Psi(w^{k+1}) \le \Psi(w^k)$ for all $k \ge k_l + 1$. All in all, we have proved the statement in (i).

(ii) Suppose that test (20) holds infinitely many times, which means that there exists $\bar{k} \ge k_l + 1$ such that (21) holds at $\bar{k}$. The algorithm uses (S.1)(i) to update w^k from $\bar{k} + 1$ onward, and (20) holds infinitely many times after $\bar{k} + 1$. Hence the relation (32) holds infinitely many times after $\bar{k} + 1$. Noting that $\{\Psi(w^k)\}_{k \ge k_l + 1}$ is monotonically decreasing, we certainly have (ii) from (32). □

The goal we want to achieve in this section is that any limit of the sequence $\{w^k\}$ is a stationary point of (3), which under reasonable conditions [11, Thm. 3.1] is already a solution of (1). Because of Lemmas 6.1 and 6.2, we need only consider the case that K contains finitely many elements and test (20) holds only finitely many times. In other words, we need only consider the case that Algorithm 5.1, after finitely many iterations, uses only rule (19) to update w^k. Without loss of generality we assume from now on that the whole sequence $\{w^k\}$ is generated according to the trust region rule (19). The convergence analysis for this part is quite standard from the trust region point of view.

LEMMA 6.3. *Suppose that the whole sequence $\{w^k\}$ was generated according to rule* (19)*, and that w^* is the limit of a subsequence $\{w^k\}_{\tilde{K}}$. If w^* is not a stationary point of* (3)*, then*

$$\lim_{k \to \infty} \inf_{k \in \bar{K}} \Delta_k > 0.$$

Proof. It is obvious that the function value sequence $\{\Psi(w^k)\}_{k \ge 1}$ is monotonically decreasing, and so is the sequence $\{\delta_k\}$. Moreover,

$$\lim_{k \to \infty} \delta_k = \delta_* := \min\{\delta, c\sqrt{\|\Phi(w^*)\|}\} > 0;$$

the last inequality uses the fact $\Psi(w^*) > 0$ as w^* is not a stationary point of (3). Now define the index set

$$\bar{K} := \{k - 1 |\ k \in \tilde{K}\}.$$

Then the subsequence $\{w^{k+1}\}_{k \in \bar{K}}$ converges to w^*. Suppose that the result of this lemma does not hold. Subsequencing if necessary we can assume that

$$\lim_{k \to \infty, k \in \bar{K}} \Delta_{k+1} = 0. \tag{33}$$

In view of the updating rule for the trust region radius (note that the lower bound

$\Delta_{\min} > 0$ plays an important role here), (33) implies that for all iterations $k \in \bar{K}$ sufficiently large, we have

$$r_k < \rho_1, \qquad w^k = w^{k+1}, \qquad \Delta_{k+1} = \sigma_1 \Delta_k. \tag{34}$$

Hence

$$\{w^k\}_{\bar{K}} \to w^* \ \text{ and } \lim_{k\to\infty, k\in\bar{K}} \Delta_k = 0. \tag{35}$$

Because of the continuity of $\nabla\Psi(\cdot)$, the first convergence in (35) implies the boundedness of $\{\|g^k_{J_k}\|\}_{\bar{K}}$. Taking into account the boundedness of $\{w^k_{J_k}\}_1^\infty$ (since $0 \le w^k_j \le \delta_k$ for all $j \in J_k$ and all k) we obtain the boundedness of $\{\|v^k_{J_k}\|\}_{\bar{K}}$.

Due to the upper semicontinuity of the generalized Jacobian, the sequence $\{\|H_k^T H_k\|\}$ is bounded for all $k \in \bar{K}$; hence the norm of its submatrix $\{\|(H^k_{\cdot\bar{J}_k})^T H^k_{\cdot\bar{J}_k}\|\}$ is also bounded for $k \in \bar{K}$. The fact

$$\lim_{k\to\infty} \rho(\Psi(w^k)) = \rho(\Psi(w^*))$$

implies that there is a constant $\kappa_1 > 0$ such that

$$\|\mathcal{B}_k\| \le \kappa_1 \tag{36}$$

for all $k \in \bar{K}$. We recall from the proof of Lemma 5.3 that $(g^k)^T v^k = (g^k_{J_k})^T v^k_{J_k} + \|g^k_{\bar{J}_k}\|^2$ and $(g^k_{J_k})^T v^k_{J_k} \ge 0$ for all k. Since w^* is not a stationary point of (3), in view of Lemma 5.2 there exists a constant $\kappa_2 > 0$ such that

$$\max\{(g^k_{J_k})^T v^k_{J_k}, \|g^k_{\bar{J}_k}\|\} \ge \kappa_2 \tag{37}$$

for all $k \in \bar{K}$. By (35) and $\delta_* > 0$, we have $\delta_k \ge \Delta_k$ for all $k \in \bar{K}$ sufficiently large. This implies that the estimate in Proposition 4.3(i) holds for all sufficiently large $k \in \bar{K}$. Hence we have, for all $k \in \bar{K}$ sufficiently large,

$$\begin{aligned} &\mathrm{Pred}_k \ge \Delta_k (g^k_{J_k})^T v^k_{J_k} + \frac{1}{4}\|g^k_{\bar{J}_k}\| \min\left\{\Delta_k, \frac{\|g^k_{\bar{J}_k}\|}{\|\mathcal{B}_k\|}\right\} \ge \frac{1}{4}\gamma_2 \Delta_k, \\ &\|\tilde{d}^k\| \le \min\{1, \Delta_k\}\|v^k_{J_k}\| + \|d^k_{\bar{J}_k}\| \le \left(1 + \|v^k_{J_k}\|\right)\Delta_k, \end{aligned} \tag{38}$$

where the last inequality in (38) uses the bounds (36)–(37) and the limit (35). Then $\{\tilde{d}^k\}_{k\in\bar{K}} \to 0$ because of the boundedness of $\{\|v^k_{J_k}\|\}_{k\in\bar{K}}$ and (35). By the mean value theorem, we have

$$\Psi(w^k + \tilde{d}^k) = \Psi(w^k) + \nabla\Psi(\xi^k)^T \tilde{d}^k \quad \text{ for some } \xi^k = w^k + \theta_k \tilde{d}^k,\ \theta_k \in (0,1).$$

Obviously, we have $\{\xi^k\}_{k\in\bar{K}} \to w^*$ as $\{\tilde{d}^k\}_{k\in\bar{K}} \to 0$. Then we obtain for $k \in \bar{K}$ sufficiently large

$$\begin{aligned} |r_k - 1| &= \left|\frac{\Psi(w^k) - \Psi(w^k + \tilde{d}^k)}{\mathrm{Pred}_k} - 1\right| \\ &= \frac{1}{\mathrm{Pred}_k}\left|\Delta_k \left((\nabla\Psi(\xi^k) - \nabla\Psi(w^k))_{J_k}\right)^T v^k_{J_k} + \left((\nabla\Psi(w^k) - \nabla\Psi(\xi^k))_{\bar{J}_k}\right)^T d^k_{\bar{J}_k}\right. \\ &\qquad \left. + \frac{1}{2}(d^k_{\bar{J}_k})^T \mathcal{B}_k d^k_{\bar{J}_k}\right| \qquad \text{(by (35) and the definition of } \tilde{d}^k) \end{aligned}$$

$$\le \frac{4}{\gamma_2\Delta_k}\left(\Delta_k\left\|(\nabla\Psi(\xi^k)-\nabla\Psi(w^k))_{J_k}\right\|\|v^k_{J_k}\|+\left\|(\nabla\Psi(w^k)-\nabla\Psi(\xi^k))_{\bar J_k}\right\|\|d^k_{\bar J_k}\|\right.$$
$$\left.+\frac{1}{2}\|\mathcal{B}_k\|\|d^k_{\bar J_k}\|^2\right) \qquad \text{(by the Cauchy–Schwarz inequality and (38))}$$
$$\le \frac{4}{\gamma_2}\left((1+\|v^k_{J_k}\|)\|\nabla\Psi(w^k)-\nabla\Psi(\xi^k)\|+\frac{1}{2}\|\mathcal{B}_k\|\Delta_k\|\right) \qquad (\text{by } \Delta_k\ge\|d^k_{\bar J_k}\|)$$
$$\to 0 \qquad \text{(by the boundedness of } \{\|v^k_{J_k}\|\}_{\bar K} \text{ and (35)).}$$

Hence the subsequence $\{r_k\}_{k\in\bar K}$ converges to 1, which is a contradiction to $r_k\le\rho_1$ in (34). □

With the help of Lemma 6.3, we are able to prove the following global convergence result. Its proof is quite standard and is omitted here. One can mimic the proof of [20, Thm. 3.1] to prepare one.

LEMMA 6.4. *Suppose that the whole sequence $\{w^k\}$ was generated according to rule* (19)*. Then any accumulation point of $\{w^k\}$ is a stationary point of* (3).

Combining the results of Lemmas 6.1, 6.2, and 6.4, we have our main result in this section.

THEOREM 6.5. *Let $\{w^k\}$ be generated by Algorithm* 5.1*, with the subproblem* (9) *being solved by the truncated CG Algorithm* 4.1*. Then any accumulation point of $\{w^k\}$ is a stationary point of* (3).

7. Local convergence. Let $\{w^k\}$ be a sequence generated by Algorithm 5.1, and let w^* be a strongly regular solution of (1). Our main result in this section is that if w^* is an accumulation point of $\{w^k\}$, then the whole sequence converges to w^* superlinearly/quadratically. The proof is based on a number of lemmas. The proof techniques of some of those lemmas are borrowed from [19]. Therefore, we will omit most of proofs in this section, but we would like to indicate their connections to [19] and refer to [28] for fully worked out proofs.

The two results of the following lemma are simple consequences of the strong regularity. The first one is about the active set J_* at w^* defined by

$$J_* := \{i\in\mathcal{J}\mid z^*_j=0\}.$$

Since our algorithm makes use of an active-set strategy, we hope that the set J_k is capable of identifying J_* correctly whenever w^k is close to w^*. This can be shown by using a recently proposed identification technique by Facchinei, Fischer, and Kanzow [10]. The second is the uniform nonsingularity of a matrix sequence [19, Lem. 5].

LEMMA 7.1. *Suppose that $\{w^k\}$ is a sequence generated by Algorithm* 5.1 *and w^* is a strongly regular solution of* (1)*. If w^* is an accumulation point of $\{w^k\}$, then the following hold:*

(i) *$J_k=J_*$ for all w^k in a sufficiently small ball around w^*.*

(ii) *There is a constant $c_2>0$ such that the matrices $(H^k_{\cdot\bar J_k})^TH^k_{\cdot\bar J_k}$ are nonsingular and*

$$\left\|\left((H^k_{\cdot\bar J_k})^TH^k_{\cdot\bar J_k}\right)^{-1}\right\|\le c_2$$

for all w^k in a sufficiently small ball around w^*.

Suppose w^* is a solution of $\Phi(w)=0$. Then $\|\Phi(w)\|=o(\sqrt{\|\Phi(w)\|})$ whenever w is close enough to w^*. Using this fact, Proposition 4.3, (6), and Lemma 7.1(ii), we

can obtain the following bound on d^k by using a proof technique similar to that of [19, Lem. 6].

LEMMA 7.2. *Suppose that $\{w^k\}$ is a sequence generated by Algorithm* 5.1, *and w^* is a strongly regular solution of* (1). *If w^* is an accumulation point of $\{w^k\}$, then there exists a constant $c_3 > 0$ such that*

$$\|d^k\| \le c_3\sqrt{\|\Phi(w^k)\|}$$

for all w^k sufficiently close to w^*, where d^k denotes the vector computed in step (S.4) *of Algorithm* 5.1.

Using Lemma 7.2, we are now able to show the convergence of the whole sequence $\{w^k\}$ (see [19, Lem. 8] for a proof.)

LEMMA 7.3. *Let $\{w^k\}$ be generated by Algorithm* 5.1, *and let w^* be a strongly regular solution of* (1). *If w^* is an accumulation point of $\{w^k\}$, then the whole sequence $\{w^k\}$ converges to w^*.*

The results developed so far allow us to establish one more technical result, which in turn implies that the iterates are eventually generated by $w^k + d^k$. The third result of the next lemma can be proved similarly to the proof of [19, Lem. 11].

LEMMA 7.4. *Let $\{w^k\}$ be generated by Algorithm* 5.1, *and let w^* be a strongly regular solution of* (1) *and an accumulation point of $\{w^k\}$. Let $\{d^k\}$ denote the directions computed in step* (S.4) *of Algorithm* 5.1. *Then the following hold:*

(i) *For all k sufficiently large, it holds that*

$$\Psi(w^k + d^k) \le \gamma\sqrt{\|\Phi(w^k)\|}.$$

(ii) *There are infinitely many iterates w^k at which* $\mathrm{ind}_k = 0$.

(iii) *It holds that* $\mathrm{ind}_k = 0$ *for all k sufficiently large.*

Proof. Under the assumed conditions, it is proved in Lemma 7.3 that the whole sequence $\{w^k\}$ converges to w^* with $\Phi(w^*) = 0$. Then Lemma 7.2 implies that there exists $c_3 > 0$ such that

$$\|d^k\| \le c_3\sqrt{\|\Phi(w^k)\|}$$

for all k sufficiently large. So the sequence $\{w^k + d^k\}$ also converges to w^*. Then for all k sufficiently large we have

$$\begin{cases} \|\Phi(w^k + d^k)\| \le L\|w^k + d^k - w^*\|, \\ \|\Phi(w^k)\| \ge c_1\|w^k - w^*\|,\ \|H_k\| \le \kappa_3, \\ \sqrt{\|\Phi(w^k)\|} \le \min\{\gamma/(2L^2c_3^2),\ c_1c_3\}, \end{cases} \tag{39}$$

where L is the Lipschitz constant of $\Phi(\cdot)$ in a small ball around w^*, c_1 is the constant used in Proposition 2.3, and κ_3 is the constant used in the proof of Lemma 7.2. It also follows from Proposition 2.2 and Lemma 7.1 that for all k sufficiently large

$$\|\Phi(w^k) - \Phi(w^*) - H_k(w^k - w^*)\| = o(\|w^k - w^*\|) \tag{40}$$

and

$$J_k = J_*. \tag{41}$$

(i) The inequalities in (39) yield (i) as follows for all k sufficiently large:

$$\begin{aligned}
&\Psi(w^k+d^k)=\frac{1}{2}\|\Phi(w^k+d^k)\|^2\\
&\le \frac{L^2}{2}\|w^k+d^k-w^*\|^2 \le \frac{L^2}{2}\left(\|w^k-w^*\|+\|d^k\|\right)^2\\
&\le \frac{L^2}{2}\left(\|\Phi(w^k)\|/c_1+c_3\sqrt{\|\Phi(w^k)\|}\right)^2 \le \frac{L^2}{2c_1^2}\|\Phi(w^k)\|\left(c_1c_3+\sqrt{\|\Phi(w^k)\|}\right)^2\\
&\le 2(Lc_3)^2\|\Phi(w^k)\| \le \gamma\sqrt{\|\Phi(w^k)\|}.
\end{aligned}$$

(ii) Suppose to the contrary that there is a $\bar{k}$ such that $\mathrm{ind}_k = 1$ for all $k \ge \bar{k}$. We again let K be defined as (22). Then K contains finitely many indices, which we denote by

$$K := \{k_0, k_1, \ldots, k_l\}$$

for some integer l. By the update rule for β_k, we have

$$\beta_k = \beta_{k_l+1} = \Psi(w^{k_l+1}) \quad \forall\, k > k_l + 1,$$

and there is no new update for $\bar{\gamma}$ after $k_l + 1$, i.e.,

$$\bar{\gamma} = \gamma_{k_l+1} = \frac{\Psi(w^{k_l+1})}{\Psi(w^{k_l})} \quad \forall\, k > k_l + 1.$$

Since $\{w^k + d^k\}$ converges to w^*, $\Psi(w^k + d^k)$ converges to $\Psi(w^*) = 0$. Hence for all k sufficiently large

$$\Psi(w^k+d^k) \le \gamma\Psi(w^{k_l}) \le \frac{\gamma}{\bar{\gamma}}\Psi(w^{k_l+1}) = \frac{\gamma}{\bar{\gamma}}\beta_k;$$

that is, test (21) is successful for all k sufficiently large. Since $\mathrm{ind}_k = 1$ for k large enough, Algorithm 5.1 (S.5)(ii) assigns $\mathrm{ind}_{k+1} = 0$, which contradicts our assumption. This establishes (ii). (iii) can be proved similarly to the proof of [19, Lem. 11] by noticing Proposition 4.2. □

We are now at the position to state the main local convergence result.

THEOREM 7.5. *Let $\{w^k\}$ be a sequence generated by Algorithm* 5.1, *and let w^* be a strongly regular solution of* (1) *and an accumulation point of $\{w^k\}$. Then the following statements hold:*

(a) *The whole sequence $\{w^k\}$ converges to w^*.*
(b) *The rate of convergence is Q-superlinear.*
(c) *The rate of convergence is Q-quadratic if, in addition, F is an LC^1 function and h, g are LC^2 functions and $\rho(\Psi(w^k)) = O(\sqrt{\Psi(w^k)})$.*

Proof. Statement (a) follows immediately from Lemma 7.3. We have proved in Lemma 7.4 that $\mathrm{ind}_k = 0$ and test (20) holds for all k sufficiently large. Hence, there exists $\bar{k} > 0$ such that for all $k \ge \bar{k}$

$$w^{k+1} := w^k + d^k.$$

Then by an argument similar to the proof of [19, Lem. 11], we can prove

$$\|w^{k+1} - w^*\| = o(\|w^k - w^*\|),$$

and under the condition in (c) and with Proposition 2.2, we can prove

$$\|w^{k+1} - w^*\| = O(\|w^k - w^*\|^2).$$

This proves (b) and (c). □

8. Numerical results. In this section, we present some numerical experiments on a subset of problems from the MCPLIB collection [7]. The details about the implementation are described as follows.

(a) *The penalized Fischer–Burmeister function.* Instead of using the Fischer–Burmeister function φ, we use its penalized version $\varphi_\alpha : \mathbb{R}^2 \to \mathbb{R}$ defined by

$$\varphi_\alpha(a,b) := \alpha\varphi(a,b) + (1-\alpha)a_+b_+, \qquad \alpha \in (0,1],$$

where $a_+ = \max\{0,a\}$ for any $a \in \mathbb{R}$. Numerical tests indicate that the penalized Fischer–Burmeister function usually leads to better numerical performance than the Fischer–Burmeister function [3, 34]. In our implementation, $\alpha = 0.7$, as recommended by Ulbrich [37].

(b) *SSOR preconditioner.* As we pointed out in the introduction, the key issue of efficient implementation of CG-type methods is the *preconditioning*. Steihaug's CG method with preconditioner C can be found in [33]. Although there is no single preconditioning that is "best" for all conceivable types of matrices, we choose C to be the SSOR preconditioner of the following type of linear equations:

$$(A^TA + \rho I)x = b, \tag{42}$$

where A, b have compatible dimension and ρ is a small positive number. This type of equation is exactly what we try to solve at each iteration. Then the SSOR preconditioner corresponds to taking

$$C = P^TP \text{ and } P = D_A^{-1/2}(D_A + \omega L_A^T), \quad 0 \le \omega < 2, \tag{43}$$

with the standard splitting $A^TA + \rho I = L_A + D_A + L_A^T$, where L_A is strictly lower triangular. The cost of matrix-vector product for each of the SSOR PCG iterations including the matrix-vector product with A^TA is in fact only $4\text{nnz}(A)$, and A^TA is not formed explicitly, where nnz is (Matlab) notation of the number of nonzero elements. See [15, Table 1], [1, p. 284], and [2] for more information about the counting. According to [15], SSOR-CG and TMRES (transformed minimal residual algorithm) are the two most efficient methods for solving linear equations of type (42) compared with several other (iterative) methods. Theory and numerical experiments indicate that $\omega = 1$ is often close to the optimum choice of ω [1]. In our implementation, $\omega = 1$.

(c) *Nonmonotone calculation of the reduction-ratio r_k.* In calculating r_k in (18), we used its nonmonotone version,

$$r_k = \left(\mathcal{W}_k - \Psi(w^k + \tilde{d}^k)\right)/\text{Pred}_\text{k},$$

where $\mathcal{W}_k := \max\{\Psi(w^j) |\ j = k+1-\ell, \ldots, k\}$ denotes the maximal function value of Ψ over the last ℓ iterations. The nonmonotone version often gives an overall better performance than its monotone version. For more discussion, see [37]. In our implementation, $\ell = 4$.

(d) *Test problems.* The test problems we used are selected from the MCPLIB collection [7] and have at most one bound per variable, i.e., $u_i - l_i = +\infty$ for all i. The collection itself is updated from time to time. As of the initial point of those problems, we follow a suggestion of Ulbrich [37] that interior starting points enable constrained algorithms to identify the correct active constraints more efficiently than starting points close to the boundary. Let $\hat{x}^0$ be the initial point returned by the initialization routine `mcpinit`. Then the initial point chosen is given by $x^0 = \max\{l + 0.1, \min\{u - 0.1, \hat{x}^0\}\}$.

The algorithm was implemented in Matlab and run on a SUN Solaris (CDE Version 1.2) workstation. The parameters used are $\Delta_0 = \min\{0.1\|g^0\|, 30\sqrt{10n}\}$, $\Delta_{\min} = 1$, $\rho_1 = 10^{-4}$, $\rho_2 = 0.75$, $\sigma_1 = 0.1$, $\sigma_2 = 10$, $c = 1$, $\gamma = 0.9$, $\delta = 10^{-4}$, and tol $= 10^{-10}$. The algorithm was terminated if one of the following conditions was met:

$$\max\{\Psi(w^k), \|\nabla\Phi(w^k)\|, \|v_k\|\} \le \text{tol} \quad \text{or} \quad \texttt{it\_outer} \ge 100,$$

where `it_outer` denotes the (outer) iteration number. The forcing function used in our implementation is $\rho(\Psi(w)) = \min\{10^{-6}, \sqrt{\Psi(w)}\}$. The stop rule (S.1) used in Algorithm 4.1 is replaced by

$$\|\mathcal{B}_k s^i - g^k_{\bar{J}_k}\|_{C_k} / \|g^k_{\bar{J}_k}\|_{C_k} \le \text{tol}$$

if $g^k_{\bar{J}_k} \neq 0$, and $s = 0$ would otherwise be the (unique) solution of the subproblem.

There are two iterative procedures in the implementation: One is the iterative procedure in Algorithm 5.1, which we call the outer iterative procedure. For each outer iteration, there is the truncated PCG iterative procedure described in Algorithm 4.1 for solving the trust region subproblem, which we call the inner iterative procedure. The average number of inner iterations per outer iteration is essential to the efficiency of our approach. The following data are reported in our numerical results: `n`, the problem size; `it_outer`, number of outer iterations when Algorithm 5.1 was terminated; `it_inner`, average number of PCG iterations per out iteration, i.e., `it_inner` = [Total numbers of PCG iterations/`it_outer`], where $[z]$ denotes the nearest integer to z; `nf`, number of evaluations of the function F; $\Psi(w^f)$, the value of $\Psi(\cdot)$ at the final iterate; $\|\nabla\Psi(x^f)\|$, the value of $\|\nabla\Psi(\cdot)\|$ at the final iterate; $\|v^f\|$, the value of $\|v^k\|$ at the final iteration. `it_outer` is also equal to the number of evaluations of the Jacobian $F'(x)$.

We tested Algorithm 5.1 with two purposes: to demonstrate the importance of preconditioning and to compare our numerical results with existing ones.

(e) *Importance of preconditioning.* For this purpose, we tested three versions of Algorithm 5.1. `tcg`: Algorithm 5.1 without preconditioning ($C = I$); `tcg_ssor`: Algorithm 5.1 with SSOR-preconditioner ($C = P^T P$ with P given by (43)); and `tcg_chol`: Algorithm 5.1 with Cholesky direct factorization ($C = R^T R$ with R being the Cholesky factor of $\mathcal{B}_k$). We expect that `tcg_ssor` is much more efficient than `tcg` and is less efficient than `tcg_chol` (we use `tcg_ssor` as benchmark). The numerical results confirm this expectation. Table 1 contains results from `tcg` and Table 2 contains results from `tcg_ssor`. On the one hand, `tcg` failed to solve four more problems (i.e., `bertsekas`, `freebert`, `games`, and `methan08`) than `tcg_ssor`. But for the remaining solved problems, they behaved quite similarly except for `colvdual`. To solve this problem, `tcg` took many more functional evaluations and outer iterations than `tcg_ssor` did. The observation clearly shows that the preconditioned version is much more efficient than the unpreconditioned version. On the other hand, `tcg_chol` is able to solve one more problem (`ne-hard`) than `tcg_ssor`, and they behaved very similarly for the rest of the problems in terms of number of functional evaluations and outer iterations. For this reason and to save space as well, we did not include the complete results for `tcg_chol`, but the final information for `ne-hard` is included in Table 2. We note that when the Cholesky preconditioner is applied in Algorithm 4.1, in theory it takes only one inner iteration to solve the trust region subproblem, and the resulting direction is of the Gauss–Newton type. Its steplength is controlled

TABLE 1
Numerical results with `tcg`.

Problem	n	it_inner	it_outer	nf	$\Psi(x^f)$	$\|\nabla\Psi(x^f)\|$	$\|v^f\|$
badfree	5	2	4	5	2.68e-13	8.87e-07	8.87e-07
bertsekas	15	–	–	–	–	–	–
billups	1	1	1	2	9.80e-05	2.94e-02	0.00e+00
bishopl	1645	–	–	–	–	–	–
colvdual	20	19	48	200	7.11e-13	6.56e-06	6.56e-06
colvnlp	15	18	6	8	3.34e-13	2.91e-05	2.91e-05
cycle	1	1	5	7	2.18e-15	2.44e-07	2.44e-07
degen	2	1	5	7	2.88e-11	7.51e-06	7.51e-06
duopoly	63	–	–	–	–	–	–
ehl_k40	41	–	–	–	–	–	–
ehl_k60	61	–	–	–	–	–	–
ehl_k80	81	–	–	–	–	–	–
ehl_kost	101	–	–	–	–	–	–
explcp	16	8	11	13	1.97e-13	4.40e-07	4.40e-07
forcebsm	184	–	–	–	–	–	–
forcedsa	186	–	–	–	–	–	–
freebert	15	–	–	–	–	–	–
games	16	–	–	–	–	–	–
hanskoop	14	10	22	32	1.65e-12	1.51e-05	1.51e-05
hydroc06	29	94	64	170	8.24e-11	2.59e-04	2.59e-04
hydroc20	99	–	–	–	–	–	–
jel	6	8	6	9	5.84e-14	6.09e-06	6.09e-06
josephy	4	3	4	5	3.38e-20	1.60e-09	1.60e-09
kojshin	4	4	3	4	1.17e-11	2.70e-05	2.70e-05
lincon	419	–	–	–	–	–	–
mathinum	3	3	5	6	8.82e-15	2.63e-07	2.63e-07
mathisum	4	3	8	10	5.83e-11	5.62e-05	5.62e-05
methan08	31	–	–	–	–	–	–
nash	10	10	5	6	1.47e-17	2.87e-07	2.87e-07
ne-hard	3	–	–	–	–	–	–
pgvon106	106	–	–	–	–	–	–
powell	16	12	5	6	6.93e-14	6.14e-06	5.46e-06
powell_mcp	8	7	2	3	3.33e-13	7.37e-06	7.37e-06
qp	4	2	9	45	1.26e-14	8.55e-07	3.19e-07
scarfanum	13	10	13	16	2.83e-16	8.57e-07	8.57e-07
scarfasum	14	12	11	45	7.31e-19	4.28e-08	4.28e-08
scarfbsum	40	–	–	–	–	–	–
shubik	45	–	–	–	–	–	–
simple-ex	17	–	–	–	–	–	–
simple-red	13	15	10	15	2.12e-11	3.10e-06	3.10e-06
sppe	27	36	4	5	1.56e-18	3.58e-09	3.57e-09
tinloi	146	5	6	63	1.30e-11	1.71e-02	1.59e-02
tobin	42	36	8	43	2.02e-22	2.08e-10	1.85e-10
trafelas	2904	–	–	–	–	–	–

by the trust region radius. In practice, it may take more than one inner iteration to produce a direction due to the accumulated roundoff. Moreover, when the linear equations of the type (42) is near singular, the Cholesky factor may not exist, leading to the failure of `tcg_chol`. In our experiments, `tcg_chol` took only one inner iteration per outer iteration. So it is not appropriate to compare `tcg_ssor` with `tcg_chol` in terms of inner iterations taken per outer iteration. However, it is safe to say that the efficiency of the preconditioned Algorithm 5.1 varies with the preconditioners used.

(f) *Comparison.* The results in Table 2 are comparable to those results obtained with existing methods [37, 35]. Moreover, for most of the tested problems the av-

TABLE 2
Numerical results with `tcg_ssor`.

Problem	n	it_inner	it_outer	nf	$\Psi(x^f)$	$\|\nabla\Psi(x^f)\|$	$\|v^f\|$
badfree	5	3	4	5	1.15e-12	1.84e-06	1.84e-06
bertsekas	15	7	18	58	8.84e-18	1.43e-07	1.43e-07
billups	1	1	1	2	9.80e-05	2.94e-02	0.00e+00
bishop	1645	–	–	–	–	–	–
colvdual	20	14	22	68	8.05e-13	9.11e-05	9.11e-05
colvnlp	15	11	6	8	3.35e-13	2.92e-05	2.92e-05
cycle	1	1	5	7	2.18e-15	2.44e-07	2.44e-07
degen	2	1	5	7	2.88e-11	7.51e-06	7.51e-06
duopoly	63	–	–	–	–	–	–
ehl_k40	41	–	–	–	–	–	–
ehl_k60	61	–	–	–	–	–	–
ehl_k80	81	–	–	–	–	–	–
ehl_kost	101	–	–	–	–	–	–
explcp	16	4	11	13	1.96e-13	4.39e-07	4.39e-07
forcebsm	184	–	–	–	–	–	–
forcedsa	186	–	–	–	–	–	–
freebert	15	7	24	117	1.18e-12	7.12e-05	7.12e-05
games	16	12	14	36	1.74e-14	5.16e-06	4.58e-06
hanskoop	14	9	24	37	1.09e-13	3.75e-06	3.72e-06
hydroc06	29	41	36	42	6.09e-11	1.86e-07	1.86e-07
hydroc20	99	–	–	–	–	–	–
jel	6	5	6	9	5.84e-14	6.09e-06	6.09e-06
josephy	4	3	4	5	3.38e-20	1.60e-09	1.60e-09
kojshin	4	4	3	4	1.17e-11	2.70e-05	2.70e-05
lincon	419	–	–	–	–	–	–
mathinum	3	3	5	6	8.82e-15	2.63e-07	2.63e-07
mathisum	4	3	8	10	5.83e-11	5.62e-05	5.62e-05
methan08	31	46	51	59	9.08e-11	9.95e-09	9.95e-09
nash	10	4	5	6	1.49e-17	2.86e-07	2.86e-07
ne-hard	3	1	25	51	5.09e-11	5.60e-02	6.60e-02
pgvon106	106	–	–	–	–	–	–
powell	16	10	14	61	5.69e-12	3.80e-05	3.31e-05
powell_mcp	8	7	2	3	3.33e-13	7.36e-06	7.36e-06
qp	4	2	9	45	1.26e-14	8.55e-07	3.19e-07
scarfanum	13	8	13	16	2.83e-16	8.57e-07	8.57e-07
scarfasum	14	9	11	45	7.31e-19	4.28e-08	4.28e-08
scarfbsum	40	–	–	–	–	–	–
shubik	45	–	–	–	–	–	–
simple-ex	17	–	–	–	–	–	–
simple-red	13	10	10	15	1.04e-11	2.18e-06	2.18e-06
sppe	27	15	4	5	1.50e-18	1.96e-09	1.96e-09
tinloi	146	5	6	63	1.30e-11	1.71e-02	1.59e-02
tobin	42	19	8	43	2.78e-23	1.93e-11	1.89e-11
trafelas	2904	–	–	–	–	–	–

erage number of inner iterations per outer iteration (column 3) in Table 2 is small compared with the problem size. Although we have a few more failed problems, we would like to point out that for some of those failed problems, say `ehl_60`, `ehl_80`, and `ehl_kost`, all of which can be solved in [37, 35], we were able to arrive at points with functional values at the order of 10^{-5} and 10^{-8} for `hydroc20` within 10 iterations, but our algorithm hardly achieved any significant improvement thereafter. The difficulties may come from two resources: On the one hand, there are many ways, all mathematically equivalent, in which to implement the CG method for (42). In *exact* arithmetic they will all generate the same sequence of approximations and all have

finite termination property, but in *finite* precision the achieved accuracy may differ substantially. On the other hand, it is hard to select the "best" *preconditioner* for all linear systems of type (42) arising from our implementation. Hence for those failed problems, an appropriate option is an efficient implementation of the CG method with a more suitable *preconditioner* (other than SSOR). Fortunately, many other *preconditioners* with corresponding efficient implementation of the CG method are available; see [1, pp. 293–311]. We also learned that the number of inner iterations may grow significantly for (very) ill-conditioned linear systems of type (42) in order to meet the required accuracy. We refer readers who are interested in the reasons to a recent paper [15] by Hager. We also observed that the sooner direction d^k is taken, the sooner the active set is identified. In fact, for most cases, the active set is identified at least two or three iterations before termination.

Based on the numerical results reported and observation in (e) and (f), we feel that the trust region PCG method proposed in this paper provides an alternative to the existing methods for solving KKT systems. More numerical experiments need to be done to evaluate the proposed approach, especially on large-scale problems with a Jacobian appearing in a certain pattern of sparsity.

9. Conclusions. In this paper, we proposed a trust region algorithm for solving KKT systems arising from VIPs. Built around those components of the current iterate, which are far from the boundary of the constrained region, the trust region subproblem is solved by the truncated PCG method. Global and local convergence analysis are provided for this method. Numerical experiments show that the proposed method is promising, mainly due to its computational inexpensiveness in that the trust region subproblem is solved by the truncated CG method. We also briefly discussed ways for improving practical efficiency of the proposed method.

Acknowledgments. The authors would like to thank the associate editor and two anonymous referees for their detailed comments, which considerably improved the presentation of the paper. In particular, one referee's expert comments on PCG methods led us to the current version of Algorithm 4.1 as well as condition (13).

REFERENCES

[1] Å. Björck, *Numerical Methods for Least Squares Problems*, SIAM, Philadelphia, PA, 1996.
[2] Å. Björck and T. Elfving, *Accelerated projection methods for computing pseudoinverse solutions of systems of linear equations*, BIT, 19 (1979), pp. 145–163.
[3] B. Chen, X. Chen, and C. Kanzow, *A penalized Fischer-Burmeister NCP-function*, Math. Program., 88 (2000), pp. 211–216.
[4] F.H. Clarke, *Optimization and Nonsmooth Analysis*, Wiley, New York, 1983.
[5] A.R. Conn, N.I.M. Gould, and P.L. Toint, *Global convergence of a class of trust region algorithms for optimization with simple bounds*, SIAM J. Numer. Anal., 25 (1988), pp. 433–460.
[6] A.R. Conn, N.I.M. Gould, and P.L. Toint, *Trust Region Methods*, MPS/SIAM Ser. on Optim. 1, SIAM, Philadelphia, 2000.
[7] S.P. Dirkse and M.C. Ferris, *MCPLIB: A collection of nonlinear mixed complementarity problems*, Optim. Methods Softw., 5 (1995), pp. 319–345.
[8] F. Facchinei, A. Fischer, and C. Kanzow, *A semismooth Newton method for variational inequalities: The case of box constraints*, in Complementarity and Variational Problems: State of the Art, Proc. Appl. Math. 92, M. C. Ferris and J.-S. Pang, eds., SIAM, Philadelphia, 1997, pp. 76–90.
[9] F. Facchinei, A. Fischer, and C. Kanzow, *Regularity properties of a semismooth reformulation of variational inequalities*, SIAM J. Optim., 8 (1998), pp. 850–869.
[10] F. Facchinei, A. Fischer, and C. Kanzow, *On the accurate identification of active constraints*, SIAM J. Optim., 9 (1998), pp. 14–32.

[11] F. Facchinei, A. Fischer, C. Kanzow, and J.-M. Peng, *A simply constrained optimization reformulation of KKT systems arising from variational inequalities*, Appl. Math. Optim., 40 (1999), pp. 19–37
[12] A. Fischer, *A special Newton-type optimization method*, Optimization, 24 (1992), pp. 269–284.
[13] A. Fischer, *Solution of monotone complementarity problems with locally Lipschitzian functions*, Math. Programming, 76 (1997), pp. 513–532.
[14] S.A. Gabriel and J.-S. Pang, *An inexact NE/SQP method for solving the nonlinear complementarity problem*, Comput. Optim. Appl., 1 (1992), pp. 67–91.
[15] W.W. Hager, *Iterative methods for nearly singular linear systems*, SIAM J. Sci. Comput., 22 (2000), pp. 747–766.
[16] H. Jiang, M. Fukushima, L. Qi, and D. Sun, *A trust region method for solving generalized complementarity problems*, SIAM J. Optim. 8 (1998), pp. 140–157.
[17] C. Kanzow, *An inexact QP-based method for nonlinear complementarity problems*, Numer. Math., 80 (1998), pp. 557–577.
[18] C. Kanzow, *Strictly feasible equation-based methods for mixed complementarity problems*, Numer. Math., 89 (2001), pp. 135–160.
[19] C. Kanzow and H.-D. Qi, *A QP-free constrained Newton-type method for variational inequality problems*, Math. Program., 85 (1999), pp. 81–106.
[20] C. Kanzow and M. Zupke, *Inexact trust-region methods for nonlinear complementarity problems*, in Reformulation: Nonsmooth, Piecewise Smooth, Semismooth and Smoothing Methods, M. Fukushima and L. Qi, eds., Kluwer Academic, Dordrecht, The Netherlands, 1998, pp. 211–233.
[21] C.-J. Lin and J.J. Moré, *Newton's method for large bound-constrained optimization problems*, SIAM J. Optim., 9 (1999), pp. 1100–1127.
[22] J. Liu, *Strong stability in variational inequalities*, SIAM J. Control Optim., 33 (1995), pp. 725–749.
[23] J.J. Moré and D.C. Sorensen, *Computing a trust region step*, SIAM J. Sci. Statist. Comput., 4 (1983), pp. 553–572.
[24] J.-S. Pang and S.A. Gabriel, *NE/SQP: A robust algorithm for the nonlinear complementarity problem*, Math. Programming, 60 (1993), pp. 295–337.
[25] J.-S. Pang and L. Qi, *Nonsmooth equations: Motivation and algorithms*, SIAM J. Optim., 3 (1993), pp. 443–465.
[26] J.-M. Peng, *Global method for monotone variational inequality problems with inequality constraints*, J. Optim. Theory Appl., 95 (1997), pp. 419–430
[27] M.J.D. Powell, *A new algorithm for unconstrained optimization*, in Nonlinear Programming 2, O. L. Mangasarian, R. R. Meyer, and S. M. Robinson, eds., Academic Press, New York, 1975, pp. 1–27.
[28] H.-D. Qi, L. Qi, and D. Sun, *Solving KKT Systems via the Trust Region and the Conjugate Gradient Methods*, AMR99/19, School of Mathematics, University of New South Wales, Sydney, Australia, 1999.
[29] L. Qi, *Convergence analysis of some algorithms for solving nonsmooth equations*, Math. Oper. Res., 18 (1993), pp. 227–244.
[30] L. Qi and H. Jiang, *Semismooth Karush-Kuhn-Tucker equations and convergence analysis of Newton and quasi-Newton methods for solving these equations*, Math. Oper. Res., 22 (1997), pp. 301–325.
[31] L. Qi and J. Sun, *A nonsmooth version of Newton's method*, Math. Programming, 58 (1993), pp. 353–368.
[32] S.M. Robinson, *Strongly regular generalized equations*, Math. Oper. Res., 5 (1980), pp. 43–62.
[33] T. Steihaug, *The conjugate gradient method and trust regions in large scale optimization*, SIAM J. Numer. Anal., 20 (1983), pp. 626–637.
[34] D. Sun and L. Qi, *On NCP-functions*, Comput. Optim. Appl., 13 (1999), pp. 201–220.
[35] D. Sun, R.S. Womersley, and H.-D. Qi, *A feasible semismooth asymptotically Newton method for mixed complementarity problems*, Math. Program., 94 (2002), pp. 167–187.
[36] P.L. Toint, *Towards an efficient sparsity exploiting Newton method for minimization*, in Sparse Matrices and Their Uses, I. Duff, ed., Academic Press, New York, 1981, pp. 57–88.
[37] M. Ulbrich, *Nonmonotone trust-region methods for bound-constrained semismooth equations with applications to nonlinear mixed complementarity problems*, SIAM J. Optim., 11 (2001), pp. 889–917.
[38] Y. Yuan, *On the truncated conjugate gradient method*, Math. Program., 87 (2000), pp. 561–573.

SIAM J. OPTIM.
Vol. 14, No. 2, pp. 464–478

PROBABILITY DISTRIBUTIONS OF ASSETS INFERRED FROM OPTION PRICES VIA THE PRINCIPLE OF MAXIMUM ENTROPY*

J. BORWEIN†, R. CHOKSI‡, AND P. MARÉCHAL§

Abstract. This article revisits the maximum entropy algorithm in the context of recovering the probability distribution of an asset from the prices of finitely many associated European call options via partially finite convex programming. We are able to provide an effective characterization of the constraint qualification under which the problem reduces to optimizing an explicit function in finitely many variables. We also prove that the value (or objective) function is lower semicontinuous on its domain. Reference is given to a website which exploits these ideas for the efficient computation of the maximum entropy solution (MES).

Key words. European options, maximum entropy, semifinite programming, Lagrangian duality, convex conjugate

AMS subject classifications. 90C25, 49N15, 91B28

DOI. 10.1137/S1052623401400324

1. Introduction.

Entropy optimization, used for recovering a probability distribution from information on a few of its moments, is well established and ubiquitous throughout the sciences [14]. Recently (cf. Buchen and Kelly [9] and Avellaneda et al. [1], [2]), this idea has been explored in the context of financial derivatives. In this *risk-neutral* model, one wishes to infer the probability distribution for the price of an asset at some future date T from the prices of European call options based upon the asset with expiration at T.

A classical approach to the application of entropy optimization has been to use the theory of Lagrange multipliers. While this formal approach does yield correct and useful results, it does not provide for a complete analysis. The purpose of this article is to analyze the option-maximum entropy problem within the framework of partially finite programming and demonstrate the extra insight and power that this approach provides. In doing so, we not only legitimize the formal calculations with Lagrange multipliers but also provide a more detailed analysis of the maximum entropy solution and the notion of admissible data. We also specifically exploit the unique structure of the piecewise linear constraints to reduce the problem to maximization of an *explicit function of finitely many variables*; hence greatly simplifying the computation of the maximum entropy solution.

The option-maximum entropy problem. Let I be an the interval of the form $[0, K)$ with either some fixed $K > 0$ or $K = +\infty$. For $0 = k_1 < k_2 < \cdots < k_m$, and

*Received by the editors December 28, 2001; accepted for publication (in revised form) January 15, 2003; published electronically October 14, 2003.
http://www.siam.org/journals/siopt/14-2/40032.html
†Department of Mathematics and Centre for Experimental and Constructive Mathematics, Simon Fraser University, Burnaby, Canada (jborwein@cecm.sfu.ca).
‡Department of Mathematics, Simon Fraser University, Burnaby, Canada (choksi@math.sfu.ca).
§Département de Sciences Mathématiques, Université Montpellier II, Montpellier, France (marechal@darboux.math.univ-montp2.fr).

$\mathbf{d} \in \mathbb{R}^m$,

$$(\mathcal{P}) \qquad \left| \begin{array}{ll} \text{minimize} & \mathcal{I}_h(p) := \displaystyle\int_I h\big(p(x)\big)\,dx \\ \text{s.t.} & 1 = \displaystyle\int_I p(x)\,dx, \\ & d_j = \displaystyle\int_I c_j(x)p(x)\,dx. \end{array} \right.$$

Here, $p(x)$ denotes the *probability density function* for the *price* x of an *asset* at a set *future time* T, and d_j represents the price of a *European call option* based on the underlying asset with *strike price* k_j and *expiration date* T. The interval I denotes the set of feasible prices for the asset at time T which may or may not be a priori constrained. The function $c_j(x)$ represent the payoffs of the jth option as a function of the asset price x at time T. Thus

$$c_j(x) \,=\, (x \,-\, k_j)^+ \,=\, \max\{0\,,\, x \,-\, k_j\}. \tag{1}$$

Finally the convex function $h : \mathbb{R} \to \mathbb{R}$ represents the entropy functional, the most common of which being the *Boltzmann–Shannon entropy*

$$h(t) := \begin{cases} t\log t - t & \text{if } t > 0, \\ 0 & \text{if } t = 0, \\ +\infty & \text{if } t < 0. \end{cases} \tag{2}$$

Note that traditionally, the entropy is taken to be $-h$, and hence maximum entropy entails solving for the minimum in $(\mathcal{P})$. We refer to the minimizer associated with $(\mathcal{P})$ as the maximum entropy solution, or simply the MES.

The particular choice of the Boltzmann–Shannon entropy yields a simple case of the *minimum cross entropy problem* using the *Kullback–Leibler* entropy. Here the idea is that given additionally a prior guess $q(x)$ for the asset price distribution at T (which one might infer from the market), one seeks to find the least prejudiced posterior density $p(x)$ consistent with the constraints which is *closest* to or least *deviant* from $q(x)$ in the following sense (see Cover and Thomas [10] for details): find a constraint satisfying $p(x)$ which minimizes

$$\int_I p(x)\log\left(\frac{p(x)}{q(x)}\right)\,dx.$$

Our problem $(\mathcal{P})$ is the simple case of the above where no prior is available and hence $q(x)$ is close to a uniform distribution and may be taken to be a constant. Of course to be precise, it will be uniform if $p(x) = 0$ for all x sufficiently large (cf. [9]). For simplicity we first carry out our analysis for the Boltzmann–Shannon entropy (i.e., uniform prior). In section 7, we briefly comment on the necessary modifications and drawbacks in the more realistic situation of including a nontrivial prior.

The constraints in $(\mathcal{P})$ may appear to be missing something. Indeed, they should read

$$d_j = DC(T)\int_I c_j(x)p(x)\,dx,$$

where $DC(T)$ represents the *riskless discount factor* up to time T. For example, one could take

$$DC(T) \,=\, e^{-rT},$$

where r is the *risk-free constant interest rate*. Without loss of generality we set $DC(T) = 1$ throughout this paper. Finally, we emphasize that this model, in which the option prices are simply the expected values of a discounted pay-off function, assumes *risk-neutrality*. See [13] (also [9] and the references therein) for further information on *risk-neutral* pricing and *arbitrage-free* models.

Convex programming approach. In this article, we reexamine problem $(\mathcal{P})$ within the general framework of convex duality and partially finite convex programming. Why this approach? To begin with, it legitimizes the calculations done in [9] and [1] which are based upon Lagrange multipliers. This standard approach is based on relaxing the hard constraints via Lagrange multipliers, reducing the problem to

$$\sup_{\boldsymbol{\lambda}\in\mathbb{R}^{m+1}} \inf_{p} L(\boldsymbol{\lambda}, p), \tag{3}$$

where

$$L(\boldsymbol{\lambda}, p) := \mathcal{I}_h(p) + \lambda_0 \left(1 - \int_0^\infty p\,dx\right) + \sum_{i=1}^{m} \lambda_i \left(d_i - \int_0^\infty c_i\, p\,dx\right). \tag{4}$$

The minimization over p is carried out via the first variation of $L(\boldsymbol{\lambda}, p)$ with respect to p; i.e., one "differentiates" the Lagrangian with respect to density functions p. There is a problem with this type of calculation. Indeed, the Lagrangian has support on the set $\{p \in L^1(I) \,|\, p \geq 0 \text{ a.e.}\}$. The complement of this set is dense in L^1 and, moreover, any reasonable definition of the Boltzmann–Shannon entropy gives a value of $+\infty$ on any function in the complement. Thus, not only is the Lagrangian nowhere differentiable, it is indeed nowhere continuous. The approach via conjugation-duality is in part to circumvent this differentiation. Moreover, with other entropies, there can be additional complications to these formal calculations resulting from a lack of weak compactness. See [8] for a fuller discussion. We emphasize, however, that the benefits of our approach are far from confined to the legitimization of the now fairly ubiquitous if flawed formal analysis with "Lagrange multipliers." Such benefits include the following:

- We transform the maximum entropy problem into a closed-form finite-dimensional maximization problem. That is, under certain explicit conditions on the data, finding the MES is equivalent to maximizing an explicit dual function (cf. (9), (12), and (16)) of finitely many real variables. The simple fact that the dual function can be written explicitly with no integrals is an advantage of using a uniform prior.
- Our approach greatly simplifies the numerical computation[1] of the MES where many of the previous numerical calculations (cf. [9]) involved in computing the optimal $\boldsymbol{\lambda}$ can now be done symbolically.
- We give a detailed analysis of the constraint qualification (CQ) and a full investigation of when the MES exists, and when the maximization with respect to $\boldsymbol{\lambda}$ in the dual (cf. (9), (12)) does indeed yield the solution. These results are pertinent when analyzing the dependence of the MES on the data $\mathbf{d}$.

[1] An interface has been set up at http://www.cecm.sfu.ca/projects/MomEnt+/moment.html which computes the MES for a variety of moment constraints, including the ones discussed in the present paper. One can test our algorithm by first pricing the list of options using, for example, a log-normal distribution, and then comparing the distribution with the computed MES based only on the option prices. In this way, one finds that the accuracy of recovering a known distribution with eight options is quite high even with a uniform prior.

- Our general approach applies to any convex entropy, not just to the standard Boltzmann–Shannon entropy used in [9], [1]. It is also amenable to natural extensions such as relaxations of the constraints, for example, requiring the moments to lie in some small finite interval.
- Partially finite duality and attainment results are usually confined to primal function spaces defined over bounded domains. The problem provides an interesting and simple example whereby a partially finite duality and attainment theorem can be proved in the case where the primal functions are defined over an infinite domain (Theorem 2). We know of no general result which would capture this.

In section 3 we prove two duality results: one for the case of a finite interval I and the other for $I = [0, \infty)$. The first (Corollary 1) is a direct consequence of a well-known duality result (Theorem 1). The latter (Theorem 2) is proved directly by exploiting the monotonicity of the constraints c_i. In either case, the MES exists if d satisfies the CQ. Conversely, for the MES to exist in its exponential form (cf. (10), and (13)), this CQ must hold. The CQ amounts to the data $\mathbf{d}$ lying in the relative interior of the *feasible set*, i.e., the set of vectors $(y_0, y_1, \ldots, y_m) \in \mathbb{R}^{m+1}$ such that

$$y_i = \int_I c_i(x)p(x)\,dx \quad \text{for} \quad i = 0, \ldots, m$$

for some distribution p with finite entropy. In section 4 we show that this condition is equivalent to the data $\mathbf{d}$ lying in some open polyhedral set which we characterize explicitly (cf. Proposition 2). It is important to note that the feasible set is not relatively open, and hence there can exist boundary points which are feasible even though the CQ fails. In such cases, the analysis via the Lagrange multipliers λ_i will fail. Indeed, as the data approaches such a boundary point, some components of the associated $\boldsymbol{\lambda}$ will become infinite.

We provide a simple—though perhaps artificial from a finance point of view—example to illustrate these points. We use only two constraints for simplicity (similar examples exist with many options) and assume the first option has strike price zero. That is, we consider strike prices $k_1 = 0, k_2$ with associated option prices d_1 and d_2 (with $d_2 \geq 1/2$). This data satisfies the CQ if and only if

$$0 \;<\; d_1 - d_2 \;<\; k_2.$$

The boundary point where $d_1 - d_2 = k_2$ is of particular interest. Clearly, this data is feasible; for example, consider

$$p = \chi_{[k_2+d_2-\frac{1}{2},\, k_2+d_2+\frac{1}{2}]}.$$

Moreover one can readily show (see (18)) that any probability distribution satisfying the associated constraints must vanish on the interval $[0, k_2]$. Hence, no MES solution of the exponential form (i.e., (10)) can exist. Indeed, as data satisfying the CQ tends to this boundary point, the associated $\boldsymbol{\lambda}$ must blow up. This simple example illustrates that an infimum associated with problem $(\mathcal{P})$ might still be finite but not attainable. In section 6, we explore this matter further by studying the *value (or objective) function* and whether or not there exists a *duality gap*.

2. Preliminaries. We first reformulate problem $(\mathcal{P})$. Let $I = [0, K)$ with either $K > 0$ fixed or $K = +\infty$. For $m \geq 1$, we assume that $0 = k_1 < \cdots < k_m < K$, and

$\mathbf{d} = (d_0, d_1, \ldots, d_m) \in \mathbb{R}^{m+1}$ with $d_0 = 1$. Consider

$$(\mathcal{P}) \qquad \inf \left\{ \mathcal{I}_h(p) + \delta \left(\mathbb{A}p - \mathbf{d} \;\middle|\; \mathbf{0} \right) \right\}, \quad \text{where} \quad \mathcal{I}_h(p) := \int_I h(p(x))\, dx,$$

δ is the indicator function defined for the set $\{\mathbf{0}\}$, i.e., for $\mathbf{y} \in \mathbb{R}^{m+1}$,

$$\delta \left(\mathbf{y} \;\middle|\; \mathbf{0} \right) := \begin{cases} 0 & \text{if } \mathbf{y} = \mathbf{0}, \\ \infty & \text{otherwise,} \end{cases}$$

and $\mathbb{A}$ is the linear operator defined by

$$\mathbb{A}p := \int_I \mathbf{c}(x) p(x)\, dx \;\in\; \mathbb{R}^{m+1},$$

with $\mathbf{c}(x) = (c_0(x), c_1(x), \ldots, c_m(x))$, $c_0(x) \equiv 1$, $c_j(x) = (x - k_j)^+ = \max\{0, x - k_j\}$. Finally, h always denotes the Boltzmann–Shannon entropy defined by (2). The space we will work in for admissible p is $L^1(I)$. We will separate the cases of bounded I and $I = [0, \infty)$. For the latter case, $\mathbb{A}$ may be infinite on some $p \in L^1([0,\infty))$, and hence for the problem at hand $\mathbb{A}$ is not a well-defined linear operator on $L^1([0,\infty))$ as it would be on, say, $L^1([0, M])$ for some fixed $M > 0$. One notes that even though the operator $\mathbb{A}$ is densely defined on $L^1([0,\infty))$, it is not closed. Hence this case requires a different approach. For the case of bounded I, we will directly apply partially finite convex programming (Theorem 1) to establish the duality relation under a CQ. A similar duality relation (Theorem 2), under the same CQ, holds true for the infinite domain $I = [0,\infty)$ and will be proved directly, bypassing the Fenchel duality of Theorem 1.

For omitted definitions and elementary facts from convex analysis in $\mathbb{R}^n$ we refer the reader to [17]. Let V and $V^\star$ be vector spaces equipped with $\langle \cdot, \cdot \rangle$, a bilinear product on $V \times V^\star$. The *convex (Fenchel) conjugate* of a convex function f on V with respect to $\langle \cdot, \cdot \rangle$ is the function $f^\star$ defined on V^* by

$$f^\star(\xi) := \sup \left\{ \langle x, \xi \rangle - f(x) \;\middle|\; x \in V \right\}.$$

We consider the functional on $L^1(I)$ (for I bounded or unbounded) defined by

$$u \longmapsto \mathcal{I}_h(u) := \int_I h(u(x))\, dx, \tag{5}$$

where the integral is interpreted in the sense of Rockafellar (cf. [18, p. 7]). Thus $\mathcal{I}_h$ is a well-defined operator from $L^1(I)$ to $[-\infty, \infty]$ and, since the entropy h is convex, also convex on $L^1(I)$.

For the conjugate of this integral functional, we take I to be a bounded interval and let $L := L^1(I)$ and $L^* := L^\infty(I)$. One can define a bilinear product on $L \times L^\star$ by

$$(u, u^\star) \longmapsto \left\langle u, u^\star \right\rangle := \int_I u(x) u^\star(x)\, dx. \tag{6}$$

To compute the convex conjugate of $\mathcal{I}_h$ with respect to (6), we may conjugate the integrand, as in the following proposition.

PROPOSITION 1. *Let I be* bounded *and consider the pair $\langle L, L^\star \rangle$ of subspaces of $L^1(I)$ as defined above with bilinear product* (6)*. Then for any $q \in L^\star$, we have*

$$\mathcal{I}_h^\star(q) = \int_I h^\star\big(q(t)\big)\, dt.$$

The proof of Proposition 1 can be found in either [19] or [16]. Finally, we recall a Fenchel duality theorem in its *partially finite* version. The proof of Theorem 1 as stated can be inferred from Theorem 4.2 in [4], with the attainment of the infimum proved via Theorems 3.7 and 3.8 of [7]. In what follows, "ri" denotes the relative interior of a subset of $\mathbb{R}^n$ and "dom" denotes the effective domain of a convex function (i.e., the set of points at which the function is finite).

THEOREM 1. *Let V and $V^\star$ be vector spaces, and let $\langle\cdot,\cdot\rangle$ be a bilinear product on $V \times V^\star$. Let $G\colon V \to \mathbb{R}^n$ be a linear map with adjoint G^T, let $F\colon V \to \bar{\mathbb{R}}$ be a proper convex function, and let $g\colon \mathbb{R}^n \to \bar{\mathbb{R}}$ be a proper concave function. Then, under the CQ*

$$\text{(CQ)} \qquad \operatorname{ri}(G \operatorname{dom} F) \cap \operatorname{ri}(\operatorname{dom} g) \neq \emptyset,$$

we have

$$\inf\left\{ F(u) - g(Gu) \;\middle|\; u \in V \right\} \;=\; \sup\left\{ g^\star(\boldsymbol{\lambda}) - F^\star(G^T\boldsymbol{\lambda}) \;\middle|\; \boldsymbol{\lambda} \in \mathbb{R}^n \right\}, \tag{7}$$

with the supremum on the right being attained when finite. Moreover for the case where V is a normed vector space with dual V^*, if F is strongly rotund (i.e., satisfies the three conclusions of Lemma 2), the infimum on the left is attained at a unique u.

3. Duality results. Theorem 1 directly applies to $(\mathcal{P})$ with I bounded. That is, $n = m+1$, $V = L^1(I)$, $V^\star = L^\infty(I)$, $\langle\cdot,\cdot\rangle$ is given by (6), $G := \mathbb{A}$, and the function g is defined by

$$g(\mathbb{A}p) \;=\; -\,\delta\left(\mathbb{A}p \,-\, \mathbf{d} \;\middle|\; \mathbf{0}\right).$$

Lastly, $F = \mathcal{I}_h$, where h is the Boltzmann–Shannon entropy functional defined by (2), which by Lemma 2 is strongly rotund. The CQ amounts to

$$\text{(CQ)} \quad \mathbf{d} \in \operatorname{ri}(\mathbb{A} \operatorname{dom} \mathcal{I}_h). \tag{8}$$

Precisely, we have the following.

COROLLARY 1. *Let I be bounded and assume (8) holds. Then $(\mathcal{P})$ has a unique solution and*

$$\begin{aligned} &\inf\left\{ \mathcal{I}_h(p) \;\middle|\; p \in L^1(I),\, \mathbb{A}p = \mathbf{d} \right\} \\ &\qquad = \sup\left\{ \sum_{i=0}^{m} \lambda_i d_i \,-\, \mathcal{I}_h^\star(\mathbb{A}^T(\boldsymbol{\lambda})) \;\middle|\; \boldsymbol{\lambda} \in \mathbb{R}^{m+1} \right\}. \end{aligned} \tag{9}$$

Moreover the solution of the primal problem (left-hand side of (9)) is

$$e^{\sum_{i=0}^{m} \bar{\lambda}_i c_i(x)}, \tag{10}$$

with $\bar{\lambda}_i$ being the unique solution to the dual problem (right-hand side of (9)).

It is straightforward to check (see section 5) that

$$\mathcal{I}_h^\star(\mathbb{A}^T(\boldsymbol{\lambda})) \;=\; \int_I e^{\mu(x)}\,dx, \qquad \mu(x) := \sum_{i=0}^{m} \lambda_i c_i(x),$$

where one can explicitly carry out the integration (cf. (16)). We also note in section 5 that the distribution given by (10) is indeed a probability distribution.

As previously mentioned, the case where $I = [0, \infty)$ is best treated differently. The duality result is identical; however, to prove it we shall bypass the direct application of Theorem 1 and exploit properties of the value function. We rewrite the CQ as

$$\text{(CQ)} \quad \mathbf{d} \in \operatorname{ri} \mathcal{A}, \tag{11}$$

where

$$\mathcal{A} := \left\{ \mathbf{x} \in \mathbb{R}^{m+1} \;\middle|\; \exists\, p \in L^1[0,\infty) \text{ with } \mathcal{I}_h(p) \text{ finite and } \mathbb{A}p = \mathbf{x} \right\}.$$

We have the following theorem.

THEOREM 2. *Let $I = [0, \infty)$ and assume* (11) *holds. Then* $(\mathcal{P})$ *has a unique solution and*

$$\inf \left\{ \mathcal{I}_h(p) \;\middle|\; p \in L^1([0,\infty)),\, \mathbb{A}p = \mathbf{d} \right\}$$

$$= \sup \left\{ \sum_{i=0}^{m} \lambda_i d_i \,-\, \int_0^\infty e^{\mu(x)}\, dx \;\middle|\; \boldsymbol{\lambda} \in \mathbb{R}^{m+1} \right\}. \tag{12}$$

Moreover, the solution of the primal problem is

$$e^{\sum_{i=0}^{m} \bar{\lambda}_i c_i(x)}, \tag{13}$$

with $\bar{\lambda}_i$ being the unique value of the right-hand side of (12).

Proof of Theorem 2. Consider the *value function*

$$\mathcal{V}(\mathbf{d}) := \inf \left\{ \mathcal{I}_h(p) \;\middle|\; \mathbb{A}p = \mathbf{d} \right\} \,=\, \inf \left\{ \mathcal{I}_h(p) + \delta \left(\mathbb{A}p \,-\, \mathbf{d} \;\middle|\; \mathbf{0} \right) \;\middle|\; p \in L^1([0,\infty)) \right\}.$$

We prove that under the (CQ) of (11),

$$\mathcal{V}(\mathbf{d}) \,=\, \sup \left\{ \sum_{i=0}^{m} \lambda_i d_i \,-\, \int_0^\infty e^{\mu(x)}\, dx \;\middle|\; \boldsymbol{\lambda} \in \mathbb{R}^{m+1} \right\}. \tag{14}$$

First note that (14) easily holds with $=$ replaced with $\geq$. To see this, note that by the definition of $h^\star$, for every $p \in \operatorname{dom} \mathcal{I}_h$ with $\int c_i(x) p(x) dx = d_i$, we have

$$\int h^\star(\mu) dx \,+\, \int h(p) dx \,\geq\, \int \sum_{i=0}^{m} \lambda_i\, c_i(x)\, p(x)\, dx \,=\, \sum_{i=0}^{m} \lambda_i d_i$$

holding for any $\boldsymbol{\lambda} \in \mathbb{R}^{m+1}$. The inequality follows by first taking the infimum over all such p, and then the supremum over $\boldsymbol{\lambda} \in \mathbb{R}^{m+1}$.

We now prove the reverse inequality. The (CQ) implies that $\mathbf{d} \in \operatorname{ri}(\operatorname{dom} \mathcal{V})$. Moreover, it is easily verified that $\mathcal{V}$ is convex on its domain. Hence (see, for example, [3]), there exists a $\bar{\boldsymbol{\lambda}} \in \mathbb{R}^{m+1}$ such that $\bar{\boldsymbol{\lambda}} \in \partial \mathcal{V}(\mathbf{d})$, the subgradient of $\mathcal{V}$ at $\mathbf{d}$. Thus for all $\mathbf{z} \in \mathbb{R}^{m+1}$, $\mathcal{V}(\mathbf{z}) \geq \mathcal{V}(\mathbf{d}) + \langle \bar{\boldsymbol{\lambda}}, \mathbf{z} - \mathbf{d} \rangle$. Fix $M > 0$. Restricting our attention to p with support in $[0, M]$, we have (by definition of $\mathcal{V}(\mathbf{z})$) for all $p \in L^1([0, M))$

$$\mathcal{V}(\mathbf{d}) \,-\, \langle \bar{\boldsymbol{\lambda}}, \mathbf{d} \rangle \,\leq\, \mathcal{I}_h(p) \,-\, \langle \bar{\boldsymbol{\lambda}}, \mathbb{A}p \rangle.$$

Setting $\bar{\mu}(x) = \sum_{i=0}^{m} \bar{\lambda}_i c_i(x)$, we have

$$\mathcal{V}(\mathbf{d}) \,-\, \sum_{i=0}^{m} \bar{\lambda}_i d_i \,\leq\, \int_0^M \left(h(p(x)) \,-\, p(x)\, \bar{\mu}(x) \right) dx,$$

and hence

$$\sup_{p\in L^1[0,M]}\left\{\int_0^M p(x)\bar{\mu}(x)\,-\,h(p(x))\,dx\right\}\,\leq\,\sum_{i=0}^m\bar{\lambda}_i d_i\,-\,\mathcal{V}(\mathbf{d}).$$

The left-hand side of the above is by definition $\mathcal{I}_h^\star(\bar{\mu})$. Hence applying Proposition 1 to $\mathcal{I}_h(p)$ on $[0,M]$, we have

$$\mathcal{I}_h^\star(\bar{\mu})\,=\,\int_0^M h^\star(\bar{\mu})dx\,=\,\int_0^M e^{\bar{\mu}(x)}dx\,\leq\,\sum_{i=0}^m\bar{\lambda}_i d_i\,-\,\mathcal{V}(\mathbf{d}),$$

or

$$\sum_{i=0}^m\bar{\lambda}_i\,d_i\,-\,\int_0^M e^{\bar{\mu}(x)}dx\,\geq\,\mathcal{V}(\mathbf{d}).$$

Since the above holds for each $M>0$, the monotone convergence theorem implies

$$\sum_{i=0}^m\bar{\lambda}_i d_i\,-\,\int_0^\infty e^{\bar{\mu}(x)}dx\,\geq\,\mathcal{V}(\mathbf{d}). \tag{15}$$

Lastly, we prove primal attainment. The (CQ) holds, and hence the supremum on the right of (12) is finite, and moreover the previous analysis shows that there exists $\bar{\boldsymbol{\lambda}}$ which attains this supremum. It remains to show that the dual function

$$D(\boldsymbol{\lambda}):=\sum_{i=0}^m\lambda_i d_i\,-\,\int_0^\infty e^{\mu(x)}\,dx$$

is differentiable at $\boldsymbol{\lambda}=\bar{\boldsymbol{\lambda}}$. To this end, we note that by (15),

$$\int_{k_m}^\infty e^{\bar{\mu}(x)}\,dx\,<\,\infty.$$

Since for $x>k_m$, $\bar{\mu}(x)\,=\,\bar{\lambda}_0\,+\,x\,\sum_{i=1}^m\bar{\lambda}_i\,-\,\sum_{i=1}^m k_i\bar{\lambda}_i$, we must have $\sum_{i=1}^m\bar{\lambda}_i<0$, and hence $D(\boldsymbol{\lambda})$ is differentiable at $\boldsymbol{\lambda}=\bar{\boldsymbol{\lambda}}$. Thus

$$d_k=\int_0^\infty c_k(x)\,e^{\bar{\mu}(x)},$$

$\bar{p}(x):=e^{\bar{\mu}(x)}$ is feasible for the primal problem, and

$$\mathcal{I}_h(e^{\bar{\mu}(x)})=\sum_{k=0}^m\bar{\lambda}_k\,d_k\,-\,\int_0^\infty e^{\bar{\mu}(x)}dx.$$

Since equality holds in (12), $e^{\bar{\mu}(x)}$ must indeed be the MES. The uniqueness follows from the strict convexity of the entropy (see, for example, [3]). □

In the following sections we complement Corollary 1 and Theorem 2 by giving an explicit characterization of the (CQ) for our problem ($\mathcal{P}$), and by computing the *dual function* D explicitly in a form with no integrals.

4. The CQ. In Proposition 2 below, we give an explicit form of the CQ for problem $(\mathcal{P})$, first for $I = [0, \infty)$ and then for $I = [0, K]$. We shall need the following simply lemma, whose proof is left as an exercise.

LEMMA 1. *Let $I = [0, \infty)$ and $\varphi(x) := [\mathbb{A}^T \boldsymbol{\lambda}](x) = \lambda_0 + \lambda_1 c_1(x) + \cdots + \lambda_m c_m(x)$. The following conditions are equivalent:*

(a) *for all p s.t. $\mathcal{I}_h(p)$ is finite and $\mathbb{A}p \in \mathbb{R}^{m+1}$, we have $\langle \boldsymbol{\lambda}, \mathbb{A}p \rangle \geq 0$;*

(b) *$\varphi(x) \geq 0$ for all $x \in \mathbb{R}_+$;*

(c) *$M\boldsymbol{\lambda} \geq \mathbf{0}$ (componentwise), where*

$$M := \begin{pmatrix} 1 & & & & \\ 1 & k_2 - k_1 & & & \\ \vdots & \vdots & \ddots & & \\ 1 & k_m - k_1 & \cdots & k_m - k_{m-1} & \\ 0 & 1 & \cdots & 1 & 1 \end{pmatrix}.$$

If A is an $(m \times n)$-matrix and K_A is the convex cone defined by $K_A := \{\mathbf{x} \in \mathbb{R}^n \mid A\mathbf{x} \geq \mathbf{0}\}$, one may easily verify that for the dual cone K_A^+, we have

$$K_A^+ := \left\{ \mathbf{y} \in \mathbb{R}^n \;\middle|\; \langle \mathbf{y}, \mathbf{x} \rangle \geq 0 \; \forall \, \mathbf{x} \in K_A \right\} = A^T \mathbb{R}_+^m,$$

where A^T denotes the adjoint of A (for example, see [3]).

PROPOSITION 2. *Let $I = [0, \infty)$ and $m > 2$. Then $(1, d_1, \ldots, d_m)$ satisfies the CQ (11) for $(\mathcal{P})$ if and only if $(d_1, \ldots, d_m)^T$ satisfies*

$$d_m > 0, \quad N^{-1} B (d_1, \ldots, d_m)^T > \mathbf{0}, \quad \text{and} \quad \langle N^{-1} B (d_1, \ldots, d_m)^T, \mathbf{u} \rangle < 1,$$

in which $\mathbf{u}$ is the vector of appropriate dimension whose components are all equal to 1, and N and B are, respectively, the $(m-1) \times (m-1)$- and $(m-1) \times m$-matrices given by

$$N := \begin{pmatrix} k_2 - k_1 & \cdots & k_m - k_1 \\ & \ddots & \vdots \\ & & k_m - k_{m-1} \end{pmatrix}, \qquad B := \begin{pmatrix} 1 & & & -1 \\ & \ddots & & \vdots \\ & & 1 & -1 \end{pmatrix}.$$

Proof. We denote by cl the closure of a subset of $\mathbb{R}^n$. A classical separation argument shows that the vector $\mathbf{d}' \in \mathbb{R}^{1+m}$ does not belong to the closed convex set $\operatorname{cl} \mathcal{A}$ if and only if there exists $\boldsymbol{\lambda} \in \mathbb{R}^{1+m}$ such that

(α) $\langle \boldsymbol{\lambda}, \mathbf{d}' \rangle < 0$, and

(β) $\langle \boldsymbol{\lambda}, \boldsymbol{\xi} \rangle \geq 0$ for all $\boldsymbol{\xi} \in \operatorname{cl} \mathcal{A}$.

Clearly, $\operatorname{cl} \mathcal{A}$ can be replaced by $\mathcal{A}$ in condition (β), which can thus be rewritten as

(β') $\langle \mathbb{A}^T \boldsymbol{\lambda}, p \rangle \geq 0$ for all p s.t. $\mathcal{I}_h(p)$ is finite and $\mathbb{A}p \in \mathbb{R}^{m+1}$.

But from Lemma 1, the latter condition is equivalent to $M\boldsymbol{\lambda} \geq 0$. In other words, we have shown that $\mathbf{d}' \in \operatorname{cl} \mathcal{A}$ if and only if for all $\boldsymbol{\lambda} \in \mathbb{R}^{1+m}$, either $\langle \boldsymbol{\lambda}, \mathbf{d}' \rangle \geq 0$ or $M\boldsymbol{\lambda} \ngeq 0$.

Let us define $C_M = \{\boldsymbol{\lambda} \in \mathbb{R}^{1+m} | M\boldsymbol{\lambda} \geq 0\}$. We have

$$\begin{aligned} \operatorname{cl} \mathcal{A} &= \left\{ \mathbf{d}' \;\middle|\; \forall \boldsymbol{\lambda}, \; M\boldsymbol{\lambda} \ngeq 0 \text{ or } \left\langle \boldsymbol{\lambda}, \mathbf{d}' \right\rangle \geq 0 \right\} \\ &= \left\{ \mathbf{d}' \;\middle|\; \forall \boldsymbol{\lambda}, \; \boldsymbol{\lambda} \notin C_M \text{ or } \left\langle \boldsymbol{\lambda}, \mathbf{d}' \right\rangle \geq 0 \right\} \\ &= \left\{ \mathbf{d}' \;\middle|\; \forall \boldsymbol{\lambda} \in C_M, \; \left\langle \boldsymbol{\lambda}, \mathbf{d}' \right\rangle \geq 0 \right\} \\ &= C_M^+. \end{aligned}$$

By the previously mentioned characterization of C_M^+ as well as by standard properties of the relative interior of convex sets (see [17], section 6), we obtain

$$\operatorname{ri}\mathcal{A} = \operatorname{ri}\operatorname{cl}\mathcal{A} = \operatorname{ri} M^T\mathbb{R}_+^{1+m} = M^T \operatorname{ri}\mathbb{R}_+^{1+m} = M^T(0,\infty)^{1+m}.$$

Consequently, $(1, d_1, \ldots, d_m)$ belongs to $\operatorname{ri}\mathcal{A}$ if and only if

$$\begin{cases} 1 &= \xi_0 + \xi_1 + \cdots + \xi_{m-1} \\ d_1 &= (k_2-k_1)\xi_1 + \cdots + (k_m-k_1)\xi_{m-1} + \xi_m \\ &\vdots \\ d_{m-1} &= (k_m-k_{m-1})\xi_{m-1} + \xi_m \\ d_m &= \xi_m \end{cases}$$

for some $\boldsymbol{\xi} > \mathbf{0}$. By subtracting the last line from lines $2, \ldots, m-1$ in the above system, we see that $(1, d_1, \ldots, d_m) \in \operatorname{ri}\mathbb{A}\operatorname{dom}\mathcal{I}_h$ if and only if

$$d_m > 0, \quad N^{-1}B(d_1, \ldots, d_m)^T > \mathbf{0}, \quad \text{and} \quad \langle N^{-1}B(d_1, \ldots, d_m)^T, \mathbf{u}\rangle < 1.$$

Notice that N is invertible since $k_m > \cdots > k_1$ by assumption. □

For the case of bounded $I = [0, K]$, one can show Proposition 2 holds with the one modification of replacing B by

$$B_K := \begin{pmatrix} 1 & & -\dfrac{K-k_1}{K-k_m} \\ & \ddots & \vdots \\ & 1 & -\dfrac{K-k_{m-1}}{K-k_m} \end{pmatrix}.$$

The proof of this is similar to that of Proposition 2.

5. Maximizing the dual function. Recall from Corollary 1 that under the CQ (8), the optimal value of $(\mathcal{P})$ is equal to the optimal value of the dual problem

$$(\mathcal{D}) \qquad \max\left\{ D(\lambda_0, \boldsymbol{\lambda}) := \lambda_0 + \sum_{i=1}^m \lambda_i d_i - \mathcal{I}_h^\star\big(\mathbb{A}^T(\lambda_0, \boldsymbol{\lambda})\big) \;\middle|\; (\lambda_0, \boldsymbol{\lambda}) \in \mathbb{R}^{1+m} \right\}.$$

The formal adjoint $\mathbb{A}^T$ of $\mathbb{A}$ is readily computed as

$$\mathbb{A}^T((\lambda_0, \boldsymbol{\lambda})) = \big\langle (\lambda_0, \boldsymbol{\lambda}), (1, \mathbf{c}(\cdot)) \big\rangle.$$

By Proposition 1, we have

$$\begin{aligned} \mathcal{I}_h^\star\big(\mathbb{A}^T(\lambda_0, \boldsymbol{\lambda})\big) &= \int_I h^\star\big(\lambda_0 + \langle \boldsymbol{\lambda}, \mathbf{c}(x)\rangle\big)\, dx \\ &= \exp\lambda_0 \times \int_0^K \exp\left[\sum_{i=1}^m \lambda_i (x-k_i)^+\right] dx \\ &= \exp\lambda_0 \times \sum_{j=1}^m \int_{k_j}^{k_{j+1}} \exp\left[\left(\sum_{i=1}^j \lambda_i\right) x - \sum_{i=1}^j \lambda_i k_i\right] dt \\ &= \exp\lambda_0 \times \sum_{j=1}^m \left(\exp(-\nu_j)\frac{\exp\mu_j k_{j+1} - \exp\mu_j k_j}{\mu_j}\right), \end{aligned} \tag{16}$$

in which $k_{m+1} := K$, $\nu_j := \sum_{i=1}^{j} \lambda_i k_i$, and $\mu_j := \sum_{i=1}^{j} \lambda_i$. The expression $\mu_j^{-1}(\exp \mu_j k_{j+1} - \exp \mu_j k_j)$ is understood to be $k_{j+1} - k_j$ when $\mu_j = 0$.

For the case $I = [0, \infty)$, Theorem 2 directly gave rise to the same dual function (with the integration carried out over the entire half line). In this case we have $\int_0^\infty e^{\mu(x)}$ equal to (16) with $k_{m+1} := +\infty$ and the understanding that $\exp(-\infty)$ is equal to zero.

We remark that $e^{-\lambda_0}$ can be taken to be

$$Z(\boldsymbol{\lambda}) := \int_0^\infty \exp\left[\sum_{i=1}^{m} \lambda_i (x - k_i)^+\right] dx,$$

and hence the dual function to be maximized can be written in terms of $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_m)$ as

$$\log Z(\boldsymbol{\lambda}) - \sum_{i=1}^{m} \lambda_i d_i. \tag{17}$$

In particular, the MES is indeed a probability distribution and has the form

$$\frac{1}{Z(\boldsymbol{\lambda})}\, e^{\sum_{i=1}^{m} \lambda_i c_i(x)},$$

with λ_i maximizing (17).

6. The value function. The *value* (or *objective*) function associated with problem $(\mathcal{P})$ is defined for $\mathbf{d} = (1, d_1, \ldots, d_m)$ by

$$\mathcal{V}(\mathbf{d}) := \inf \left\{ \mathcal{I}_h(p) \;\middle|\; \mathbf{d} = \mathbb{A}p \right\}.$$

While it is known that the value function is continuous on the interior of the CQ set, it is not in general on its closure. It turns out that if $\mathcal{V}$ is lower semicontinuous on its domain (the set of all feasible data), then there is no duality gap, i.e., (9) and (12) hold whenever the left-hand side is finite.

We will prove $\mathcal{V}$ is lower semicontinuous on its domain for the case $I = [0, \infty)$. The proof for bounded I follows verbatim from the first part of the proof. Our proof of lower semicontinuity only requires the entropy functional (over a bounded domain) to have weakly compact level sets. The following result from [7] (Theorem 3.8) implies that our proof holds not just for h but also for any entropy whose convex conjugate is everywhere finite and differentiable.

LEMMA 2. *Let I be bounded and let $\phi : \mathbb{R} \to \bar{\mathbb{R}}$ be such that ϕ^* is everywhere finite and differentiable; then*

$$I_\phi(p) \;=\; \int_I \phi(p(x))\, dx$$

(i) *is strictly convex,* (ii) *has weakly compact level sets in $L^1(I)$, and* (iii) *$p_n \to p$ in $L^1(I)$ whenever $I_\phi(p_n) \to I_\phi(p)$ and $p_n \to p$ weakly in $L^1(I)$.*

We will also need the following useful lemma, which explicitly gives the MES for the case of two constraints.

LEMMA 3. *The two-constraint problem,*[2] *i.e.,*

$$\left|\begin{array}{rl} \textit{minimize} & \mathcal{I}_h(p) := \displaystyle\int_0^\infty h\big(p(x)\big)\,dx \\ \textit{s.t.} & d_0 = \displaystyle\int_0^\infty p(x)\,dx, \\ & d_1 = \displaystyle\int_0^\infty x\,p(x)\,dx, \end{array}\right.$$

has the explicit solution

$$\hat{p}(x) \;=\; \frac{d_0^2}{d_1} e^{-(d_0/d_1)x}.$$

Proof of Lemma 3. Let $\lambda_0 = \log \frac{d_0^2}{d_1}$ and $\lambda_1 = -\frac{d_0}{d_1}$. One readily checks that $\hat{p}(x)$ is feasible (satisfies the two constraints), and $\mathcal{I}_h(\hat{p}) = d_0 \log d_0^2/d_1 - 2d_0$. On the other hand,

$$D(\lambda_0, \lambda_1) = \lambda_0 d_0 \;+\; \lambda_1\, d_1 \;-\; \int_0^\infty e^{\lambda_0 + \lambda_1 x}\,dx \;=\; d_0 \log d_0^2/d_1 - 2d_0.$$

The result follows by (12)—in fact, the result would follow simply from weak duality, i.e., (12) with equality replaced by $\geq$, which always holds true. □

THEOREM 3. *The value function $\mathcal{V}$ is lower semicontinuous on its domain.*

Proof. The basis for our proof lies in the fact that the particular structure of the constraint functions allows us to rewrite all but the first two constraints as integrals over a finite domain. To this end, observe that for $j = 2, \ldots, m$, we have

$$\begin{aligned} d_j &= \int_{k_j}^\infty (x - k_j) p(x)\,dx \\ &= \int_0^\infty x p(x)\,dx - k_j \int_0^\infty p(x)\,dx + \int_0^{k_j} (k_j - x) p(x)\,dx \\ &= d_1 - k_j + \int_0^{k_j} (k_j - x) p(x)\,dx. \end{aligned} \tag{18}$$

Consequently, all constraints corresponding to $j > 1$ can be rewritten as

$$\int_0^M (k_j - x)^+ p(x)\,dx = \delta_j := d_j - d_1 + k_j,$$

where M is any constant greater than or equal to k_m.

With this in hand, suppose $\mathbf{d}, \mathbf{d}^{(n)} \in \mathrm{dom}\mathcal{V}$ ($d_0 = d_0^{(n)} = 1$) with $\mathbf{d}^{(n)} \to \mathbf{d}$ and for some constant C, $\mathcal{V}(\mathbf{d}^{(n)}) \leq C$ for all n. We prove that $\mathcal{V}(\mathbf{d}) \leq C$. To this end, pick a sequence $p^{(n)}$ such that $\mathbb{A}p^{(n)} = \mathbf{d}^{(n)}$ and $\mathcal{I}_h(p^{(n)}) \leq C + 2^{-n}$. Fix $M > k_m$ and define

$$d_{M,0}^{(n)} = \int_0^M p^{(n)}(x)\,dx \quad \text{and} \quad d_{M,1}^{(n)} = \int_0^M x\,p^{(n)}(x)\,dx.$$

[2]This constrained problem is used as a tool in our analysis. In the context of options, not only would $d_0 = 1$, but d_1 would also be predetermined by the risk-free interest rate.

Then $1 = d_{M,0}^{(n)} + \varepsilon_{M,0}^{(n)}$ and $d_1^{(n)} = d_{M,1}^{(n)} + \varepsilon_{M,1}^{(n)}$, where

$$\varepsilon_{M,0}^{(n)} = \int_M^\infty p^{(n)}(x)\,dx \quad \text{and} \quad \varepsilon_{M,1}^{(n)} = \int_M^\infty x\,p^{(n)}(x)\,dx.$$

Clearly, $0 \le d_{M,0}^{(n)} \le 1$ and $0 \le d_{M,1}^{(n)} \le d_1^{(n)} \to d_1$, so we have, up to taking a subsequence (not relabeled), that $d_{M,0}^{(n)}$ tends to some $d_{M,0}$ and $d_{M,1}^{(n)}$ tends to some $d_{M,1}$. Then $\varepsilon_{M,0}^{(n)} \to \varepsilon_{M,0} := 1 - d_{M,0}$ and $\varepsilon_{M,1}^{(n)} \to \varepsilon_{M,1} := d_1 - d_{M,1}$.

Assume for the moment that for some constant c,

$$\int_0^M h(p^{(n)}(x))\,dx < c. \tag{19}$$

Since $h^\star$ is everywhere finite and differentiable, Lemma 2 implies that there exists a subsequence (not relabeled) such that $p^{(n)}$ weakly converges to some p_M on $[0, M]$. Furthermore, p_M satisfies

$$\int_0^M (k_j - x)^+ p_M(x)\,dx = \delta_j, \qquad j > 1,$$

$$\int_0^M p_M(x)\,dx = d_{M,0} \le 1,$$

$$\int_0^M x\,p_M(x)\,dx = d_{M,1} \le d_1.$$

We note that either

(a) $d_{M,0} < 1$ and $d_{M,1} < d_1$, or

(b) $d_{M,0} = 1$ and $d_{M,1} = d_1$.

For case (a), we consider the two-constraint problem

$$\left|\begin{array}{rl} \text{minimize} & \mathcal{I}_h(\tilde{p}) := \displaystyle\int_M^\infty h\big(\tilde{p}(x)\big)\,dx \\ \text{s.t.} & \varepsilon_{M,0}^{(n)} = \displaystyle\int_M^\infty \tilde{p}(x)\,dx, \\ & \varepsilon_{M,1}^{(n)} = \displaystyle\int_M^\infty x\,\tilde{p}(x)\,dx. \end{array}\right.$$

By Lemma 3, this has an *explicit* solution

$$\tilde{p}^{(n)}(x) \;=\; \frac{(\varepsilon_{M,0}^{(n)})^2}{\varepsilon_{M,1}^{(n)} - M\varepsilon_{M,0}^{(n)}}\; e^{-\frac{\varepsilon_{M,0}^{(n)}}{\varepsilon_{M,1}^{(n)} - M\varepsilon_{M,0}^{(n)}}(x-M)}.$$

Note that on $[M, \infty)$ the entropy of $\tilde{p}^{(n)}$ is

$$\varepsilon_{M,0}^{(n)} \log\left(\frac{(\varepsilon_{M,0}^{(n)})^2}{\varepsilon_{M,1}^{(n)} - M\varepsilon_{M,0}^{(n)}}\right) - 2\varepsilon_{M,0}^{(n)}, \tag{20}$$

which, since $\varepsilon_{M,0}^{(n)}, \varepsilon_{M,1}^{(n)}$ are bounded, is bounded below. Moreover, $\tilde{p}^{(n)}(x)$ converges pointwise to

$$\tilde{p}(x) := \frac{\varepsilon_{M,0}^2}{\varepsilon_{M,1} - M\varepsilon_{M,0}}\, e^{-\frac{\varepsilon_{M,0}}{\varepsilon_{M,1}-M\varepsilon_{M,0}}(x-M)}.$$

Note that for case (a), $\varepsilon_{M,1} - M\varepsilon_{M,0} > 0$. Define $\hat{p}$ to be p_M on $[0, M]$ and $\tilde{p}$ on $[M, \infty)$. Then $\hat{p}$ is feasible for $\mathbf{d}$, and by taking a subsequence (not relabeled) of $p^{(n)}$, we have

$$\begin{aligned}\mathcal{I}_h(\hat{p}) &= \int_0^M h(p_M)dx + \int_M^\infty h(\tilde{p})dx \\ &\le \int_0^M h(p^{(n)})dx + \int_M^\infty h(\tilde{p}^{(n)})dx + 2^{-n} \\ &\le \int_0^\infty h(p^{(n)}) + 2^{-n} \le C + 2^{-n} + 2^{-n}.\end{aligned}$$

Above we used the weak lower semicontinuity of $\mathcal{I}_h$ on $[0, M]$ in the first inequality, and the fact that $\tilde{p}^{(n)}$ was optimal with respect to its constraints on $[M, \infty)$ in the second inequality. Letting $n \to \infty$ gives $\mathcal{V}(\mathbf{d}) \le C$.

In case (b), p_M (extended to be 0 on $[M, \infty)$) is feasible for $\mathbf{d}$. We have $p^{(n)} \to 0$ in L^1 on $[M, \infty)$, but since we do not know that I_h is lower semicontinuous on the infinite domain, we cannot immediately conclude anything about the limit of $\int_M^\infty h(p^{(n)})$. It suffices to prove that the liminf $\int_M^\infty h(p^{(n)}) = A$, with A for some finite $A \ge 0$. Then, by weak lower semicontinuity of $\mathcal{I}_h$ on $[0, M]$, we may pick a subsequence to find

$$\mathcal{V}(\mathbf{d}) \le \mathcal{I}_h(p_M) \le \int_0^M h(p_M)\,dx \,+\, A \le \int_0^\infty h(p^{(n)})dx \,+\, 2^{-n} \le C \,+\, 2^{-n+1}.$$

To this end, we note that since $\mathcal{I}_h(p^{(n)}) < C + 2^{-n}$ and $\int_0^M h(p^{(n)})$ is bounded below, liminf $\int_M^\infty h(p^{(n)})$ cannot be $+\infty$. Moreover, since both $\varepsilon_{M,0}^{(n)}$ and $\varepsilon_{M,1}^{(n)}$ tend to 0, the liminf of the entropies of the optimal $\tilde{p}^{(n)}$ (i.e., (20)) is greater than or equal to zero. Since $p^{(n)}$ restricted to $[M, \infty)$ always has greater entropy than $\tilde{p}^{(n)}$, liminf $\int_M^\infty h(p^{(n)})$ is some finite number $A \ge 0$.

Finally we address assumption (19). Suppose this did not hold; then (up to taking a subsequence) the entropy of $p^{(n)}$ on $[M, \infty)$ would have to approach $-\infty$. But this is impossible since we have shown above that the optimal (lowest entropy) distribution on $[M, \infty)$, over constraints for which $p^{(n)}$ restricted to $[M, \infty)$ is admissible, has entropy bounded below. □

COROLLARY 2. *Equality holds in* (9) *and* (12) *whenever the left-hand side is finite.*

Proof. See [18]. □

7. Remark. We briefly comment on the presence of a prior distribution. For a fixed distribution q (i.e., $q \in L^1(I)$, $\int_I q(x)\,dx = 1$), consider

$$(\mathcal{P}_q) \qquad \left|\begin{array}{l} \text{minimize} \quad \displaystyle\int_I p(x)\log\left(\frac{p(x)}{q(x)}\right)dx \\ \qquad \text{s.t.} \quad 1 = \displaystyle\int_I p(x)\,dx \quad \text{and} \quad d_j = \int_I c_j(x)p(x)\,dx. \end{array}\right.$$

Here we minimize the "entropic" distance to a prior distribution $q(x)$. This gives a more realistic approach to recovering the price distribution, as our previous model

is based upon the assumption that the only a priori guess for $p(x)$ is uniform. In practice, one may have a priori information that the unknown distribution could be, say, log-normal.

For the analysis to carry over, we require q to be bounded away from zero at $x = 0$. Particularly, we would require

$$e^{-ax} < q(x) < e^{bx} \qquad \text{a.e. for some positive constants } a, b.$$

This assumption may seem rather odd but it is simply a consequence of the structure of MESs. Note, for example, that the MES is never zero when $x = 0$ regardless of the moment constraints.

The main modification in the results would be that the measure dx in the dual function D is replaced with $q(x)dx$, with the corresponding adjustment in the closed form of the primal solution. Note that this would prevent one from carrying out the integration performed in (16) for an explicit representation. In this way, the uniform prior is rather special.

REFERENCES

[1] M. Avellaneda, *The minimum-entropy algorithm and related methods for calibrating asset-pricing models*, in Proceedings of the International Congress of Mathematicians, Vol. III, Doc. Math., Berlin, 1998, pp. 545–563.

[2] M. Avellaneda, C. Friedman, R. Holmes, and D. Samperi, *Calibrating volatility surfaces via relative entropy minimization*, Appl. Math. Finance, 4 (1997), pp. 37–64.

[3] J. M. Borwein and A. S. Lewis, *Convex Analysis and Nonlinear Optimization: Theory and Examples*, CMS Books Math./Ouvrages Math. SMC 3, Springer-Verlag, New York, 2000.

[4] J. M. Borwein and A. S. Lewis, *Partially finite convex programming. Part* I: *Quasi relative interiors and duality theory*, Math. Programming, 57 (1992), pp. 15–48.

[5] J. M. Borwein and A. S. Lewis, *Partially finite convex programming. Part* II: *Explicit lattice models*, Math. Programming, 57 (1992), pp. 49–83.

[6] J. M. Borwein and A. S. Lewis, *Partially-finite programming in L_1 and the existence of maximum entropy estimates*, SIAM J. Optim., 3 (1993), pp. 248–267.

[7] J. M. Borwein and A. S. Lewis, *Strong rotundity and optimization*, SIAM J. Optim., 4 (1994), pp. 146–158.

[8] J. M. Borwein and M. A. Limber, *Underdetermined Moment Problems: A Case for Convex Analysis*, invited, SIAM Conference on Optimization, SIAM, Philadelphia, 1994.

[9] P. W. Buchen and M. Kelley, *The maximum entropy distribution of an asset inferred from option prices*, J. Financial and Quantitative Analysis, 31 (1996), pp. 143–159.

[10] J. Cover and J. A. Thomas, *Elements of Information Theory*, John Wiley, New York, 1991.

[11] J. B. Hiriart-Urruty and C. Lemaréchal, *Convex Analysis and Minimization Algorithms. Vol.* I: *Fundamentals*, Springer-Verlag, Berlin, 1993.

[12] J. B. Hiriart-Urruty and C. Lemaréchal, *Convex Analysis and Minimization Algorithms. Vol.* II: *Advanced Theory and Bundle Methods*, Springer-Verlag, Berlin, 1993.

[13] J. C. Hull, *Options, Futures, and Other Derivative Securities*, Prentice-Hall, Englewood Cliffs, NJ, 1993.

[14] D. M. Lin and E. K. Wong, *A survey on the maximum entropy method and parameter spectral estimation*, Phys. Rep., 193 (1990), pp. 41–135.

[15] P. Maréchal, *On the principle of maximum entropy as a methodology for solving linear inverse problems*, in Probability Theory and Mathematical Statistics, B. Grigelionis et al., eds., VPS/TEV, Zeist, The Netherlands, 1999, pp. 481–492.

[16] P. Maréchal, *A note on entropy optimization*, in Approximation, Optimization and Mathematical Economics, M. Lassonde, ed., Physica-Verlag, Heidelberg, 2001, pp. 205–211.

[17] R. T. Rockafellar, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.

[18] R. T. Rockafellar, *Conjugate Duality and Optimization*, CBMS-NSF Regional Conf. Ser. in Appl. Math. 16, SIAM, Philadelphia, 1974.

[19] R. T. Rockafellar, *Convex integral functionals and duality*, in Contributions to Nonlinear Functional Analysis, E. H. Zarantonello, ed., Academic Press, New York, 1971, pp. 215–236.

SIAM J. OPTIM.
Vol. 14, No. 2, pp. 479–499

AN INTERIOR POINT METHOD WITH A PRIMAL-DUAL QUADRATIC BARRIER PENALTY FUNCTION FOR NONLINEAR OPTIMIZATION*

HIROSHI YAMASHITA† AND HIROSHI YABE‡

Abstract. In this paper, we are concerned with a primal-dual interior point method for solving nonlinearly constrained optimization problems, in which Newton-like methods are applied to the shifted barrier KKT conditions. We propose a new primal-dual merit function, called the primal-dual quadratic barrier penalty function, framework of line search methods, and show the global convergence properties of our method. Asymptotic superlinear convergence of the method is achieved by carefully controlling the parameters. Some numerical experiments are presented to show the performance of our method.

Key words. constrained optimization, primal-dual interior point method, primal-dual quadratic barrier penalty function, global convergence, superlinear convergence

AMS subject classifications. 90C30, 90C51, 90C53

DOI. S1052623499355533

1. Introduction.

In this paper, we consider the constrained optimization problem

$$\begin{array}{lll} \text{minimize} & f(x), & x \in \mathbf{R}^n, \\ \text{subject to} & g(x) = 0, & x_i \geq 0,\ i \in I_P, \end{array} \tag{1}$$

where we assume that the functions $f : \mathbf{R}^n \to \mathbf{R}$ and $g : \mathbf{R}^n \to \mathbf{R}^m$ are twice continuously differentiable, and I_P is a subset of the index set $\{1, 2, \ldots, n\}$. Let $p = |I_P| > 0$ and E be a $p \times n$ matrix whose rows consist of $e_i^t,\ i \in I_P$, where $e_i \in \mathbf{R}^n$ denotes the ith column vector of the identity matrix. Then problem (1) is written as

$$\begin{array}{lll} \text{minimize} & f(x), & x \in \mathbf{R}^n, \\ \text{subject to} & g(x) = 0, & Ex \geq 0. \end{array}$$

In what follows, we use the notation

$$x' \equiv Ex \ \in \mathbf{R}^p$$

for simplicity.

Let the Lagrangian function of the above problem be defined by

$$L(w) = f(x) - y^t g(x) - z^t Ex = f(x) - y^t g(x) - z^t x',$$

where $w = (x, y, z)^t$, and $y \in \mathbf{R}^m$ and $z \in \mathbf{R}^p$ are the Lagrange multiplier vectors which correspond to the equality and inequality constraints, respectively. Then

*Received by the editors April 30, 1999; accepted for publication (in revised form) May 21, 2003; published electronically November 6, 2003.
http://www.siam.org/journals/siopt/14-2/35553.html
†Mathematical Systems, Inc., 2-4-3, Shinjuku, Shinjuku-ku, Tokyo, Japan (hy@msi.co.jp).
‡Department of Mathematical Information Science, Faculty of Science, Tokyo University of Science, 1-3, Kagurazaka, Shinjuku-ku, Tokyo, Japan (yabe@rs.kagu.tus.ac.jp).

Karush–Kuhn–Tucker (KKT) conditions for optimality of the above problem are given by

$$r_0(w) \equiv \begin{pmatrix} \nabla_x L(w) \\ g(x) \\ X'Ze \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} \tag{2}$$

and

$$x' \geq 0, \qquad z \geq 0, \tag{3}$$

where

$$\begin{aligned} \nabla_x L(w) &= \nabla f(x) - A(x)^t y - E^t z, \\ A(x) &= \begin{pmatrix} \nabla g_1(x)^t \\ \vdots \\ \nabla g_m(x)^t \end{pmatrix}, \\ X' &= \operatorname{diag}\left(x'_1, \ldots, x'_p\right), \\ Z &= \operatorname{diag}\left(z_1, \ldots, z_p\right), \\ e &= (1, \ldots, 1)^t \in \mathbf{R}^p. \end{aligned}$$

To solve the above problem by a primal-dual interior point method, many researchers have applied Newton's method to the equality part of the barrier KKT conditions

$$\begin{pmatrix} \nabla_x L(w) \\ g(x) \\ X'Ze - \mu e \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \quad x' > 0, \quad \text{and} \quad z > 0, \tag{4}$$

where $\mu > 0$ is a barrier parameter. In this case, the Newton step $\Delta w = (\Delta x, \Delta y, \Delta z)^t$ is defined by a solution of the Newton equation

$$\begin{pmatrix} G & -A(x)^t & -E^t \\ A(x) & 0 & 0 \\ ZE & 0 & X' \end{pmatrix} \Delta w = - \begin{pmatrix} \nabla_x L(w) \\ g(x) \\ X'Ze - \mu e \end{pmatrix},$$

where we use the relation $X'Ze = X'z = ZEx$. The matrix G is $\nabla_x^2 L(w)$ or a quasi-Newton approximation to the Hessian matrix.

To globalize the algorithm, Yamashita [18] introduced the barrier penalty function $\Phi(\bullet, \mu) : S \to \mathbf{R}$, which is defined by

$$\Phi(x, \mu) = f(x) - \mu \sum_{i=1}^{p} \log x'_i + \rho \sum_{i=1}^{m} |g_i(x)|, \tag{5}$$

where μ and ρ are given positive constants, and

$$S = \{x \in \mathbf{R}^n \,|\, x' > 0\}. \tag{6}$$

Yamashita proposed using the function (5) as a merit function, based on the fact that if ρ is sufficiently large, the necessary condition for the optimality of the barrier penalty function minimization problem for a given $\mu > 0$ is the barrier KKT conditions. The

function (5) is a merit function in the primal space. Using this function for primal variable x, Yamashita [18], Yamashita and Tanabe [20], and Yamashita, Yabe, and Tanabe [22] showed the global convergence properties of their primal-dual interior point methods within the framework of line search strategy and trust region strategy, respectively. For the variable z, the step size is controlled by a box constraint, and for the variable y, several step sizes could be theoretically possible, but the one equal to the step size of z is adopted in this paper. Both algorithms are shown to be quite efficient through numerical experiments. Different primal merit functions were used, for example, by Akrotirianakis and Rustem [1, 2] and Vanderbei and Shanno [16] within the framework of line search strategies. Primal merit functions within the framework of trust region strategies have also been dealt with by, for example, Byrd, Gilbert, and Nocedal [4], Byrd, Hribar, and Nocedal [5], Conn et al. [7], and Dennis, Heinkenschloss, and Vicente [9]. Some researchers have considered primal-dual merit functions within the framework of line search strategies (see, for example, Argaez and Tapia [3], El-Bakry et al. [10], and Forsgren and Gill [12]). Specifically, Argaez and Tapia combined the augmented Lagrangian function and the barrier function as the merit function. El-Bakry et al. used the residual function of the KKT conditions for optimality. Forsgren and Gill derived a differentiable primal-dual merit function based on shifted barrier KKT conditions, which will be discussed below. On the other hand, superlinear convergence properties of primal-dual methods based on solving the barrier KKT conditions have been studied by several authors, for example, Martinez, Parada, and Tapia [14], El-Bakry et al. [10], Yamashita and Yabe [21], Yabe and Yamashita [17], Yamashita, Yabe, and Tanabe [22], and Byrd, Liu, and Nocedal [6].

In this paper, we consider a more conventional merit function,

$$F_0(x,\mu) = f(x) - \mu \sum_{i=1}^{p} \log x'_i + \frac{1}{2\mu} \sum_{i=1}^{m} g_i(x)^2, \tag{7}$$

which is extensively described in a book by Fiacco and McCormick [11]. We also call this function the barrier penalty function. To discriminate this function from (5), we may call this the quadratic barrier penalty function, whereas the function defined in (5) may be called the l_1 barrier penalty function.

The necessary condition for the optimality of the problem

$$\text{minimize} \quad F_0(x,\mu), \quad x \in S,$$

is

$$\nabla F_0(x,\mu) = \nabla f(x) - \mu E^t (X')^{-1} e + \frac{1}{\mu} \sum_{i=1}^{m} g_i(x) \nabla g_i(x) = 0 \tag{8}$$

and $x' > 0$, where S is defined by (6). As in [11, 12, 18], we introduce the variables y and z by $y = -g(x)/\mu$ and $z = \mu(X')^{-1}e$. Then the above conditions are written as

$$r(w,\mu) \equiv \begin{pmatrix} \nabla f(x) - A(x)^t y - E^t z \\ g(x) + \mu y \\ X'Ze - \mu e \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} \tag{9}$$

and $x' > 0, z > 0$ (see, for example, [11]). We call these conditions the shifted barrier KKT (SBKKT) conditions. It should be noted that we treat x, y, and z as

independent variables. These conditions are also considered by Forsgren and Gill [12]. Based on these conditions, they proposed a differentiable primal-dual merit function[1]

$$M_{FG}(x,y,z) = F_0(x,\mu) + \frac{\rho}{2\mu}\|g(x)+\mu y\|^2 - \mu\rho\sum_{i=1}^{p}\left(\log\frac{x_i'z_i}{\mu} + 1 - \frac{x_i'z_i}{\mu}\right), \tag{10}$$

where $F_0(x,\mu)$ is defined by (7) and ρ is a positive penalty parameter. They showed that the Newton step for the SBKKT conditions becomes a descent search direction for this merit function, but global convergence property was not proved in [12].

The main purpose of this paper is to analyze global convergence and local behavior of the primal-dual interior point method based on the SBKKT conditions (9). This paper is organized as follows. We will first propose a differentiable primal-dual merit function different from (10) in section 2.2 and show global convergence property within the framework of the line search strategy in section 2.3. Furthermore in section 3, superlinear convergence of the method to a point satisfying the SBKKT conditions will be proved. Finally preliminary numerical results and concluding remarks will be presented in section 4.

We call w satisfying $x' > 0$ and $z > 0$ an interior point. The algorithm in this paper will generate such interior points. In what follows, the subscript k denotes an iteration count in the inner iteration or in the outer iteration. Let $\|\cdot\|$ denote the l_2 norm for vectors and the operator norm induced from the l_2 vector norm for matrices. Let $\mathbf{R}^p_+ = \{z \in \mathbf{R}^p |\ z > 0\}$.

2. Algorithm and its global convergence.

2.1. Outer iteration. A prototype of the algorithm that uses the SBKKT conditions is described as follows.

ALGORITHM IP.

Step 0. (Initialize) Set $\varepsilon > 0$, $M_c > 0$, and $k = 0$. Let a positive sequence $\{\mu_k\}, \mu_k \downarrow 0$ be given.

Step 1. (Approximate SBKKT point) Find an interior point w_{k+1} that satisfies

$$\|r(w_{k+1},\mu_k)\| \le M_c\mu_k. \tag{11}$$

Step 2. (Termination) If $\|r_0(w_{k+1})\| \le \varepsilon$, then stop.

Step 3. (Update) Set $k := k+1$ and go to Step 1.

We note that the barrier parameter sequence $\{\mu_k\}$ in Algorithm IP need not be determined beforehand. The value of each μ_k may be set adaptively as the iteration proceeds. We call condition (11) the approximate SBKKT condition, and we call a point that satisfies this condition an approximate SBKKT point.

The following theorem shows the global convergence property of Algorithm IP.

THEOREM 2.1. *Let $\{w_k\}$ be an infinite sequence generated by Algorithm* IP. *Suppose that the sequences $\{x_k\}$ and $\{y_k\}$ are bounded. Then $\{z_k\}$ is bounded, and any accumulation point of $\{w_k\}$ satisfies KKT conditions* (2) *and* (3).

Proof. Assume that $\{z_k\}$ is not bounded, i.e., that there exists an i such that $(E^t z_k)_i \to \infty$. Equation (11) yields

$$\left|\frac{(\nabla f(x_k) - A(x_k)^t y_k)_i}{(E^t z_k)_i} - 1\right| \le M_c\frac{\mu_{k-1}}{(E^t z_k)_i}.$$

[1] Forsgren and Gill [12] originally dealt with the minimization problem min $f(x)$ subject to $g(x) = 0$ and $h(x) \ge 0$, where $f : \mathbf{R}^n \to \mathbf{R}$, $g : \mathbf{R}^n \to \mathbf{R}^m$, $h : \mathbf{R}^n \to \mathbf{R}^p$. In this paper, their merit function is rewritten for problem (1).

The sequences $\{x_k\}$ and $\{y_k\}$ are bounded, and f and g are twice continuously differentiable, and $\mu_k \to +0$ as $k \to \infty$. This implies that $1 \leq 0$, which is a contradiction. Thus the sequence $\{z_k\}$ is bounded.

Let $\hat{w}$ be any accumulation point of $\{w_k\}$. Since the sequences $\{w_k\}$ and $\{\mu_k\}$ satisfy (11) for each k and μ_k approaches zero, $r_0(\hat{w}) = 0$ follows from the definition of $r(w, \mu)$. Therefore the proof is complete. □

2.2. Solving the SBKKT conditions. In this subsection we consider a method for solving the SBKKT conditions approximately for a given $\mu > 0$ (Step 1 of Algorithm IP). Therefore the index k denotes the inner iteration count for a given $\mu > 0$. We note that $x'_k > 0$ and $z_k > 0$ for all k in the following. The Newton-like iteration for solving (9) is defined by

$$J_k \Delta w_k = -r(w_k, \mu), \tag{12}$$

where the matrix J_k is given by

$$J_k = \begin{pmatrix} G_k & -A(x_k)^t & -E^t \\ A(x_k) & \mu I & 0 \\ Z_k E & 0 & X'_k \end{pmatrix},$$

and the matrix G_k is $\nabla_x^2 L(w_k)$ or its approximation. If $G_k = \nabla_x^2 L(w_k)$, then J_k becomes the Jacobian matrix of $r(w, \mu)$ at w_k. We note that (12) can be represented by the forms

$$G_k \Delta x_k - A(x_k)^t \Delta y_k - E^t \Delta z_k = -\nabla_x L(w_k), \tag{13}$$

$$A(x_k) \Delta x_k + \mu \Delta y_k = -g(x_k) - \mu y_k, \tag{14}$$

$$Z_k E \Delta x_k + X'_k \Delta z_k = \mu e - X'_k z_k. \tag{15}$$

The following lemma gives a sufficient condition for (12) to be solvable.

LEMMA 2.2. *If the matrix* $G_k + E^t (X'_k)^{-1} Z_k E + \frac{1}{\mu} A(x_k)^t A(x_k)$ *is positive definite, then the matrix* J_k *is nonsingular.*

Proof. Consider the equation

$$J_k \begin{pmatrix} v_x \\ v_y \\ v_z \end{pmatrix} = 0$$

for $(v_x, v_y, v_z)^t \in \mathbf{R}^n \times \mathbf{R}^m \times \mathbf{R}^p$. Then we have

$$\left(G_k + E^t (X'_k)^{-1} Z_k E + \frac{1}{\mu} A(x_k)^t A(x_k) \right) v_x = 0,$$

$$v_y = -\mu^{-1} A(x_k) v_x,$$

$$v_z = -(X'_k)^{-1} Z_k E v_x.$$

By the assumption we obtain $v_x = 0$, and therefore $v_y = 0$ and $v_z = 0$. This proves the lemma. □

From (13), (14), and (15), we have

$$\left(G_k + E^t (X'_k)^{-1} Z_k E + \frac{1}{\mu} A(x_k)^t A(x_k) \right) \Delta x_k = -\nabla F_0(x_k, \mu). \tag{16}$$

Therefore it is easy to see that under appropriate assumptions the function $F_0(x, \mu)$ can be used as a merit function as in [18]. Because $F_0(x, \mu)$ depends only on the primal variables, we should use a method similar to the one which is given in [18] for controlling the step sizes for dual variables. Instead of following this possibility, we consider a merit function in the primal-dual space in this paper. As noted in the introduction, primal-dual merit functions based on the barrier KKT conditions (4) have been proposed by Argaez and Tapia [3] and El-Bakry et al. [10], while a primal-dual merit function based on the SBKKT conditions (9) has been proposed by Forsgren and Gill [12].

To have a merit function which has a minimum point at the SBKKT point, and which gives a descent direction with a Newton step, it is natural to consider

$$F_0(x, \mu) + \frac{\rho}{2} \|g(x) + \mu y\|^2 + \frac{\rho}{2} \|X'z - \mu e\|^2,$$

where ρ is a positive constant. We note that the second and third terms correspond to the second and third components in $r(w, \mu)$, respectively. However, this function does not prevent each component of z from tending to zero, and therefore cannot give a globally convergent algorithm unless an appropriate procedure is devised. Thus we need a sort of barrier term for the variable z. In this paper we propose the following function, called the primal-dual barrier penalty function:

$$F(w, \mu) = F_0(x, \mu) + \rho \log \frac{(x')^t z/p + \|g(x) + \mu y\|^2 + \|X'z - \mu e\|^2}{\left(\prod_{i=1}^p x_i' z_i\right)^{1/p}}, \tag{17}$$

where p is defined in section 1 and $\rho > 0$ is a constant. This function is a modification of the primal-dual merit function proposed by Yamashita [19]. The denominator in the second term is to prevent each z_i from tending to 0. For notational convenience we denote the expression in the last term in (17) by $\rho\phi(w)$, i.e.,

$$\begin{aligned} \phi(w) &\equiv \log \frac{(x')^t z/p + \|g(x) + \mu y\|^2 + \|X'z - \mu e\|^2}{\left(\prod_{i=1}^p x_i' z_i\right)^{1/p}} \\ &= \log \left((x')^t z/p + \|g(x) + \mu y\|^2 + \|X'z - \mu e\|^2 \right) - \frac{1}{p} \sum_{i=1}^p \log x_i' z_i. \end{aligned} \tag{18}$$

For later convenience we quote two well-known relations between arithmetic and geometric means:

$$\frac{(x')^t z}{p} \geq \left(\prod_{i=1}^p x_i' z_i \right)^{1/p}, \tag{19}$$

$$\sum_{i=1}^p \frac{1}{p x_i' z_i} \geq \frac{1}{\left(\prod_{i=1}^p x_i' z_i\right)^{1/p}}, \tag{20}$$

where $x' > 0$ and $z > 0$. In the above inequalities, the equalities hold if and only if $x_1' z_1 = \cdots = x_p' z_p$.

The function defined by (18) has the following properties.

LEMMA 2.3. *Suppose that $x' > 0$ and $z > 0$. The following relationships hold:*

(i) $\phi(w) \geq 0$.

(ii) $\phi(w) = 0$ *if and only if* $g(x) + \mu y = 0$ *and* $X'z - \mu e = 0$.

Proof. (i) Inequality (19) yields

$$\frac{(x')^t z}{p} + \|g(x) + \mu y\|^2 + \|X'z - \mu e\|^2 \geq \left(\prod_{i=1}^{p} x_i' z_i\right)^{1/p}.$$

Thus we have

$$\phi(w) \geq \log 1 = 0.$$

(ii) Suppose that $g(x) + \mu y = 0$ and $X'z - \mu e = 0$. Since $x_1' z_1 = \cdots = x_p' z_p = \mu$, we have

$$\frac{(x')^t z}{p} = \left(\prod_{i=1}^{p} x_i' z_i\right)^{1/p},$$

which implies $\phi(w) = \log 1 = 0$.

Conversely suppose that $\phi(w) = 0$. It follows from (19) that

$$\frac{(x')^t z}{p} + \|g(x) + \mu y\|^2 + \|X'z - \mu e\|^2 = \left(\prod_{i=1}^{p} x_i' z_i\right)^{1/p} \leq \frac{(x')^t z}{p},$$

which implies

$$\|g(x) + \mu y\|^2 + \|X'z - \mu e\|^2 \leq 0.$$

Therefore the proof is complete. □

Now we calculate the derivatives of the merit function:

$$\nabla F(w, \mu) = \begin{pmatrix} \nabla F_0(x, \mu) + \rho \nabla_x \phi(w) \\ \rho \nabla_y \phi(w) \\ \rho \nabla_z \phi(w) \end{pmatrix}, \tag{21}$$

where

$$\begin{aligned}
\nabla_x \phi(w) &= \frac{E^t z/p + 2A(x)^t (g(x) + \mu y) + 2E^t Z(X'z - \mu e)}{(x')^t z/p + \|g(x) + \mu y\|^2 + \|X'z - \mu e\|^2} - \frac{E^t (X')^{-1} e}{p}, \\
\nabla_y \phi(w) &= \frac{2\mu(g(x) + \mu y)}{(x')^t z/p + \|g(x) + \mu y\|^2 + \|X'z - \mu e\|^2}, \\
\nabla_z \phi(w) &= \frac{x'/p + 2X'(X'z - \mu e)}{(x')^t z/p + \|g(x) + \mu y\|^2 + \|X'z - \mu e\|^2} - \frac{Z^{-1} e}{p}.
\end{aligned}$$

The following lemma shows that an SBKKT point is equivalent to a stationary point of the function $F(w, \mu)$ and gives the relationship between minimizers of functions F and F_0.

LEMMA 2.4. *Suppose that $x' > 0$ and $z > 0$.*

(i) *The following statements are equivalent:*
- (a) $r(w, \mu) = 0$.
- (b) $\nabla F_0(x, \mu) = 0,\ g(x) + \mu y = 0,\ X'z - \mu e = 0$.
- (c) $\nabla F(w, \mu) = 0$.

(ii) *A point $w = (x, y, z)$ is an unconstrained local minimizer of the function $F(w, \mu)$ if and only if x is an unconstrained local minimizer of the function $F_0(x, \mu)$ and w satisfies $\phi(w) = 0$.*

Proof. (i) The equivalence of (a) and (b) is obvious from (8) and (9).

The equivalence of (b) and (c) comes from (21). If $\nabla F_0(x, \mu) = 0$, $g(x) + \mu y = 0$, and $X'z - \mu e = 0$, then we have $\nabla F(w, \mu) = 0$. Conversely assume that $\nabla F(w, \mu) = 0$. Then it follows from the relations $\nabla_y \phi(w) = 0$ and $\nabla_z \phi(w) = 0$ that

$$g(x) + \mu y = 0$$

and

$$\frac{x'/p + 2X'(X'z - \mu e)}{(x')^t z/p + \|X'z - \mu e\|^2} - \frac{Z^{-1}e}{p} = 0. \tag{22}$$

Multiplying (22) through by $Z(X')^{-1}$, one obtains

$$\frac{z/p + 2Z(X'z - \mu e)}{(x')^t z/p + \|X'z - \mu e\|^2} - \frac{(X')^{-1}e}{p} = 0,$$

which implies $\nabla_x \phi(w) = 0$, and we have

$$\nabla F_0(x, \mu) = \nabla_x F(w, \mu) = 0.$$

Equation (22) also yields

$$2(X'z - \mu e) = \frac{1}{p}\left(\frac{(x')^t z}{p} + \|X'z - \mu e\|^2\right)(X'Z)^{-1}e - \frac{1}{p}e.$$

Multiplying both sides of the above equation by $(X'z - \mu e)^t$, we have

$$\begin{aligned} 2\|X'z - \mu e\|^2 &= \left(\frac{(x')^t z}{p} + \|X'z - \mu e\|^2\right) - \frac{(x')^t z}{p} \\ &\quad - \frac{\mu}{p}\left(\frac{(x')^t z}{p} + \|X'z - \mu e\|^2\right) e^t (X'Z)^{-1}e + \mu \\ &= \|X'z - \mu e\|^2 + \mu - \frac{\mu}{p}\left(\frac{(x')^t z}{p} + \|X'z - \mu e\|^2\right) e^t (X'Z)^{-1}e. \end{aligned}$$

Then we have

$$\begin{aligned} \left(1 + \frac{\mu}{p} e^t (X'Z)^{-1}e\right)\|X'z - \mu e\|^2 &= \mu - \mu \frac{(x')^t z}{p} \frac{e^t (X'Z)^{-1}e}{p} \\ &\le \mu - \mu \frac{\left(\prod_{i=1}^p x_i' z_i\right)^{1/p}}{\left(\prod_{i=1}^p x_i' z_i\right)^{1/p}} \quad \text{(from (19) and (20))} \\ &= 0, \end{aligned}$$

which implies $X'z - \mu e = 0$ since $\frac{\mu}{p} e^t (X'Z)^{-1}e > 0$.

(ii) For fixed x, let

$$\hat{y}(x) = -\frac{1}{\mu} g(x) \qquad \text{and} \qquad \hat{z}(x) = \mu (X')^{-1} e.$$

Since a point $(x, \hat{y}(x), \hat{z}(x))$ minimizes the function $\phi(w)$, and its minimum value is zero by Lemma 2.3, we have

$$\min_{y,z} F(w,\mu) = F(x, \hat{y}(x), \hat{z}(x), \mu) = F_0(x,\mu).$$

Therefore by combining the above with result (i) of this lemma, the proof is complete. □

In the following, we set $\Delta x' = E\Delta x$. To derive an upper bound on the directional derivative of F, we first calculate the one for ϕ.

(23)

$$\begin{aligned}
&\nabla\phi(w)^t\Delta w \\
&= \frac{\left(z^t\Delta x' + (x')^t\Delta z\right)/p + 2(A(x)\Delta x + \mu\Delta y)^t(g(x)+\mu y) + 2\left(Z\Delta x' + X'\Delta z\right)^t(X'z-\mu e)}{(x')^t z/p + \|g(x)+\mu y\|^2 + \|X'z-\mu e\|^2} \\
&\quad - \frac{1}{p}\sum_{i=1}^{p}\frac{z_i\Delta x_i' + x_i'\Delta z_i}{x_i' z_i}.
\end{aligned}$$

LEMMA 2.5. *If* Δw_k *solves* (12), *then*

$$\nabla\phi(w_k)^t\Delta w_k \le \frac{-\|g(x_k)+\mu y_k\|^2 - \|X_k' z_k - \mu e\|^2}{(x_k')^t z_k/p + \|g(x_k)+\mu y_k\|^2 + \|X_k' z_k - \mu e\|^2}. \tag{24}$$

Proof. Since (15) yields

$$\begin{aligned}
z_k^t\Delta x_k' + (x_k')^t\Delta z_k &= e^t(Z_k\Delta x_k' + X_k'\Delta z_k) \\
&= p\mu - (x_k')^t z_k,
\end{aligned}$$

we have by (14), (15), and (23)

$$\begin{aligned}
\nabla\phi(w_k)^t\Delta w_k &= \frac{\mu - (x_k')^t z_k/p - 2\|g(x_k)+\mu y_k\|^2 - 2\|X_k' z_k - \mu e\|^2}{(x_k')^t z_k/p + \|g(x_k)+\mu y_k\|^2 + \|X_k' z_k - \mu e\|^2} \\
&\quad - \sum_{i=1}^{p}\frac{\mu - (x_k')_i(z_k)_i}{p(x_k')_i(z_k)_i} \\
&= \frac{\mu - \|g(x_k)+\mu y_k\|^2 - \|X_k' z_k - \mu e\|^2}{(x_k')^t z_k/p + \|g(x_k)+\mu y_k\|^2 + \|X_k' z_k - \mu e\|^2} \\
&\quad - \sum_{i=1}^{p}\frac{\mu}{p(x_k')_i(z_k)_i}.
\end{aligned}$$

From relations (19) and (20), we obtain

$$\begin{aligned}
&\frac{\mu - \|g(x_k)+\mu y_k\|^2 - \|X_k' z_k - \mu e\|^2}{(x_k')^t z_k/p + \|g(x_k)+\mu y_k\|^2 + \|X_k' z_k - \mu e\|^2} - \sum_{i=1}^{p}\frac{\mu}{p(x_k')_i(z_k)_i} \\
&\le \frac{p\mu}{(x_k')^t z_k} - \frac{\mu}{\left(\prod_{i=1}^{p}(x_k')_i(z_k)_i\right)^{1/p}} - \frac{\|g(x_k)+\mu y_k\|^2 + \|X_k' z_k - \mu e\|^2}{(x_k')^t z_k/p + \|g(x_k)+\mu y_k\|^2 + \|X_k' z_k - \mu e\|^2} \\
&\le \frac{-\|g(x_k)+\mu y_k\|^2 - \|X_k' z_k - \mu e\|^2}{(x_k')^t z_k/p + \|g(x_k)+\mu y_k\|^2 + \|X_k' z_k - \mu e\|^2}.
\end{aligned}$$

This proves the lemma. □

LEMMA 2.6. *If* Δw_k *solves* (12), *then*

$$\nabla F(w_k,\mu)^t\Delta w_k \le -\Delta x_k^t(G_k + E^t(X_k')^{-1}Z_kE + \frac{1}{\mu}A(x_k)^tA(x_k))\Delta x_k \\ -\rho\frac{\|g(x_k)+\mu y_k\|^2 + \|X_k'z_k-\mu e\|^2}{(x_k')^tz_k/p + \|g(x_k)+\mu y_k\|^2 + \|X_k'z_k-\mu e\|^2}.$$

Proof. From (16) and (21), we obtain

$$\nabla F(w_k,\mu)^t\Delta w_k = -\Delta x_k^t(G_k + E^t(X_k')^{-1}Z_kE + \frac{1}{\mu}A(x_k)^tA(x_k))\Delta x_k \\ +\rho\nabla\phi(w_k)^t\Delta w_k, \tag{25}$$

which proves the lemma in view of (24). □

LEMMA 2.7. *Assume that* Δw_k *solves* (12). *If* $\Delta x_k = 0$, $g(x_k)+\mu y_k = 0$, *and* $X_k'z_k - \mu e = 0$, *then* w_k *is an SBKKT point.*

Proof. $\Delta x_k = 0$ means $\nabla F_0(x_k,\mu) = 0$ from (16). Thus from Lemma 2.4(i), $r(w_k,\mu) = 0$ follows. □

Note that this lemma shows that if the matrix $G_k+E^t(X_k')^{-1}Z_kE+\frac{1}{\mu}A(x_k)^tA(x_k)$ is positive definite and w_k is not an SBKKT point, then the direction Δw_k is a descent direction for the primal-dual barrier penalty function from Lemma 2.6. Furthermore under these conditions, it follows from (16) that Δx_k is a descent direction for the barrier penalty function $F_0(x,\mu)$.

2.3. Line search algorithm. To obtain a globally convergent algorithm to an SBKKT point for a fixed $\mu > 0$, we modify the basic Newton iteration. Our iterations take the form

$$w_{k+1} = w_k + \alpha_k\Delta w_k,$$

where α_k is a step size determined by the line search procedure described below.

The main iteration is to decrease the value of the primal-dual barrier penalty function $F(w,\mu)$ for fixed μ. Thus the step size is determined by the sufficient decrease rule of the merit function. We adopt Armijo's rule. At the point w_k, we calculate the maximum allowed step to the boundary of the feasible region by

$$\alpha_{k\max} = \min\left\{\min_i\left\{-\frac{(x_k')_i}{(\Delta x_k')_i}\,\middle|\,(\Delta x_k')_i < 0\right\}, \min_i\left\{-\frac{(z_k)_i}{(\Delta z_k)_i}\,\middle|\,(\Delta z_k)_i < 0\right\}\right\}.$$

A step to the next iterate is given by

$$\alpha_k = \bar\alpha_k\beta^{l_k}, \qquad \bar\alpha_k = \min\{\gamma\alpha_{k\max}, 1\},$$

where $\gamma \in (0,1)$ and $\beta \in (0,1)$ are fixed constants and l_k is the smallest nonnegative integer such that

$$F(w_k+\bar\alpha_k\beta^{l_k}\Delta w_k,\mu) - F(w_k,\mu) \le \varepsilon_0\bar\alpha_k\beta^{l_k}\nabla F(w_k,\mu)^t\Delta w_k,$$

where $\varepsilon_0 \in (0,1)$.

Now we give the line search algorithm, which is called Algorithm LS. This algorithm can be regarded as the inner iteration of Algorithm IP (see Step 1 of Algorithm IP). We also note that ε' given below corresponds to $M_c\mu$ in Algorithm IP.

ALGORITHM LS.

Step 0. (Initialize) Let $w_0 \in S \times \mathbf{R}^m \times \mathbf{R}^p_+$, and $\mu > 0$, $\rho > 0$. Set $\varepsilon' > 0$, $\gamma \in (0,1)$, $\beta \in (0,1)$, $\varepsilon_0 \in (0,1)$. Let $k = 0$.

Step 1. (Termination) If $\|r(w_k, \mu)\| \leq \varepsilon'$, then stop.

Step 2. (Compute direction) Calculate the direction Δw_k by (12).

Step 3. (Step size) Calculate

$$\alpha_{k\max} = \min\left\{ \min_i \left\{ -\frac{(x'_k)_i}{(\Delta x'_k)_i} \middle| (\Delta x'_k)_i < 0 \right\}, \min_i \left\{ -\frac{(z_k)_i}{(\Delta z_k)_i} \middle| (\Delta z_k)_i < 0 \right\} \right\},$$
$$\bar{\alpha}_k = \min\{\gamma \alpha_{k\max}, 1\}.$$

Find the smallest nonnegative integer l_k that satisfies

$$F(w_k + \bar{\alpha}_k \beta^{l_k} \Delta w_k, \mu) - F(w_k, \mu) \leq \varepsilon_0 \bar{\alpha}_k \beta^{l_k} \nabla F(w_k, \mu)^t \Delta w_k. \tag{26}$$

Calculate

$$\alpha_k = \bar{\alpha}_k \beta^{l_k}.$$

Step 4. (Update variables) Set

$$w_{k+1} = w_k + \alpha_k \Delta w_k.$$

Step 5. Set $k := k + 1$ and go to Step 1.

To prove global convergence of Algorithm LS, we need the following assumptions.

Assumption G.

(G1) The functions f and $g_i, i = 1, \ldots, m$, are twice continuously differentiable.

(G2) The sequence $\{w_k\}$ generated by Algorithm LS remains in a compact set Ω of $S \times \mathbf{R}^m \times \mathbf{R}_+$.

(G3) The matrix G_k is uniformly bounded and the matrix $G_k + E^t (X'_k)^{-1} Z_k E + \frac{1}{\mu} A(x_k)^t A(x_k)$ is uniformly positive definite.

Assumption (G2) ensures the existence of a limit point of the generated sequence as shown in the next theorem. This compactness of the generated sequence is derived if we assume the compactness of the level set of the function $F(w, \mu)$ at the initial point, for example, because the iterates give decreasing function values.

It follows from Assumptions (G1) and (G2) that the function $F(w, \mu)$ is bounded below on Ω. It is known that the step size rule (26) is well-defined under Assumptions (G1) and (G2) (see, for example, Theorem 6.3.2 in [8]). We note that if a quasi-Newton approximation is used for computing the matrix G_k, then we need the continuity of only the first-order derivatives of functions in Assumption (G1). A specific updating formula will be given in section 4 to show numerical experiments.

The following theorem gives a convergence of an infinite sequence generated by Algorithm LS.

THEOREM 2.8. *Suppose that Assumption* G *holds. Let an infinite sequence* $\{w_k\}$ *be generated by Algorithm* LS. *Then there exists at least one accumulation point of* $\{w_k\}$, *and any accumulation point of the sequence* $\{w_k\}$ *is an SBKKT point.*

Proof. By Assumption (G2), the sequence $\{w_k\}$ remains in a compact set and thus has at least one limit point. The compactness of $\{w_k\}$ implies that each component of x'_k and z_k is bounded above. Thus the term

$$\left(\prod_{i=1}^{p} x'_i z_i \right)^{1/p}$$

in the denominator and $\|X'z - \mu e\|$ in the numerator in (17) guarantee that each component of x'_k and z_k is bounded away from zero. Using this boundedness property and Assumption (G3), there exists a positive number M such that

$$\frac{\|v\|^2}{M} \leq v^t \left(G_k + E^t (X'_k)^{-1} Z_k E + \frac{1}{\mu} A(x_k)^t A(x_k) \right) v \leq M \|v\|^2 \quad \forall v \in \mathbf{R}^n \tag{27}$$

for all k. From (25) and (27), we have

$$\nabla F(w_k, \mu)^t \Delta w_k \leq -\frac{\|\Delta x_k\|^2}{M} + \rho \nabla \phi(w_k)^t \Delta w_k < 0, \tag{28}$$

and from (26),

$$\begin{aligned} F(w_{k+1}, \mu) - F(w_k, \mu) &\leq \varepsilon_0 \bar{\alpha}_k \beta^{l_k} \nabla F(w_k, \mu)^t \Delta w_k \\ &\leq -\varepsilon_0 \bar{\alpha}_k \beta^{l_k} \left(\frac{\|\Delta x_k\|^2}{M} - \rho \nabla \phi(w_k)^t \Delta w_k \right) \\ &< 0. \end{aligned} \tag{29}$$

Because the sequence $\{F(w_k, \mu)\}$ is decreasing and bounded below, the left-hand side of (29) converges to 0. Since the inverse of the coefficient matrix of (16) is uniformly bounded by (27), $\|\Delta x_k\|$ is uniformly bounded above. Then it follows from (14) and (15) that Δy_k and Δz_k are also uniformly bounded. Thus we conclude that $\|\Delta w_k\|$ is uniformly bounded above. Since $\liminf_{k\to\infty} (x'_k)_i > 0$ and $\liminf_{k\to\infty} (z_k)_i > 0$ for $i = 1, \ldots, p$, we have $\liminf_{k\to\infty} \bar{\alpha}_k > 0$.

We will prove that

$$\lim_{k\to\infty} \nabla F(w_k, \mu)^t \Delta w_k = 0. \tag{30}$$

Suppose that there exist an infinite subsequence $K \subset \{0, 1, \ldots\}$ and a δ such that

$$\left| \nabla F(w_k, \mu)^t \Delta w_k \right| \geq \delta > 0 \quad \forall k \in K. \tag{31}$$

Then we have $l_k \to \infty$, $k \in K$, from (29) because the left-most expression tends to zero, and therefore we can assume $l_k > 0$ for sufficiently large $k \in K$ without loss of generality. In particular, the point $w_k + \alpha_k \Delta w_k / \beta$ does not satisfy condition (26). Thus, we have

$$F(w_k + \alpha_k \Delta w_k / \beta, \mu) - F(w_k, \mu) > \varepsilon_0 \alpha_k \nabla F(w_k, \mu)^t \Delta w_k / \beta. \tag{32}$$

By the mean value theorem, there exists a $\theta_k \in (0, 1)$ such that

$$F(w_k + \alpha_k \Delta w_k / \beta, \mu) - F(w_k, \mu) = \alpha_k \nabla F(w_k + \theta_k \alpha_k \Delta w_k / \beta, \mu)^t \Delta w_k / \beta. \tag{33}$$

Then, from (32) and (33), we have

$$\varepsilon_0 \nabla F(w_k, \mu)^t \Delta w_k < \nabla F(w_k + \theta_k \alpha_k \Delta w_k / \beta, \mu)^t \Delta w_k.$$

This inequality yields

$$\begin{aligned} &\nabla F(w_k + \theta_k \alpha_k \Delta w_k / \beta, \mu)^t \Delta w_k - \nabla F(w_k, \mu)^t \Delta w_k \\ &\quad > (\varepsilon_0 - 1) \nabla F(w_k, \mu)^t \Delta w_k > 0. \end{aligned} \tag{34}$$

Thus by the property $l_k \to \infty$, we have $\alpha_k \to 0$ and thus $\|\theta_k \alpha_k \Delta w_k/\beta\| \to 0, k \in K$, because $\|\Delta w_k\|$ is uniformly bounded above. Thus the left-hand side of (34) and therefore $\nabla F(w_k,\mu)^t \Delta w_k$ converge to zero when $k \to \infty, k \in K$. This contradicts assumption (31). Therefore we have proved (30).

It follows from (24) and (28) that

$$\nabla F(w_k,\mu)^t \Delta w_k \le -\frac{\|\Delta x_k\|^2}{M} - \rho \frac{\|g(x_k)+\mu y_k\|^2 + \|X_k' z_k - \mu e\|^2}{(x_k')^t z_k/p + \|g(x_k)+\mu y_k\|^2 + \|X_k' z_k - \mu e\|^2} < 0.$$

Since the compactness of the sequence $\{w_k\}$ guarantees that the denominator of the above equation does not approach infinity, (30) implies that

$$\Delta x_k \to 0, \quad g(x_k)+\mu y_k \to 0, \quad X_k' z_k - \mu e \to 0. \tag{35}$$

We should note that the existence of an accumulation point of the sequence $\{w_k\}$ is ensured by Assumption (G2). Let an arbitrary accumulation point of the sequence $\{w_k\}$ be $\hat{w} = (\hat{x},\hat{y},\hat{z})^t \in S \times \mathbf{R}^m \times \mathbf{R}^p_+$. Then from (35), we have

$$\hat{y} = -\frac{g(\hat{x})}{\mu} \quad \text{and} \quad \hat{z} = \mu(\hat{X}')^{-1} e,$$

where $\hat{X}' = \operatorname{diag}(\hat{x}_1', \ldots, \hat{x}_p')$. Because $\Delta x_k \to 0$ implies $\nabla F_0(\hat{x},\mu) = 0$ from (16) and assumption (G3), we have $r(\hat{w},\mu) = 0$ from Lemma 2.4(i). □

3. Q-superlinear convergence. In this section, we discuss under which condition Algorithm IP can possess the superlinear convergence property. For this purpose, we consider the following local algorithm, which is called Algorithm IPlocal. By appropriately controlling the parameters μ_k ($\mu_k \downarrow 0$) and γ_k ($\gamma_k \uparrow 1$) at each step near a KKT point, we can show that the unit Newton-like step from an approximate SBKKT point yields a next approximate SBKKT point that corresponds to the new updated barrier parameter, and that the sequence $\{w_k\}$ generated by Algorithm IPlocal converges Q-superlinearly to the KKT point.

ALGORITHM IPLOCAL.

Step 0. (Initialize) Set $w_0 \in S \times \mathbf{R}^m \times \mathbf{R}^p_+$ and $\varepsilon > 0$. Let $k = 0$.

Step 1. (Termination) If $\|r_0(w_k)\| \le \varepsilon$, then stop.

Step 2. (Update the parameters) Choose the parameters $\mu_k > 0$ and $0 < \gamma_k < 1$.

Step 3. (Compute direction) Calculate the direction Δw_k by the linear system of equations

$$J_k \Delta w_k = -r(w_k,\mu_k), \tag{36}$$

where the matrix J_k is given by

$$J_k = \begin{pmatrix} G_k & -A(x_k)^t & -E^t \\ A(x_k) & \mu_k I & 0 \\ Z_k E & 0 & X_k' \end{pmatrix}. \tag{37}$$

Step 4. (Step size) Set

$$\alpha_{k\max} = \min\left\{ \min_i \left\{ -\frac{(x_k')_i}{(\Delta x_k')_i} \,\middle|\, (\Delta x_k')_i < 0 \right\}, \min_i \left\{ -\frac{(z_k)_i}{(\Delta z_k)_i} \,\middle|\, (\Delta z_k)_i < 0 \right\} \right\},$$
$$\alpha_k = \min\{\gamma_k \alpha_{k\max}, 1\}.$$

Step 5. (Update variables) Set

$$w_{k+1} = w_k + \alpha_k \Delta w_k.$$

Step 6. Set $k := k + 1$ and go to Step 1.

Denote the Jacobian matrix of $r(w, \mu)$ by

$$\nabla r(w, \mu) = \begin{pmatrix} \nabla_x^2 L(w) & -A(x)^t & -E^t \\ A(x) & \mu I & 0 \\ ZE & 0 & X' \end{pmatrix}.$$

Let $w^* = (x^*, y^*, z^*)^t$ be a KKT point of (1). In the following, we assume that k is sufficiently large and μ_k is sufficiently close to 0. In order to prove superlinear convergence, we need Assumption L.

Assumption L.

(L1) The sequence $\{w_k\}$ converges to w^*.

(L2) The second derivatives of the functions f and g are Lipschitz continuous at x^*.

(L3) The second-order sufficient conditions for optimality and strict complementarity hold at w^*. Moreover, the active constraint gradients are linearly independent.

(L4) μ_k and γ_k are updated by the rules

$$\mu_k = \xi_k \|r_0(w_k)\|^{1+\tau_1} \quad \text{and} \quad 1 - \gamma_k = \sigma \xi_k \|r_0(w_k)\|^{\tau_2},$$

where τ_1, τ_2, and σ are positive constants such that $\min(1, \tau_2) > \tau_1$ and $0 < \sigma < 1$, and ξ_k is a positive number that satisfies $\frac{1}{M'} \le \xi_k \le M'$ for a positive constant M'.

(L5) The matrix G_k satisfies, for k sufficiently large,

$$\|G_k - \nabla_x^2 L(w^*)\| < \delta$$

for a positive constant $\delta > 0$ such that $\|\nabla r_0(w^*)^{-1}\|\delta < 1$, and

$$\|(G_k - \nabla_x^2 L(w_k))\Delta x_k\| = \mathrm{O}(\|\Delta w_k\|^{1+\tau_3}) \tag{38}$$

for some positive constant τ_3 such that $\tau_3 > \tau_1$.

First we note that the positive definiteness of the matrix $G_k + E^t(X_k')^{-1}Z_k E + \frac{1}{\mu}A(x_k)^t A(x_k)$ is not assumed. We should note that by (L3), the Jacobian matrix $\nabla r_0(w^*)$ is nonsingular. Then by (L2), (L4), and (L5), we have

$$\begin{aligned} \|J_k - \nabla r_0(w^*)\| &\le \|\nabla r_0(w_k) - \nabla r_0(w^*)\| + \|\nabla_x^2 L(w_k) - \nabla_x^2 L(w^*)\| \\ &\quad + \|G_k - \nabla_x^2 L(w^*)\| + \mu_k \\ &\le \|\nabla r_0(w_k) - \nabla r_0(w^*)\| + \|\nabla_x^2 L(w_k) - \nabla_x^2 L(w^*)\| \\ &\quad + \delta + M' \|r_0(w_k)\|^{1+\tau_1}. \end{aligned}$$

Thus there is some positive constant ζ_1 such that

$$\|\nabla r_0(w^*)^{-1}\| \|J_k - \nabla r_0(w^*)\| \le \zeta_1 < 1,$$

the matrix J_k in (37) is nonsingular, and we have

$$\|J_k^{-1}\| \le \zeta_2$$

for a positive constant ζ_2 by the Banach perturbation lemma. Thus the linear system of equations (36) has a unique solution. We also note that condition (38) is stronger than the condition

$$\lim_{k\to\infty} \frac{\|(G_k - \nabla_x^2 L(w_k))\Delta x_k\|}{\|\Delta w_k\|} = 0,$$

which was discussed by Martinez, Parada, and Tapia [14] and Yabe and Yamashita [17] for superlinear convergence. (Note that this corresponds to the Dennis–Moré condition for unconstrained optimization.)

Now we give the following theorem, which is very important for proving the superlinear convergence property of Algorithm IPlocal.

THEOREM 3.1. *Suppose that Assumption* L *holds. Let M_c be a constant such that $0 < M_c < \sqrt{p}$.*

(i) *If a point $\hat{w} = (\hat{x}, \hat{y}, \hat{z})^t \in S \times \mathbf{R}^m \times \mathbf{R}^p_+$ satisfies $\|r(\hat{w}, \mu_k)\| \le M_c\mu_k$, then*

$$\nu_1\|r_0(w_k)\|^{1+\tau_1} \le \|r_0(\hat{w})\| \le \nu_2\|r_0(w_k)\|^{1+\tau_1}$$

for positive constants ν_1 and ν_2.

(ii) *If $\|r(w_k, \mu_{k-1})\| \le M_c\mu_{k-1}$, then $\alpha_k = 1$.*

(iii) *The following holds:*

$$\|r(w_k + \Delta w_k, \mu_k)\| \le M_c\mu_k. \tag{39}$$

Proof. (i) Since $\|r(\hat{w}, \mu_k)\| \le M_c\mu_k$, we have

$$\|r_0(\hat{w})\| = \left\| r(\hat{w}, \mu_k) + \mu_k \begin{pmatrix} 0 \\ -\hat{y} \\ e \end{pmatrix} \right\| = \mathrm{O}(\mu_k) = \mathrm{O}(\|r_0(w_k)\|^{1+\tau_1}).$$

The last equality follows from Assumption (L4). Furthermore we obtain

$$\begin{aligned} \|r_0(\hat{w})\| &= \left\| r(\hat{w}, \mu_k) + \mu_k \begin{pmatrix} 0 \\ -\hat{y} \\ e \end{pmatrix} \right\| \ge \mu_k \left\| \begin{pmatrix} 0 \\ -\hat{y} \\ e \end{pmatrix} \right\| - \|r(\hat{w}, \mu_k)\| \\ &= \mu_k\sqrt{\|\hat{y}\|^2 + \|e\|^2} - \|r(\hat{w}, \mu_k)\| \ge (\sqrt{p} - M_c)\mu_k \\ &\ge \frac{\sqrt{p} - M_c}{M'}\|r_0(w_k)\|^{1+\tau_1}. \end{aligned}$$

(ii) We will show that

$$\gamma_k \min_i \left\{ -\frac{(x'_k)_i}{(\Delta x'_k)_i} \,\middle|\, (\Delta x'_k)_i < 0 \right\} \ge 1. \tag{40}$$

Assumption (L4) implies $\gamma_k \to 1$. For i such that $(Ex^*)_i > 0$, it follows from $(\Delta x'_k)_i \to 0$ and $\gamma_k \to 1$ that

$$-\gamma_k \frac{(x'_k)_i}{(\Delta x'_k)_i} > 1 \quad \text{for} \quad (\Delta x'_k)_i < 0.$$

Now we consider an index i such that $(Ex^*)_i = 0$. In this case we note that $(z^*)_i > 0$ by Assumption (L3), and thus $(z_k)_i > \frac{1}{2}(z^*)_i$. By (36), we have

$$(x'_k)_i + (\Delta x'_k)_i = \frac{\mu_k}{(z_k)_i} - \frac{(x'_k)_i(\Delta z_k)_i}{(z_k)_i}. \tag{41}$$

Since $\|r(w_k, \mu_{k-1})\| \leq M_c \mu_{k-1}$, we have

$$\mu_k \geq \frac{1}{M'} \|r_0(w_k)\|^{1+\tau_1} \geq \frac{\nu_1^{1+\tau_1}}{M'} \|r_0(w_{k-1})\|^{(1+\tau_1)^2} \tag{42}$$

by result (i), and

$$|(x'_k)_i (z_k)_i - \mu_{k-1}| \leq M_c \mu_{k-1}.$$

The latter yields

$$(x'_k)_i \leq \frac{(1+M_c)\mu_{k-1}}{(z_k)_i} = \frac{1+M_c}{(z_k)_i} \xi_{k-1} \|r_0(w_{k-1})\|^{1+\tau_1}.$$

Since the uniform boundedness of J_k^{-1} and the result of (i) imply

$$|(\Delta z_k)_i| \leq \|\Delta w_k\| = \mathrm{O}(\|r(w_k, \mu_k)\|) = \mathrm{O}(\|r_0(w_k)\|) = \mathrm{O}(\|r_0(w_{k-1})\|^{1+\tau_1}),$$

we have

$$(x'_k)_i |(\Delta z_k)_i| = \mathrm{O}\left(\|r_0(w_{k-1})\|^{2(1+\tau_1)}\right). \tag{43}$$

It follows from (41) and (43) that

$$(x'_k)_i + (\Delta x'_k)_i > \frac{\mu_k}{(z_k)_i} - \frac{\psi}{(z_k)_i} \|r_0(w_{k-1})\|^{2(1+\tau_1)}$$

for some positive constant ψ. Since Assumption (L4) implies $(1+\tau_1)^2 < 2(1+\tau_1)$, (42) implies that the first term of the right-hand side is dominant in the above equation. Thus for any constant $\hat{\sigma} \in (0,1)$, the following holds for k sufficiently large:

$$(x'_k)_i + (\Delta x'_k)_i > \hat{\sigma} \frac{\mu_k}{(z_k)_i}.$$

Letting $\hat{\sigma} = \sigma$ given by (L4), we have

$$(x'_k)_i + (\Delta x'_k)_i > \sigma \frac{\mu_k}{(z_k)_i}. \tag{44}$$

Since $(x'_k)_i (z_k)_i \leq \|r_0(w_k)\|$, Assumption (L4) guarantees

$$\begin{aligned} \frac{\mu_k}{(z_k)_i} = \frac{\xi_k \|r_0(w_k)\|^{1+\tau_1}}{(z_k)_i} &\geq \xi_k (x'_k)_i \|r_0(w_k)\|^{\tau_1} \\ &\geq \xi_k (x'_k)_i \|r_0(w_k)\|^{\tau_2} = \frac{1}{\sigma} (x'_k)_i (1-\gamma_k); \end{aligned}$$

then we have

$$\sigma \frac{\mu_k}{(z_k)_i} \geq (x'_k)_i (1-\gamma_k). \tag{45}$$

Thus by (44) and (45) we obtain

$$(x'_k)_i + (\Delta x'_k)_i > (1-\gamma_k)(x'_k)_i,$$

which implies

$$\gamma_k \left(-\frac{(x'_k)_i}{(\Delta x'_k)_i} \right) > 1 \quad \text{for} \quad (\Delta x'_k)_i < 0.$$

Hence (40) holds.

In the same way as above, we can prove that

$$\gamma_k \min_i \left\{ -\frac{(z_k)_i}{(\Delta z_k)_i} \middle| \, (\Delta z_k)_i < 0 \right\} \geq 1.$$

Therefore the result follows.

(iii) From Assumptions (L1), (L4), and (L5), we directly obtain

$$\begin{aligned}
\|r(w_k + \Delta w_k, \mu_k)\| &= \|r(w_k, \mu_k) + \nabla r(w_k, \mu_k)\Delta w_k + \mathrm{O}(\|\Delta w_k\|^2)\| \\
&\leq \|r(w_k, \mu_k) + J_k \Delta w_k\| + \mathrm{O}(\|\Delta w_k\|^2) \\
&\qquad + \|(J_k - \nabla r(w_k, \mu_k))\Delta w_k\| \\
&= \|(G_k - \nabla_x^2 L(w_k))\Delta x_k\| + \mathrm{O}(\|\Delta w_k\|^2) \\
&= \mathrm{O}(\|\Delta w_k\|^{\min(1+\tau_3,\ 2)}) \\
&= \mathrm{O}(\|r(w_k, \mu_k)\|^{\min(1+\tau_3,\ 2)}) \\
&= \mathrm{O}(\|r_0(w_k)\|^{\min(1+\tau_3,\ 2)}) \\
&= \mathrm{o}(\|r_0(w_k)\|^{1+\tau_1}) \\
&= \mathrm{o}(\mu_k) \\
&\leq M_c \mu_k.
\end{aligned}$$

This proves (39).

Therefore the proof of this theorem is complete. □

Theorem 3.1 shows that if w_k satisfies the approximate SBKKT condition for μ_{k-1}, then α_k is set to be unit in Step 4 of Algorithm IPlocal and $w_{k+1} = w_k + \Delta w_k$ also satisfies the approximate SBKKT condition for μ_k. Thus by result (i) of Theorem 3.1, we have

$$\nu_1 \|r_0(w_k)\|^{1+\tau_1} \leq \|r_0(w_{k+1})\| \leq \nu_2 \|r_0(w_k)\|^{1+\tau_1} \tag{46}$$

for positive constants ν_1 and ν_2. This implies that the R-superlinear convergence property of Algorithm IPlocal can be obtained if we choose an approximate SBKKT point for μ_0 as an initial point. Furthermore, it follows from Assumption L that there exist positive constants ν'_1 and ν'_2 such that

$$\nu'_1 \|w_k - w^*\| \leq \|r_0(w_k)\| \leq \nu'_2 \|w_k - w^*\|. \tag{47}$$

Combining Assumption (L4), (46), and (47), we obtain the following corollary.

COROLLARY 3.2. *Suppose that Assumption* L *holds. Then the sequences* $\{w_k\}$ *and* $\{\mu_k\}$ *generated by Algorithm* IPlocal *converge Q-superlinearly to the KKT point* w^* *and zero, respectively, and the relationships*

$$\nu''_1 \|w_k - w^*\|^{1+\tau_1} \leq \|w_{k+1} - w^*\| \leq \nu''_2 \|w_k - w^*\|^{1+\tau_1}$$

and

$$\nu''_1 \mu_k^{1+\tau_1} \leq \mu_{k+1} \leq \nu''_2 \mu_k^{1+\tau_1}$$

hold for positive constants ν''_1 *and* ν''_2.

4. Numerical experiment and concluding remarks. Before stating concluding remarks, we show a preliminary numerical experiment of the algorithm of this paper. A test code was written by Takahito Tanabe, and the following experiment was executed by him. The matrix G_k is updated by the quasi-Newton method using the BFGS formula. We use updating formula suggested by Powell [15] for the SQP method,

$$G_{k+1} = G_k - \frac{G_k s_k s_k^t G_k}{s_k^t G_k s_k} + \frac{u_k u_k^t}{s_k^t u_k},$$

where u_k is calculated by

$$\begin{aligned}
s_k &= x_{k+1} - x_k, \\
v_k &= \nabla_x L(x_{k+1}, y_{k+1}, z_{k+1}) - \nabla_x L(x_k, y_{k+1}, z_{k+1}), \\
u_k &= \theta_k v_k + (1-\theta_k) G_k s_k, \\
\theta_k &= \begin{cases} 1, & s_k^t v_k \geq 0.2 s_k^t G_k s_k, \\ \frac{0.8 s_k^t G_k s_k}{s_k^t G_k s_k - s_k^t v_k}, & s_k^t v_k < 0.2 s_k^t G_k s_k, \end{cases}
\end{aligned}$$

to satisfy $s_k^t u_k > 0$ for the hereditary positive definiteness of the update.

If the barrier parameter is large ($\mu_k > \epsilon_0$, where $\epsilon_0 = 10^4\sqrt{\epsilon_m}$, $\sqrt{\epsilon_m} \simeq 1.4 \times 10^{-8}$, ϵ_m is the machine epsilon), we use Algorithms IP and LS, and the barrier parameter is updated by

$$\mu_{k+1} = \max\left(\frac{\|r(w_{k+1}, \mu_k)\|}{M_\mu}, \frac{\mu_k}{M_{\mu 0}}\right)$$

when the condition

$$\|r(w_{k+1}, \mu_k)\| \leq M_c \mu_k \tag{48}$$

is satisfied. The following values of parameters are used in our experiment:

$$M_\mu = 4, \quad M_{\mu 0} = 10^6, \quad M_c = 3.$$

Suppose at w_{k+1}, (48) is satisfied and the barrier parameter is considered small ($\mu_k \leq \epsilon_0$); then we try to follow the steps described in Algorithm IPlocal hereafter. In Algorithm IPlocal we use the values $\tau_1 = 0.6, \tau_2 = 1$. If the next iterate computed by the method in IPlocal failed to satisfy (48), we would resort to Algorithm LS to satisfy the condition, and then proceed to try the next iteration by Algorithm IPlocal from the point obtained by Algorithm LS. If the iterate computed by the method in IPlocal satisfies (48), we proceed to the next iteration in Algorithm IPlocal. The barrier parameter is updated by the method described in Algorithm IPlocal in both cases.

The test problems are chosen from the Hock and Schittkowski test set [13]. Of the tested 39 problems, the code failed to solve 6 problems. The failed problems are marked with *i and *d in Table 1. The mark *i shows that the method stopped because of the iteration limits. The mark *d means that the method failed to produce a descent direction. Of the failed problems, two cases stopped with nondescent direction produced, and the remaining 4 problems stopped with iteration count over. In the table, "objective" means the final objective function value, "residual" means the final

TABLE 1

Problem	n	m	Objective	Residual	#iter	#eval	
HS41	4	1	1.92593	6.6e-07	20	24	
HS42	4	2	13.8579	1.6e-07	9	12	
HS43	4	3	-44	8.5e-07	11	14	
HS44	4	6	-4.73317	1.7e-07	23	28	
HS45	5	0	1.00001	9.2e-07	29	32	
HS46	5	2	3.1181e-07	3.6e-05	101	645	*i
HS47	5	3	4.21871e-09	5.1e-07	35	73	
HS48	5	2	6.02339e-14	1.1e-07	7	9	
HS49	5	2	3.6357e-07	1.1e-06	28	30	
HS50	5	3	4.25159e-09	6.7e-07	19	22	
HS51	5	3	1.34743e-12	9.5e-07	5	8	
HS52	5	3	5.32665	1.5e-07	15	18	
HS53	5	3	4.09302	6.7e-08	11	13	
HS54	6	1	-0.903488	1.2e-06	60	97	
HS55	6	6	6.33334	9.0e-07	14	17	
HS56	7	4	-8.06032e+41	1.2e+14	7	27	*d
HS57	2	1	0.0306463	1.2e-07	11	14	
HS59	2	3	-7.80279	7.1e-07	25	29	
HS60	3	1	0.032568	9.5e-07	8	10	
HS61	3	2	-143.646	1.3e-06	7	9	
HS62	3	1	-26272.5	1.3e-06	16	18	
HS63	3	2	961.715	5.8e-07	13	16	
HS64	3	1	6458.18	2.7e-04	101	619	*i
HS65	3	1	0.953548	5.8e-07	14	16	
HS66	3	2	0.518164	2.1e-07	22	34	
HS67	3	14	-910.017	1.0e-01	101	625	*i
HS68	4	2	2.62086e-05	6.8e-07	23	25	
HS69	4	2	0.00401042	5.5e-08	24	26	
HS70	4	1	0.00749846	1.1e-06	61	82	
HS71	4	2	17.0141	1.4e-06	16	20	
HS72	4	2	727.679	1.7e-07	69	144	
HS73	4	3	28.4203	2.0e-08	20	23	
HS74	4	5	6112.23	7.9e+04	101	297	*i
HS75	4	5	5431.57	7.0e+03	75	170	*d
HS76	4	3	-2.16186	1.1e-06	14	16	
HS77	5	2	0.241503	5.1e-07	14	17	
HS78	5	3	-2.9197	4.4e-07	11	13	
HS79	5	3	0.0787768	2.3e-07	10	12	
HS80	5	3	0.0539498	1.2e-06	11	13	

value of $\|r_0(w)\|$, "#iter" means number of total inner iterations needed, and "#eval" means number of function evaluations needed.

In summary, we have proposed a new differentiable primal-dual merit function in this paper. Theorem 2.8 ensures the global convergence of Algorithm LS to an SBKKT point for a fixed μ and therefore the global convergence of Algorithm IP to a KKT point of problem (1), while Corollary 3.2 implies the Q-superlinear convergence of Algorithm IPlocal to a KKT point of problem (1). However, this does not necessarily imply the superlinear convergence of Algorithm IP, because the Armijo line search criterion required in the inner iteration (Algorithm LS) may prevent it from choosing a unit step size even if the iterates are near a KKT point. This phenomenon is known as the Maratos effect. However, if we adopt a unit step size when the current point w_k (the initial point for the kth inner iteration) satisfies the approximate SBKKT

condition for sufficiently small μ_{k-1}, and $w_k + \Delta w_k$ (the first step for the kth inner iteration) satisfies the approximate SBKKT condition for μ_k, even when the merit function value does not satisfy the Armijo rule, then Theorems 2.8 and 3.1 assure that we can have the global and superlinear convergence of Algorithm IP by appropriately controlling the parameters μ_k and γ_k at the final stage of iterations. The numerical experiment above adopts this strategy.

We could devise an algorithm for avoiding the Maratos effect explicitly. For this purpose, we could use a nonmonotone strategy like the primal-dual interior point trust region method given by Yamashita, Yabe, and Tanabe [22], for example. Since we think it is another theme, we do not elaborate on a specific exposition of the algorithm in the present paper. Further research can be expected.

Acknowledgments. We are grateful to anonymous referees for their valuable suggestions, and Takahito Tanabe for implementing the algorithm.

REFERENCES

[1] I. Akrotirianakis and B. Rustem, *A Globally Convergent Interior Point Algorithm for General Nonlinear Programming Problems*, Tech. Report 97-14, Department of Computing, Imperial College of Science, Technology and Medicine, 1998 (revised March 1999).

[2] I. Akrotirianakis and B. Rustem, *A Primal-Dual Interior Point Algorithm with an Exact and Differentiable Merit Function for General Nonlinear Programming Problems*, Tech. Report 98-09, Department of Computing, Imperial College of Science, Technology and Medicine, 1998.

[3] M. Argaez and R. A. Tapia, *On the Global Convergence of a Modified Augmented Lagrangian Linesearch Interior Point Newton Method for Nonlinear Programming*, Tech. Report TR95-38, Department of Computational and Applied Mathematics, Rice University, Houston, TX, 1995 (revised February 1997).

[4] R. H. Byrd, J. C. Gilbert, and J. Nocedal, *A trust region method based on interior point techniques for nonlinear programming*, Math. Program., 89 (2000), pp. 149–185.

[5] R. H. Byrd, M. E. Hribar, and J. Nocedal, *An interior point algorithm for large-scale nonlinear programming*, SIAM J. Optim., 9 (1999), pp. 877–900.

[6] R. H. Byrd, G. Liu, and J. Nocedal, *On the local behaviour of an interior point method for nonlinear programming*, in Numerical Analysis 1997, D. F. Griffiths, D. J. Higham, and G. A. Watson, eds., Longman, Harlow, UK, 1998, pp. 37–56.

[7] A. R. Conn, N. I. M. Gould, D. Orban, and P. L. Toint, *A primal-dual trust-region algorithm for non-convex nonlinear programming*, Math. Program., 87 (2000), pp. 215–249.

[8] J. E. Dennis, Jr., and R. B. Schnabel, *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*, Prentice-Hall, Englewood Cliffs, NJ, 1983.

[9] J. E. Dennis, Jr., M. Heinkenschloss, and L. N. Vicente, *Trust-region interior-point SQP algorithms for a class of nonlinear programming problems*, SIAM J. Control Optim., 36 (1998), pp. 1750–1794.

[10] A. S. El-Bakry, R. A. Tapia, T. Tsuchiya, and Y. Zhang, *On the formulation and theory of the Newton interior-point method for nonlinear programming*, J. Optim. Theory Appl., 89 (1996), pp. 507–541.

[11] A. V. Fiacco and G. P. McCormick, *Nonlinear Programming: Sequential Unconstrained Minimization Techniques*, Classics Appl. Math. 4, SIAM, Philadelphia, 1990.

[12] A. Forsgren and P. E. Gill, *Primal-dual interior methods for nonconvex nonlinear programming*, SIAM J. Optim., 8 (1998), pp. 1132–1152.

[13] W. Hock and K. Schittkowski, *Test Examples for Nonlinear Programming Codes*, Lecture Notes in Econ. and Math. Systems 187, Springer-Verlag, Berlin, 1981.

[14] H. J. Martinez, Z. Parada, and R. A. Tapia, *On the characterization of Q-superlinear convergence of quasi-Newton interior-point methods for nonlinear programming*, Bol. Soc. Mat. Mexicana, 1 (1995), pp. 137–148.

[15] M. J. D. Powell, *A fast algorithm for nonlinearly constrained optimization calculations*, in Numerical Analysis, Dundee 1977, G. A. Watson, ed., Lecture Notes in Math. 630, Springer-Verlag, New York, 1978, pp. 144–157.

[16] R. J. Vanderbei and D. F. Shanno, *An interior-point algorithm for nonconvex nonlinear programming*, Comput. Optim. Appl., 13 (1999), pp. 231–252.

[17] H. Yabe and H. Yamashita, *Q-superlinear convergence of primal-dual interior point quasi-Newton methods for constrained optimization*, J. Oper. Res. Soc. Japan, 40 (1997), pp. 415–436.

[18] H. Yamashita, *A globally convergent primal-dual interior point method for constrained optimization*, Optim. Methods Softw., 10 (1998), pp. 443–469.

[19] H. Yamashita, *A primal-dual exact merit function for constrained optimization*, in Optimization—Modeling and Algorithms 8, Cooperative Research Report 84, Institute of Statistical Mathematics, 1996, pp. 119–127.

[20] H. Yamashita and T. Tanabe, *A primal-dual interior point trust region method for large scale constrained optimization*, in Optimization—Modeling and Algorithms 6, Cooperative Research Report 73, Institute of Statistical Mathematics, 1995, pp. 1–25.

[21] H. Yamashita and H. Yabe, *Superlinear and quadratic convergence of some primal-dual interior point methods for constrained optimization*, Math. Programming, 75 (1996), pp. 377–397.

[22] H. Yamashita, H. Yabe, and T. Tanabe, *A Globally and Superlinearly Convergent Primal-Dual Interior Point Trust Region Method for Large Scale Constrained Optimization*, Tech. Report, 1997 (revised May 2003).

SIAM J. OPTIM.
Vol. 14, No. 2, pp. 500–506

RELATING HOMOGENEOUS CONES AND POSITIVE DEFINITE CONES VIA T-ALGEBRAS*

CHEK BENG CHUA†

Abstract. T-algebras are nonassociative algebras defined by Vinberg in the early 1960s for the purpose of studying homogeneous cones. Vinberg defined a cone $K(\mathcal{A})$ for each T-algebra $\mathcal{A}$ and proved that every homogeneous cone is isomorphic to one such $K(\mathcal{A})$. We relate each T-algebra $\mathcal{A}$ with a space of linear operators in such a way that $K(\mathcal{A})$ is isomorphic to the cone of positive definite self-adjoint operators. Together with Vinberg's result, we conclude that every homogeneous cone is isomorphic to a "slice" of a cone of positive definite matrices.

Key words. homogeneous cones, T-algebras, positive definite cones

AMS subject classification. 90C25

DOI. S1052623402406765

1. Introduction. Due to the generality of interior-point methods, they have been successfully applied to a wide class of conic programming problems; one of the more prominent of these classes is semidefinite programming (SDP), whose underlying cone is the set of positive semidefinite symmetric matrices.

Positive semidefinite cones are examples of homogeneous cones. A full-dimensional cone K in $\mathbb{R}^n$ is homogeneous if the group of automorphisms of the cone acts transitively on it (i.e., for every $x, y \in K$, there exists a linear automorphism A of K such that $Ax = y$). Homogeneous cones were studied by Vinberg [4], who associated homogeneous cones with certain nonassociative algebras called T-algebras. Through T-algebras, Vinberg classified all homogeneous self-dual cones.

From the association of homogeneous cones with T-algebras, we show that homogeneous cones are "slices" of positive definite cones. More precisely, we show that for some $m \leq n$, there exists an injective linear map $M : \mathbb{R}^n \to \mathbb{S}^{m\times m}$ such that $M(K) = \mathbb{S}^{m\times m}_{++} \cap M(\mathbb{R}^n)$, where $\mathbb{S}^{m\times m}$ is the space of m-by-m symmetric matrices and $\mathbb{S}^{m\times m}_{++}$ is the cone of positive definite symmetric m-by-m matrices.[1]

(After the first version of this paper appeared, Faybusovich pointed out that the same conclusion follows from his work [2]. Indeed, by recognizing the cone $K(\mathcal{A})$ as a cone of "squares" in the context of [2], it follows from the construction in [2] that $K(\mathcal{A})$ is a "slice" of a positive definite cone. However, this construction requires

*Received by the editors April 30, 2002; accepted for publication (in revised form) May 12, 2003; published electronically November 6, 2003. This research was performed as part of the author's Ph.D. study at Cornell University.

http://www.siam.org/journals/siopt/14-2/40676.html

†Department of Combinatorics and Optimization, University of Waterloo, Waterloo, ON, Canada N2L 3G1 (cbchua@math.uwaterloo.ca).

[1]The converse is not true. For example, consider the cone $K = \{(x_1, x_2, x_3)^T \in \mathbb{R}^3 : x_1 > 0, x_3 > \sqrt{x_1^2 + x_2^2}\}$. This is a "slice" of the positive definite cone $\mathbb{S}^{3\times 3}_{++}$, as can be seen by taking $M : \mathbb{R}^3 \to \mathbb{S}^{3\times 3}$ to be the injective linear map

$$M : (x_1, x_2, x_3)^T \mapsto \begin{bmatrix} x_1 & 0 & 0 \\ 0 & x_3 - x_1 & x_2 \\ 0 & x_2 & x_3 + x_1 \end{bmatrix}.$$

From Vinberg's classification of homogeneous cones (see [4]), a three-dimensional homogeneous cone is linearly isomorphic to either the positive orthant or the second-order cone. Therefore, K is not homogeneous.

an order n positive definite cone, i.e., a cone of symmetric positive definite n-by-n matrices. In this paper, our construction may produce a cone of a lower order with the proper choice of some index set I.)

One consequence of this result is that we can model conic programming problems over homogeneous cones as SDP problems, which are studied much more thoroughly than homogeneous cone programming (see, e.g., [1]). However, from a practical point of view, modeling a conic programming problem over a homogeneous cone as an SDP may not be the best thing to do. For example, to optimize over an n-dimensional second-order cone (i.e., Lorentz cone), we can use the standard logarithmic barrier, which has a complexity value of 2. Modeling it as an SDP would embed the second-order cone into the cone of positive definite $(n-1)$-by-$(n-1)$ matrices. Thus we would be using a barrier of complexity value $n-1$ instead of 2 if we solve a second-order programming as an SDP. In fact, Güler and Tuncel [3] showed that the best barrier parameter for a homogeneous cone is the same as the rank r of the cone, which is an algebraic property of the cone. In the same paper, a barrier of complexity value r is given. However, the applicability of this barrier in implementations of interior-point methods for optimization over homogeneous cones depends on the efficient computability of its gradient and Hessian, which is still not addressed.

This paper is organized as follows. We begin by describing T-algebras as defined in [4]. We then state the main result in [4] that associates homogeneous cones with T-algebras. In section 3, we associate T-algebras with spaces of linear operators; in particular, we define, for each T-algebra, an injective linear map L that maps elements in the T-algebra to linear operators. The special structure of T-algebras allows us to derive important properties of L, which is used in the proof of our main theorem. In the last section, we prove the main theorem: every homogeneous cone is a "slice" of some cone of positive definite linear operators.

2. *T*-algebras and homogeneous cones. This section is devoted to the description of T-algebras and the association of homogeneous cones with T-algebras.

A *homogeneous cone* K is a full-dimensional convex pointed cone in a finite-dimensional space such that the group of linear automorphisms of K acts transitively on it (i.e., for every $x, y \in K$, there exists a linear map A such that $Ax = y$ and $AK = K$).

A *matrix algebra of rank* r is an algebra $\mathcal{A} = \bigoplus_{i,j=1}^{r} \mathcal{A}_{ij}$ such that

$$\mathcal{A}_{ij}\mathcal{A}_{\ell k} \subset \begin{cases} \mathcal{A}_{ik} & \text{if } j = \ell, \\ \{0\} & \text{if } j \neq \ell. \end{cases}$$

Denote the dimension of $\mathcal{A}_{ij}$ by n_{ij}.

If we represent each $a \in \mathcal{A}$ by the generalized matrix $(a_{ij})_{i,j=1}^{r}$, where a_{ij} denotes the projection of a onto $\mathcal{A}_{ij}$, then the representation of ab is given by the matrix product $(a_{ij})(b_{ij})$. For example, suppose $\mathcal{A}$ is a matrix algebra of rank 2 and $a = a_{11} + a_{12} + a_{21} + a_{22}$. It is easy to see that $(ab)_{ij} = \sum_{k=1}^{2} a_{ik}b_{kj}$, which corresponds to the usual matrix multiplication.

An *involution of a matrix algebra* $\mathcal{A}$ is an linear map * of $\mathcal{A}$ onto itself such that

1. $a^{**} = a$,
2. $(ab)^* = b^*a^*$, and
3. $\mathcal{A}_{ij}^* \subset \mathcal{A}_{ji}$.

In its matrix representation, an involution corresponds to taking the transpose, i.e., $(a^*)_{ij} = a_{ji}^*$. A consequence of the existence of an involution is that $n_{ij} = n_{ji}$.

EXAMPLE 2.1 (real matrices). *The algebra $\mathbb{R}^{r\times r}$ of real r-by-r matrices is a matrix algebra of rank r. In this matrix algebra, $\mathcal{A}_{ij}$ is the subspace of matrices that are zero outside the (i,j)th entry, and $n_{ij}=1$. The transposition of matrices is an involution for the matrix algebra.*

EXAMPLE 2.2 (real vectors). *When $n_{ij}=0$ for all $i\neq j$, we get the algebra of real r-vectors, where the multiplication of two vectors is given by their componentwise product. The involution is the identity map.*

Henceforth, $\mathcal{A}$ will be a matrix algebra with involution.

Let

$$\mathcal{T} := \sum_{i\leq j} \mathcal{A}_{ij}$$

be the subspace of $\mathcal{A}$ whose elements are represented by upper-triangular matrices, and let

$$\mathcal{H} := \{a\in\mathcal{A} : a = a^*\}$$

be the subspace of $\mathcal{A}$ whose elements are represented by "symmetric" matrices.

Suppose $\mathcal{A}_{ii}$ is isomorphic to the field $\mathbb{R}$ of real numbers for each i. We let $\rho_i : \mathcal{A}_{ii}\to\mathbb{R}$ denote the isomorphism and e_i denote the unit element of $\mathcal{A}_{ii}$. Since the function $f:\mathbb{R}\to\mathbb{R} : x\mapsto \rho_i(\rho_i^{-1}(x)^*)$ is a linear automorphism on $\mathbb{R}$, it is the identity map. Hence $a_{ii}^* = a_{ii}$ for all $a_{ii}\in\mathcal{A}$. The *trace* of an element $a\in\mathcal{A}$ is defined as

$$\operatorname{tr} a := \sum_{i=1}^{r} \rho_i(a_{ii}).$$

A *T-algebra* is a matrix algebra $\mathcal{A}$ of rank r with involution * that satisfies the following axioms:

(I) $\mathcal{A}_{ii}$ is isomorphic to $\mathbb{R}$.
(II) $e_i a_{ij} = a_{ij}e_j = a_{ij}$ for all $a_{ij}\in\mathcal{A}_{ij}$.
(III) $\operatorname{tr} ab = \operatorname{tr} ba$.
(IV) $\operatorname{tr} a(bc) = \operatorname{tr}(ab)c$.
(V) $\operatorname{tr} a^*a > 0$ unless $a=0$.
(VI) $t(uw) = (tu)w$ for all $t,u,w\in\mathcal{T}$.
(VII) $t(uu^*) = (tu)u^*$ for all $t,u\in\mathcal{T}$.

In a T-algebra $\mathcal{A}$, the element with $a_{ii}=e_i$ and $a_{ij}=0$, $i\neq j$, is the unit element e of $\mathcal{A}$.

From axiom (V), we see that $\langle a,b\rangle := \operatorname{tr} a^*b$ is an inner product on $\mathcal{A}$. Under this inner product, $\mathcal{A}_{ij}$ is orthogonal to $\mathcal{A}_{k\ell}$ unless $(i,j)=(k,\ell)$.

Let

$$\mathcal{I} := \{t\in\mathcal{T} : \rho_i(t_{ii})>0 \text{ for } 1\leq i\leq r\}$$

be the subgroup of upper-triangular matrices whose diagonal elements are positive, and let

$$K(\mathcal{A}) := \{tt^* : t\in\mathcal{I}\}\subset\mathcal{H}.$$

Vinberg [4] proved the following important result that relates homogeneous cones with the cones $K(\mathcal{A})$.

THEOREM 2.3. *A cone K is homogeneous if and only if there exists a T-algebra $\mathcal{A}$ such that K is isomorphic to $K(\mathcal{A})$.*

3. T-algebras and linear operators. Let

$$\hat{\mathcal{V}} := \sum_{i=1}^{r} \mathcal{A}_{i1}$$

be the subspace of "vectors." Each $a \in \mathcal{A}$ defines a linear operator $\hat{L}_a : \hat{\mathcal{V}} \to \hat{\mathcal{V}}$ by $v \mapsto av$. Since $\mathcal{A}$ is nonassociative in general, we cannot expect $\hat{L}_a\hat{L}_b = \hat{L}_{ab}$ to hold in general, where $\hat{L}_a\hat{L}_b$ is the composition of $\hat{L}_a$ and $\hat{L}_b$. Still, T-algebras have enough structure to allow us to prove the following useful proposition.

PROPOSITION 3.1. *Let $\hat{L} : \mathcal{A} \to L[\hat{\mathcal{V}}, \hat{\mathcal{V}}]$ be as defined above. For every $a \in \mathcal{A}$ and $t, u \in \mathcal{T}$,*

(i) $\hat{L}_{a^*} = \hat{L}_a^*$, *where $\hat{L}_a^*$ denotes the adjoint of $\hat{L}_a$ under $\langle\cdot,\cdot\rangle$;*
(ii) $\hat{L}_t\hat{L}_u = \hat{L}_{tu}$; *and*
(iii) $\hat{L}_t\hat{L}_{t^*} = \hat{L}_{tt^*}$.

Furthermore, $\hat{L}_a$ is the zero map if and only if $a_{ji} = a_{ij} = 0$ for all i with $n_{i1} \neq 0$ and all $j \geq i$.

Proof. (i) For any $u, v \in \hat{\mathcal{V}}$, $\langle \hat{L}_a^* u, v\rangle = \langle u, \hat{L}_a v\rangle = \operatorname{tr} u^*(av) = \operatorname{tr}(u^*a)v = \operatorname{tr}(a^*u)^*v = \langle a^*u, v\rangle = \langle \hat{L}_{a^*}u, v\rangle$ by axiom (IV). It follows that $\hat{L}_a^* = \hat{L}_{a^*}$.

(ii) By axiom (VI), $\hat{L}_{u^*}\hat{L}_{t^*}v = \hat{L}_{u^*}(t^*v) = u^*(t^*v) = (u^*t^*)v = \hat{L}_{u^*t^*}v$. This implies that $\hat{L}_{u^*}\hat{L}_{t^*} = \hat{L}_{u^*t^*}$. Taking * on both sides, we get $\hat{L}_t\hat{L}_u = \hat{L}_{tu}$.

(iii) By axiom (VII), $\hat{L}_t\hat{L}_{t^*}v = \hat{L}_t(t^*v) = t(t^*v) = (tt^*)v = \hat{L}_{tt^*}v$. So, $\hat{L}_t\hat{L}_{t^*} = \hat{L}_{tt^*}$.

Suppose that $\hat{L}_a$ is the zero map and $n_{i1} \neq 0$. Then, for any $v_{i1} \in \mathcal{A}_{i1}$ with $v_{i1} \neq 0$, $a_{ji}v_{i1} = (\hat{L}_a v_{i1})_{j1} = 0$. So, for any $j \geq i$,

$$\begin{aligned}
0 =& \operatorname{tr}(a_{ji}v_{i1})(a_{ji}v_{i1})^* \\
=& \operatorname{tr} a_{ji}(v_{i1}(v_{i1}^* a_{ji}^*)) && \text{by axiom (IV)} \\
=& \operatorname{tr}((v_{i1}v_{i1}^*)a_{ji}^*)a_{ji} && \text{by axioms (III) and (VII)} \\
=& \operatorname{tr}(v_{i1}v_{i1}^*)(a_{ji}^*a_{ji}) && \text{by axiom (IV)} \\
=& \rho_i(v_{i1}v_{i1}^*)\rho_i(a_{ji}^*a_{ji}),
\end{aligned}$$

implying that $\rho_i(a_{ji}^*a_{ji}) = 0$ since $\rho_i(v_{i1}v_{i1}^*) \neq 0$ when $v_{i1} \neq 0$. Therefore, we conclude that $a_{ji} = 0$. Since $\hat{L}_{a^*} = \hat{L}_a^*$ is also the zero map, the same argument shows that $(a^*)_{ji} = 0$, from which we conclude that $a_{ij} = (a_{ij}^*)^* = ((a^*)_{ji})^* = 0^* = 0$.

Conversely, suppose that $a \in \mathcal{A}$ is such that $a_{ij} = a_{ji} = 0$ for all i with $n_{i1} \neq 0$ and all $j \geq i$. Let $v \in \hat{\mathcal{V}}$ be arbitrary. Consider $L_a v_{i1}$ for each $1 \leq i \leq r$. Clearly, $L_a v_{i1} = 0$ if $n_{i1} = 0$. If $n_{i1} \neq 0$, consider $(L_a v_{i1})_{j1}$ for each $1 \leq j \leq r$. If $n_{j1} = 0$, then $(L_a v_{i1})_{j1} \in \mathcal{A}_{j1} \implies (L_a v_{i1})_{j1} = 0$. Otherwise, we have either $i \leq j$ or $j \leq i$ (or $i = j$). In either case, $a_{ij} = a_{ji} = 0$ by assumption. Hence, $(L_a v_{i1})_{j1} = a_{ji}v_{i1} = 0$. Consequently, $L_a v_{i1} = 0$ when $n_{i1} \neq 0$. Thus, $L_a v = \sum_{i=1}^r L_a v_{i1} = 0$ for any $v \in \hat{\mathcal{V}}$. ∎

For each i, $\mathcal{A}^{(i)} := \sum_{k,l=i}^r \mathcal{A}_{kl}$ is clearly a subalgebra of $\mathcal{A}$. In fact, it is a T-algebra with involution *. Thus, we can define the subspace of "vectors" $\mathcal{V}^{(i)}$ in $\mathcal{A}^{(i)}$, and the linear operator $L_a^{(i)} : \mathcal{V}^{(i)} \to \mathcal{V}^{(i)}$ by $v \mapsto av$ for each $a \in \mathcal{A}^{(i)}$. Note that $\mathcal{A}^{(1)} = \mathcal{A}$, $\mathcal{V}^{(1)} = \hat{\mathcal{V}}$, and $L^{(1)} = \hat{L}$. For each subset $I \subset \{1, \ldots, r\}$, let $\mathcal{V}^I$ denote the subspace $\sum_{i\in I} \mathcal{V}^{(i)} \subset \mathcal{T}^*$. Define the map $L_a^I : \mathcal{V}^I \to \mathcal{V}^I$ by $L_a^I v = \sum_{i\in I} L_{a^{(i)}}^{(i)} v^{(i)}$, where $a^{(i)}$ and $v^{(i)}$ denote projections of a and v onto $\mathcal{A}^{(i)}$ and $\mathcal{V}^{(i)}$, respectively. By

observing that $(L_a v)^{(i)} = L^{(i)}_{a^{(i)}} v^{(i)}$, we can easily see that the first three statements in the above proposition hold for the map $L^I : \mathcal{A} \to L[\mathcal{V}^I, \mathcal{V}^I]$.

Suppose that I is chosen to satisfy the following condition:

$(*)$ For all $1 \le j \le r$, there exists $i \in I$ such that $i \le j$ and $n_{ji} \neq 0$.

Clearly, the choice $I = \{1, \dots, r\}$ satisfies this condition. Whenever I satisfies $(*)$, we call the map L^I the *real matrix representation of $\mathcal{A}$ with respect to I*.

EXAMPLE 3.2 (real matrices (cont'd)). *When $\mathcal{A}$ is the algebra of real r-by-r matrices, the choice $I = \{1\}$ satisfies $(*)$ since $n_{ji} = 1 \neq 0$ for all $1 \le i, j \le r$. With this choice, $\mathcal{V}^I$ can be regarded as the space of real r-vectors, and L^I_a is the map represented by the matrix a.*

EXAMPLE 3.3 (real vectors (cont'd)). *When $\mathcal{A}$ is the algebra of real r-vectors, the only I satisfying $(*)$ is $I = \{1, \dots, r\}$ since $n_{ji} = 0$ for all $j \neq i$.*

Suppose that L^I_a is the zero map. Then, for each $i \in I$, $L^{(i)}_{a^{(i)}}$ is the zero map. Now, fix an arbitrary $1 \le j \le r$ and choose an $i \in I$, $i \le j$, for which $n_{ji} \neq 0$. By applying the above proposition to $L^{(i)}$, we conclude that $a_{kj} = a_{jk} = 0$ for all $k \ge j$. Since j is arbitrary, we have $a = 0$. Thus, L^I is injective when I satisfies $(*)$.

Conversely, suppose that for some $1 \le j \le r$, $n_{ji} = 0$ for all $i \in I$ such that $i \le j$. It follows that $L^I_{e_j} v = \sum_{i \in I} L^{(i)}_{e_j^{(i)}} v^{(i)} = \sum_{i \in I,\, i \le j} L^{(i)}_{e_j} v^{(i)} = \sum_{i \in I,\, i \le j} e_j v_{ji} = 0$ for any $v \in \mathcal{V}^I$. Hence, L^I is not injective when I violates $(*)$.

Thus, we have proven the following proposition.

PROPOSITION 3.4. *Let $L^I : \mathcal{A} \to L[\mathcal{V}^I, \mathcal{V}^I]$ be as defined above. For every $a \in \mathcal{A}$ and $t, u \in \mathcal{T}$,*

(i) $L^I_{a^*} = (L^I_a)^*$;
(ii) $L^I_t L^I_u = L^I_{tu}$ *(equivalently, $L^I|_{\mathcal{T}^*}$ is an isomorphism of algebras); and*
(iii) $L^I_t L^I_{t^*} = L^I_{tt^*}$.

Furthermore, L^I is injective if and only if I satisfies $(*)$.

Henceforth, we will fix an I that satisfies $(*)$. To simplify notation, we shall drop the superscript I from L^I and $\mathcal{V}^I$.

We end this section with two remarks on the map L.

REMARK 3.5. *By observing that each $t \in \mathcal{I}$ has a right inverse $u \in \mathcal{I}$ such that $tu = e$, we see that L_t is invertible for any $t \in \mathcal{I}$. Since $L|_{\mathcal{T}^*}$ is an isomorphism of algebras, t is also invertible with inverse t^{-1} satisfying $L_{t^{-1}} = L_t^{-1}$. It follows from $L_{(t^*)^{-1}} = L^{-1}_{t^*} = (L^*_t)^{-1} = (L_t^{-1})^* = (L_{t^{-1}})^* = L_{(t^{-1})^*}$ that $(t^*)^{-1} = (t^{-1})^*$.*

REMARK 3.6. *It is easy to see that $t = e$ is the only $t \in \mathcal{I}$ that satisfies $tt^* = e$. Suppose $tt^* = uu^*$ for some $t, u \in \mathcal{I}$. Then $L_{(t^{-1}u)(t^{-1}u)^*} = L_{t^{-1}u} L^*_{t^{-1}u} = L_t^{-1} L_u L_u^* (L_t^{-1})^* = L_t^{-1} L_{uu^*} (L_t^{-1})^* = L_t^{-1} L_{tt^*} (L_t^{-1})^* = L_t^{-1} L_t L_t^* (L_t^{-1})^* = L_{t^{-1}t} L^*_{t^{-1}t}$ is the identity map, implying that $t = u$. Hence, the relation $a = tt^*$ sets up a one-to-one correspondence between each $a \in K(\mathcal{A})$ and $t \in \mathcal{I}$.*

4. Homogeneous cones and cones of positive definite operators. Before we proceed to the main theorem, let us apply the result of the previous section to produce an easy proof of the fact that $K(\mathcal{A})$ is homogeneous.

For each $t \in \mathcal{I}$, define the map $\tau(t) : uu^* \mapsto (tu)(tu)^*$. By Remark 3.6, $\tau(t)$ is well defined. $\tau(t)$ is clearly a map of $K(\mathcal{A})$ into itself. In fact, by observing that every $u \in \mathcal{I}$ has an inverse in $\mathcal{I}$, we see that $\tau(t)$ maps $K(\mathcal{A})$ onto itself and $\{\tau(t) : t \in \mathcal{I}\}$ acts transitively on $K(\mathcal{A})$. From Proposition 3.4, $L_{(tu)(tu)^*} = L_t L_u L_u^* L_t^* = L_t L_{uu^*} L_t^*$, which implies that $\tau(t)$ acts linearly on $K(\mathcal{A})$. By extending $\tau(t)$ to a linear automorphism of the subspace $\mathcal{H}$, we can prove the "if" part of Theorem 2.3.

THEOREM 4.1. *For each $t \in \mathcal{I}$, let $\bar{\tau}(t)$ be the extension of $\tau(t)$ to the subspace $\mathcal{H}$. The subgroup of automorphisms $\{\bar{\tau}(t) : t \in \mathcal{I}\}$ of $\mathcal{H}$ is an invariant and transitive subgroup for the cone $K(\mathcal{A})$. Consequently, $K(\mathcal{A})$ is homogeneous.*

Finally, we give the main theorem.

THEOREM 4.2. *For each $a \in \mathcal{A}$, $a \in K(\mathcal{A})$ if and only if L_a is positive definite and self-adjoint. Consequently, L embeds $K(\mathcal{A})$ into some cone of positive definite self-adjoint linear operators.*

Proof. For the "only if" part, suppose that $a = tt^* \in K(\mathcal{A}) \subset \mathcal{H}$ for some $t \in \mathcal{I}$. Then, by Proposition 3.4, $L_a^* = L_{a^*} = L_a$ and $\langle v, L_a v\rangle = \langle v, L_{tt^*} v\rangle = \langle v, L_t L_t^* v\rangle = \langle L_t^* v, L_t^* v\rangle > 0$ for all $v \in \mathcal{V}$, $v \neq 0$, since L_t is nonsingular, and so $L_t^* v \neq 0$.

For the "if" part, we shall proceed by induction on the rank of $\mathcal{A}$.[2] If $\mathcal{A}$ has rank 1, then $\mathcal{A}$ is isomorphic to the algebra of the reals, and every positive definite a can be written as $(\sqrt{\rho_1(a_1)}e_1)(\sqrt{\rho_1(a_1)}e_1)^*$ with $\rho_1(a_1) > 0$. Suppose that $\mathcal{A}$ has rank $r > 1$, and that the "if" part is true for all T-algebras of rank less than r. Suppose L_a is positive definite and self-adjoint. Let $\bar{\mathcal{A}} := \sum_{i,j=1}^{r-1} \mathcal{A}_{ij}$ be a rank $r-1$ T-algebra. Let $\bar{a} = \sum_{i,j=1}^{r-1} a_{ij} \in \bar{\mathcal{A}}$ and $a_r = \sum_{i=1}^{r-1} a_{ir}$. Let $\bar{\mathcal{V}} := \sum_{i\in I} \sum_{j=1}^{r-1} \mathcal{A}_{ji} \subset \mathcal{V}$. The orthogonal complement of $\bar{\mathcal{V}}$ in $\mathcal{V}$ is $\bar{\bar{\mathcal{V}}} := \sum_{i\in I} \mathcal{A}_{ri}$. For any $v \in \mathcal{V}$, there exist $\bar{v} \in \bar{\mathcal{V}}$ and $\bar{\bar{v}} \in \bar{\bar{\mathcal{V}}}$ such that $v = \bar{v} + \bar{\bar{v}}$; and

$$\begin{aligned} L_a v =& \sum_{i\in I} a^{(i)} v^{(i)} = \sum_{i\in I} (\bar{a}^{(i)} + a_r^{(i)}) v^{(i)} + \sum_{i\in I} ((a_r^{(i)})^* + a_{rr}) v^{(i)} \\ =& \sum_{i\in I} \bar{a}^{(i)} (\bar{v})^{(i)} + \sum_{i\in I} a_r^{(i)} (\bar{\bar{v}})^{(i)} + \sum_{i\in I} (a_r^{(i)})^* (\bar{v})^{(i)} + \sum_{i\in I} a_{rr} (\bar{\bar{v}})^{(i)} \\ =& L_{\bar{a}} \bar{v} + L_{a_r} \bar{\bar{v}} + L_{a_r}^* \bar{v} + \rho_r(a_{rr}) \bar{\bar{v}}, \end{aligned}$$

where $L_{\bar{a}} \bar{v} + L_{a_r} \bar{\bar{v}} \in \bar{\mathcal{V}}$ and $L_{a_r}^* \bar{v} + \rho_r(a_{rr}) \bar{\bar{v}} \in \bar{\bar{\mathcal{V}}}$. By $(*)$, both $\bar{\mathcal{V}}$ and $\bar{\bar{\mathcal{V}}}$ have positive dimensions. So, $\hat{L}_a$ is positive definite and self-adjoint only if $\rho_r(a_{rr}) > 0$ and $L_{\bar{a}} - \rho_r(a_{rr})^{-1} L_{a_r} L_{a_r}^*$ is positive definite over $\bar{\mathcal{V}}$. Therefore,

$$\begin{aligned} L_{\rho_r(a_{rr})\bar{a} - a_r a_r^*} &= \rho_r(a_{rr}) L_{\bar{a}} - L_{a_r a_r^*} \\ &= \rho_r(a_{rr}) L_{\bar{a}} - L_{a_r} L_{a_r}^* \qquad \text{(by Proposition 3.4(iii))} \\ &= \rho_r(a_{rr}) (L_{\bar{a}} - \rho_r(a_{rr})^{-1} L_{a_r} L_{a_r}^*) \end{aligned}$$

is positive definite over $\bar{\mathcal{V}}$. It is clearly self-adjoint. Let $\bar{I} = I \backslash \{r\}$, which satisfies $(*)$ for $\bar{\mathcal{A}}$. Let $\bar{L}$ be the real matrix representation of $\bar{\mathcal{A}}$ with respect to $\bar{I}$. It is easy to check that $L_a|_{\bar{\mathcal{V}}} = \bar{L}_a$ for all $a \in \bar{\mathcal{A}}$. So, by the induction hypothesis, $\rho_r(a_{rr})\bar{a} - a_r a_r^* = tt^*$ for some $t \in \mathcal{I} \cap \bar{\mathcal{A}}$. Therefore,

$$(t + a_r + a_{rr})(t + a_r + a_{rr})^* = tt^* + a_r a_r^* + a_{rr}^2 + a_r a_{rr} + a_{rr} a_r^* = \rho_r(a_{rr}) a,$$

which implies that $a = uu^*$ with $u = (t + a_r + a_{rr})/\sqrt{\rho_r(a_{rr})}$.

Finally, since L is injective, it is an embedding of $K(\mathcal{A})$ into the cone of positive definite self-adjoint linear operators over $\mathcal{T}^*$. □

COROLLARY 4.3. *If K is a homogeneous cone in $\mathbb{R}^n$, then there exist an $m \leq n$ and an injective linear map $M : \mathbb{R}^n \to \mathbb{S}^{m\times m}$ such that $M(K) = \mathbb{S}_{++}^{m\times m} \cap M(\mathbb{R}^n)$,*

[2] The proof of this part resembles a proof of the Cholesky factorization of symmetric positive definite matrices. Indeed, in the case where our T-algebra is the algebra of real r-by-r matrices and $L = L^I$ with $I = \{1\}$, the proof of this part would be a proof of Cholesky factorization.

where $\mathbb{S}^{m\times m}$ is the space of m-by-m symmetric matrices and $\mathbb{S}^{m\times m}_{++}$ is the cone of positive definite symmetric m-by-m matrices.

Proof. By Theorem 2.3, there exists a T-algebra $\mathcal{A}$ for which $K(\mathcal{A})$ is isomorphic to K. This isomorphism can be extended linearly to a linear bijection from $\mathbb{R}^n$ to $\mathcal{H}$, the space of "symmetric" matrices in $\mathcal{A}$. Pick an I that satisfies $(*)$ for $\mathcal{A}$. Then the real matrix representation of $\mathcal{A}$ with respect to I embeds $K(\mathcal{A})$ into the cone of positive definite self-adjoint linear operators on $\mathcal{V}^I$, which is of dimension $m := \sum_{i\in I}\sum_{j\geq i} n_{ji} \leq \sum_{i=1}^{r}\sum_{j\geq i} n_{ji} = n$. M is then obtained by composing the bijection from $\mathbb{R}^n$ to $\mathcal{H}$ with the real matrix representation of $\mathcal{A}$ with respect to I. ∎

Acknowledgments. I would like to express my gratitude toward my thesis advisor Professor James Renegar for his motivation and encouragement. I would also like to thank Professor Adrian Lewis for his helpful comments which led to a shorter proof of the main proposition. I thank the referees for their very useful comments. I dedicate this paper to my wife.

REFERENCES

[1] A. Ben-Tal and A. Nemirovski, *Lectures on Modern Convex Optimization: Analysis, Algorithms, and Engineering Applications*, MPS/SIAM Ser. Optim., SIAM, Philadelphia, PA, 2001.

[2] L. Faybusovich, *On Nesterov's approach to semi-definite programming*, Acta Appl. Math., 74 (2002), pp. 195–215.

[3] O. Güler and L. Tuncel, *Characterization of the barrier parameter of homogeneous convex cones*, Math. Programming, 81 (1998), pp. 55–76.

[4] È. B. Vinberg, *The theory of convex homogeneous cones*, Trans. Moscow Math. Soc., 12 (1965), pp. 340–403.

SIAM J. OPTIM.
Vol. 14, No. 2, pp. 507–516

CONDITION NUMBER THEOREMS IN OPTIMIZATION*

T. ZOLEZZI†

Abstract. Condition numbers for optimization problems in Banach spaces are considered. Lower and upper estimates of the (suitably defined) distance from ill-conditioning are obtained in terms of the reciprocal of condition numbers. An approach is presented based on the metric regularity of the inverse to the arg min map. These results are extensions of the Eckart–Young distance theorem to optimization problems.

Key words. conditioning, sensitivity, distance theorem, condition number theorem

AMS subject classifications. 49K40, 90C31

DOI. S1052623402411885

1. Introduction. Condition numbers of a given mathematical problem are measures of sensitivity of the solutions with respect to small changes in the problem's data. Therefore condition numbers depend on the class of (appropriately chosen) perturbations of the given problem, under which solvability persists. The simplest case arises when each perturbed problem has exactly one solution.

Consider the set D of the data of all perturbed problems and the set S of their solutions. Assume that both are subsets of normed linear spaces. Denote by $m(p) \in S$ the unique solution corresponding to the problem defined by the datum $p \in D$. Let $p^* \in D$ define the (unperturbed) problem whose condition number we want to consider. The (absolute) condition number of problem p^* can be defined by

$$\limsup_{p \to p^*} \frac{\|m(p) - m(p^*)\|}{\|p - p^*\|};$$

see, e.g., [2] for related definitions.

A possibly different (but again standard) definition of condition number is given by

$$\limsup_{(p,q) \to (p^*,p^*)} \frac{\|m(p) - m(q)\|}{\|p - q\|},$$

namely the Lipschitz modulus of the solution mapping m at p^*; see, e.g., [7, p. 43].

It is well known that for many problems of numerical analysis, the distance (appropriately measured) of a given problem to the set of ill-conditioned problems is proportional to (or bounded by a multiple of) the reciprocal of the condition number; see [2]. This has interesting implications on computational complexity issues; see [3]. For an abstract version see [1].

A suitable notion of condition number of a matrix leads to the well-known distance theorem of Eckart and Young [6], whose variational interpretation has been generalized in [16] to the setting of convex quadratic forms on Banach spaces. It turns out that, in such a framework, the relevant notion of condition number is any of the previous two, taken with respect to linear continuous additive perturbations.

*Received by the editors July 22, 2002; accepted for publication (in revised form) April 29, 2003; published electronically November 6, 2003. A preliminary version of this paper was presented at the workshop "Equilibrium Problems and Variational Models," Erice, Italy 2001. This work was partially supported by Università di Genova.

http://www.siam.org/journals/siopt/14-2/41188.html

†DIMA, Università di Genova, via Dodecaneso 35, 16146 Genova, Italy (zolezzi@dima.unige.it).

The role the condition number theorem plays in optimization is enforced by the results of [12] and [13], where conditioning measures are introduced related to the reciprocal of the distance to infeasibility in the framework of linear programming problems. In [9] an extension of the distance theorem is obtained for convex processes.

In this paper we consider optimization problems with objective functions defined on a ball in a Banach space. Motivated by the results of [16] we define condition numbers of such optimization problems with respect to linear continuous additive (tilt) perturbations. In the setting of differentiable functions with Lipschitz continuous gradient, we prove new results about upper and lower estimates of the distance to ill-conditioning, making use of the reciprocal of the condition number. The relevant notion of condition number turns out to be the Lipschitz modulus of the arg min map for the lower estimate and a related quantity for the upper estimate. The definition of (pseudo-)distance makes use of the Lipschitz constant of the gradients of the objective functions. In this way we obtain two generalizations of the condition number theorem in [16], as presented in sections 3 and 4 for free optimization problems. Both results can be considered as preliminary steps towards condition number theorems for more general optimization problems. Here we rely on first order optimality conditions in a crucial way. The two estimates we prove about the distance to ill-conditioning do not match, due to the technically different notions of (well-)conditioning we employ, even though the two settings are rather similar.

These partial results show that reasonable definitions of a condition number exist for optimization problems, with the standard meaning of sensitivity measure, leading to the geometrical link with the distance to ill-conditioning through versions of a condition number theorem. (In a sense, we reverse here the approach followed in [12].)

In section 5 we compare our results with those of [16] about optimization of convex quadratic forms. In section 6 we further enforce the role of the Lipschitz modulus of the arg min map by pointing out its connection with metric regularity properties of the inverse multifunction to arg min, as a direct consequence of known results. This approach does not require smoothness of the objective functions involved and applies to constrained problems as well. Making use of some results in [5] we obtain a (weak) version of the condition number theorem in a very general setting, whose (not necessarily variational) meaning is discussed at the end of the paper.

We remark that links between well-posedness and well-conditioning are obtained in [15]. Let us note that well-posedness and well-conditioning are often used interchangeably as equivalent terms in the literature (contrary to [15] and this paper).

2. Notations and problem setting. Throughout the paper, E is a real Banach space with dual E^*. The duality pairing is denoted by $\langle \cdot, \cdot \rangle$. $B(0,r)$ is the closed ball of center 0 and positive radius r. We fix $L > 0$. Given $p \in E^*$ and

$$f : B(0,L) \subset E \to R$$

we write

$$f_p(x) = f(x) - \langle p, x \rangle, \quad \|x\| \leq L, \tag{1}$$

and denote by

$$m(f,p) \tag{2}$$

its unique global minimizer on $B(0,L)$ whenever this makes sense. If f is fixed and no confusion arises we write simply $m(p)$ instead of $m(f,p)$ in (2). In general, $(B(0,L), g)$

denotes the optimization problem of globally minimizing the function g on $B(0,L)$, and

$$\arg\min\,(B(0,L),g)$$

denotes the (possibly empty) set of all global minimizers of g on $B(0,L)$.

Let $f: B(0,L)\subset E\to R$ be such that every f_p has exactly one global minimizer $m(p)$ on $B(0,L)$ for every p sufficiently small. Then three extended real numbers will be considered as follows:

$$c_1(f)=\limsup_{p\to 0}\frac{\|m(p)-m(0)\|}{\|p\|}; \tag{3}$$

$$c_2(f)=\limsup_{(p,q)\to(0,0)}\frac{\|m(p)-m(q)\|}{\|p-q\|};$$

$$c_3(f)=\limsup_{p\to 0}\frac{\langle p,m(p)-m(0)\rangle}{\|p\|^2}.$$

In (3), $c_1(f)$ is a measure of sensitivity of $m(p)$ as compared to $m(0)$, corresponding to small changes of the data p, having fixed the unperturbed value $p=0$. The same can be said of $c_2(f)$, except that now small perturbations q of $p=0$ are allowed.

We see that $c_1(f)$ and $c_2(f)$ represent definitions of condition numbers for the optimization problems $(B(0,L),f)$ with respect to the class of linear continuous additive perturbations defined by $f_p, p\in E^*$ sufficiently small. As already remarked, this is in agreement with general definitions of conditioning.

In (3), $c_3(f)$ represents a new measure of sensitivity, based on the pairing between the relative error $[m(p)-m(0)]/\|p\|$ and the corresponding data perturbation $p/\|p\|$. The introduction of $c_3(f)$ is motivated by the results of sections 4 and 5.

We denote by $C^{1,1}[B(0,L)]$ the set of all real-valued functions

$$f: B(0,L)\to R,$$

which are Fréchet differentiable at each interior point of $B(0,L)$, and whose gradient Df can be extended to the closed ball in such a way that it is Lipschitz continuous on $B(0,L)$. We endow $C^{1,1}[B(0,L)]$ with the pseudodistance

$$d(f,g)=\sup\left\{\frac{\|Df(x)-Dg(x)-Df(y)+Dg(y)\|}{\|x-y\|}:\right. \tag{4}$$
$$\left. x\neq y,\|x\|\le L,\|y\|\le L\right\};$$

$f,g\in C^{1,1}[B(0,L)]$.

We remark that for quadratic functions

$$f(x)=\frac{1}{2}\langle Ax,x\rangle,\quad g(x)=\frac{1}{2}\langle Bx,x\rangle$$

with A,B linear bounded symmetric operators between E and E^*, their pseudodistance (4) is

$$d(f,g)=\|A-B\|.$$

In this paper we consider two notions of well-conditioned optimization problems with objective function f. The first requires $c_2(f) < +\infty$, while the second requires $0 \leq c_3(f) < +\infty$. Under suitable regularity assumptions, partially depending on the notion employed, we shall estimate in each case the distance to ill-conditioning, making use of the appropriate condition numbers $c_2(f), c_3(f)$.

3. Estimate from below. In this section we prove a lower bound of the distance from ill-conditioning of a given well-conditioned optimization problem $(B(0,L), f)$ in terms of the reciprocal of $c_2(f)$ defined in (3).

To this aim we consider the set T_1 of all $f \in C^{1,1}[B(0,L)]$ such that

$$\text{arg min } (B(0,L), f_p) \neq \emptyset \text{ for every sufficiently small } p; \tag{5}$$

$$\text{arg min } (B(0,L), f) = \{0\}; \tag{6}$$

$$p \to \text{ arg min } (B(0,L), f_p) \text{ is upper semicontinuous at } p = 0. \tag{7}$$

How small p is in (5) may depend on f. Upper semicontinuity of the arg min multifunction in (7) is meant with respect to the strong topologies.

If $U \subset T_1$ and $f \in T_1$ we write

$$\text{dist } (f, U) = \text{ inf } \{d(f,g) : g \in U\},$$

where d is given by (4).

Our starting point is the following well-known result.

LEMMA 3.1. *Let P, Q be real Banach spaces and B a nonempty subset of P. Let $H : B \to Q$ be one-to-one, with a Lipschitz continuous inverse of Lipschitz constant K. Let $G : B \to Q$ be Lipschitz continuous with Lipschitz constant α. If $\alpha K < 1$, then*

$$F = H + G : B \to Q$$

is one-to-one, and its inverse function

$$F^{-1} : F(B) \to B$$

is Lipschitz (of constant $K/(1 - \alpha K)$).

The proof of Lemma 3.1 can be obtained from that of [14, Lemma 1.18, p. 14] by standard modifications.

Denote by W_1 the set of all functions $f : B(0,L) \to R$ such that

$$\text{arg min } (B(0,L), f_p) \text{ is a singleton} \tag{8}$$

for every sufficiently small $p \in E^*$ (depending on f);

$$c_2(f) < +\infty. \tag{9}$$

Then consider

$$I_1 = \{g \in T_1 : g \notin W_1\}.$$

The definition of W_1 (respectively, I_1) isolates those objective functions which give rise to a well-conditioned (respectively, ill-conditioned) optimization problem on $B(0,L)$ within T_1.

LEMMA 3.2. *If $f : B(0,L) \to R$ fulfills* (8)*, then*

$$c_3(f) \le c_1(f) \le c_2(f). \tag{10}$$

Moreover, if $f \in T_1$, the extended real number $c_2(f) > 0$.

Proof. Obviously, $c_1(f) \le c_2(f)$. If $p \neq 0$,

$$\frac{\langle p, m(p) - m(0)\rangle}{\|p\|^2} \le \frac{\|m(p) - m(0)\|}{\|p\|};$$

hence (10) follows. Let $f \in T_1$. If $c_2(f) = 0$, then $c_1(f) = 0$ as well by (10); hence $\|m(p)\|/\|p\| \to 0$ as $p \to 0$. Then $\|m(p)\| \to 0$; thus $p = Df[m(p)]$ for sufficiently small p, whence

$$\frac{\|Df[m(p)]\|}{\|m(p)\|} \to +\infty \text{ as } p \to 0. \tag{11}$$

Since $Df(0) = 0$ by (6), formula (11) contradicts the Lipschitz continuity of Df. ∎

LEMMA 3.3. *Let $f \in T_1 \cap W_1$ be such that*

$$Df \text{ is one-to-one near } 0. \tag{12}$$

Let $g \in T_1$ be such that $d(f,g) < 1/c_2(f)$. Then $g \in W_1$.

Proof. We remark that $1/c_2(f)$ makes sense by Lemma 3.2. We need to prove that

(13) for every sufficiently small p, g_p has a unique global minimizer on $B(0,L)$

and

$$c_2(g) < +\infty. \tag{14}$$

Proof of (13). Since $g \in T_1$, $\arg\min(B(0,L), g_p)$ is nonempty for all sufficiently small p. Given $p \in E^*$ let

$$u_1, u_2 \in \arg\min(B(0,L), g_p).$$

By (6) and (7), if p is sufficiently small, then u_1, u_2 are interior minimizers of g_p; hence

$$Dg(u_1) = p = Dg(u_2). \tag{15}$$

By (12) and $f \in T_1 \cap W_1$,

$$c_2(f) = \limsup_{(p,q)\to(0,0)} \frac{\|Df^{-1}(p) - Df^{-1}(q)\|}{\|p - q\|} < +\infty;$$

hence

$$K(\delta) = \sup\left\{\frac{\|Df^{-1}(p) - Df^{-1}(q)\|}{\|p - q\|} : p \neq q, \|p\| < \delta, \|q\| < \delta\right\} < +\infty,$$

provided $\delta > 0$ is sufficiently small. Since

$$K(\delta) \to c_2(f) \text{ as } \delta \to 0 \text{ and } d(f,g) < \frac{1}{c_2(f)},$$

it follows that $d(f,g) < 1/K(\delta)$ for all sufficiently small $\delta > 0$. Of course the Lipschitz constant of $Df - Dg$ on $B(0,\delta)$ is $\leq d(f,g)$. Moreover, Df is one-to-one on $B(0,\delta)$, with a Lipschitz continuous inverse whose Lipschitz constant is $K(\delta)$. □

By writing $Dg = Dg - Df + Df$ we are in position to apply Lemma 3.1; hence Dg is one-to-one on $B(0,\delta)$. It follows by (15) and (7) that $u_1 = u_2$, whence (13).

Again by Lemma 3.1, Dg^{-1} is Lipschitz continuous near 0, whence (14). □

The following statement, an obvious corollary of Lemma 3.3, is the main result of this section.

THEOREM 3.1. *Let $f \in T_1 \cap W_1$ with Df one-to-one near 0. Then*

$$\operatorname{dist}(f, I_1) \geq \frac{1}{c_2(f)}. \tag{16}$$

The main information contained in (16) can be interpreted in the following two ways. If $c_2(f)$ is small, then the optimization problem $(B(0,L), f)$ lies at large distance from ill-conditioning. Equivalently, if $(B(0,L), f)$ is close to ill-conditioning, then its condition number $c_2(f)$ must be proportionally large.

Remark 3.1. If $f \in T_1$, then $c_2(f) > 0$ by Lemma 3.2, so that in our setting the case $c_2(f) = 0$ never occurs, and the corresponding limit case of (16), namely $\operatorname{dist}(f, I_1) = +\infty$, is ruled out.

If Df fails to be Lipschitz, then $c_2(f) = 0$ is possible. For example, with $E = R$ and $f(x) = 3x^{4/3}/4$ we have $m(p) = p^3$; hence $c_2(f) = 0$.

However, the limit case of an estimate of the form (16) cannot hold for classes of problems with a non-Lipschitzian gradient of the objective function (whichever definition of distance is adopted), provided the set of ill-conditioned problems is nonempty. See [11, p. 30] for pointing out some shortcomings of the definition of condition number in the limit cases.

Remark 3.2. In the finite-dimensional case ($E = R^N$) the well-conditioned problems defined by objective functions $f \in W_1$ of class C^2, namely those with $c_2(f) < +\infty$, are characterized in [10, p. 288] by the strictly positive definite character of their Hessian matrix at 0. A similar result is true in the Hilbert space setting, as proved in [4, Theorem 4.1].

4. Estimates from above. In order to obtain an upper bound of the distance to ill-conditioning in terms of the reciprocal of a suitable condition number, we modify in this section the definition of well-conditioning as follows.

Denote by T_2 the set of all $f \in C^{1,1}[B(0,L)]$ such that (5) holds. We fix (any) $\delta > 0$ and consider the set W_2 of those $f \in T_2$ such that $c_3(f) \geq 0$ and

$$Df \text{ is one-to-one on } B(0,\delta). \tag{17}$$

Then put

$$I_2 = T_2 \setminus W_2.$$

In the following theorem we interpret $1/0 = +\infty$ and $1/+\infty = 0$.

THEOREM 4.1. *Let $f \in W_2$ fulfill (6) and (7). Moreover, let E be finite-dimensional, or let E be an infinite-dimensional reflexive Banach space with f weakly sequentially lower semicontinuous. Then*

$$\operatorname{dist}(f, I_2) \leq \frac{1}{c_3(f)}. \tag{18}$$

Proof. Only the case $c_3(f) > 0$ needs proof. By (7) and (17), the singleton

$$m(p) = Df^{-1}(p)$$

for all p sufficiently small. There exists a sequence $q_n \to 0$ in E^*, $q_n \neq 0$, such that

$$\frac{\langle q_n, Df^{-1}(q_n)\rangle}{\|q_n\|^2} \to c_3(f). \tag{19}$$

Let

$$w_n = Df^{-1}(q_n),$$

and consider the linear bounded symmetric operator $G_n : E \to E^*$ given by

$$G_n(x) = \frac{\langle q_n, x\rangle q_n}{\langle q_n, w_n\rangle}, \quad x \in E.$$

The above definition of G_n makes sense for every n sufficiently large. Now define for such n

$$f_n(x) = f(x) - \frac{1}{2}\langle G_n(x), x\rangle, \quad x \in B(0, L).$$

If E is finite-dimensional, the continuous function f_{np} attains its global minimum value of the compact set $B(0, L)$ for every p. If E is infinite-dimensional, the same conclusion holds by reflexivity, weak sequential lower semicontinuity of f, and weak sequential continuity of

$$x \to \langle G_n(x), x\rangle = \frac{\langle q_n, x\rangle^2}{\langle q_n, w_n\rangle}.$$

Hence $f_n \in T_2$ for every n. We have

$$Df_n(0) = Df(0) = 0;$$

moreover,

$$Df_n(w_n) = Df(w_n) - G_n(w_n) = q_n - q_n = 0. \tag{20}$$

If some $w_n = 0$, then

$$q_n = Df(w_n) = Df(0) = 0,$$

which is a contradiction. It follows that $w_n \neq 0$ for every n. Since $w_n \to 0$ by (7), equation (20) implies that Df_n fails to be one-to-one on $B(0, \delta)$ for all sufficiently large n. Thus $f_n \in I_2$; hence

$$\text{dist } (f, I_2) \leq d(f, f_n) = \|G_n\| = \frac{\|q_n\|^2}{\langle q_n, w_n\rangle}.$$

In the limit as $n \to +\infty$ we get (18) by (19). □

5. Convex quadratic forms. Considerably sharper results are available in the special case of objective functions

$$f(x) = \frac{1}{2}\langle Ax, x\rangle, \quad x \in E, \tag{21}$$

where A is any linear symmetric bounded nonnegative operator between E and E^*. Denote by T the set of all quadratic forms (21), and call a convex quadratic form f well-conditioned iff f_p has a unique global minimizer on E for all p and $c_1(f) < +\infty$. Denote by I the set of ill-conditioned forms in T. According to the main result of [16], if $f \in T$, f given by (21), is well-conditioned, then

$$\text{dist } (f, I) = \frac{1}{\|A^{-1}\|}.$$

This a generalization of the classical Eckart–Young theorem.

PROPOSITION 5.1. *Let f be given by* (21) *and be well-conditioned. Then*

$$c_1(f) = c_2(f) = c_3(f) = \|A^{-1}\|.$$

Proof. Well-conditioning implies that $A : E \to E^*$ is an isomorphism by [16, Proposition 3.1]. Now

$$c_3(f) = \lim_{\delta \to 0} \sup \left\{ \frac{\langle A^{-1}p, p\rangle}{\|p\|^2} : 0 < \|p\| < \delta \right\}$$

$$= \lim_{\delta \to 0} \sup \{\langle A^{-1}q, q\rangle : \|q\| = 1\} = \|A^{-1}\|.$$

Moreover,

$$c_2(f) = \lim_{\delta \to 0} \sup \left\{ \frac{\|A^{-1}(p-q)\|}{\|p-q\|} : p \neq q, \|p\| < \delta, \|q\| < \delta \right\} = \|A^{-1}\|,$$

and the conclusion is proved by Lemma 3.2 (whose conclusion (10) holds of course in our case as well). □

Proposition 5.1 shows that Theorems 3.1 and 4.1 are extensions of the condition number theorem in convex quadratic optimization. Work is in progress about condition number theorems for special classes of convex optimization problems, where better results are available (as we plan to present elsewhere).

6. An approach by metric regularity. In this section we show that the condition number $c_2(f)$ is, in a general setting, the modulus of metric regularity of the inverse multifunction of the arg min map. As a corollary we get a weak form of the condition number theorem in the finite-dimensional setting, making use of some result in [5].

The following known definitions (see [5] and [8]) will be needed. We consider two real Banach spaces P, Q, a set-valued mapping

$$F : P \to Q,$$

and a point $(x_0, y_0) \in P \times Q$ such that $y_0 \in F(x_0)$. The mapping F is called *metrically regular* at (x_0, y_0) if there exists a constant $M > 0$ such that

$$\text{dist } [x, F^{-1}(y)] \leq M \text{ dist } [y, F(x)] \tag{22}$$

for all (x, y) close to (x_0, y_0). Here distances are taken with respect to the norms of P, Q; for example,

$$\text{dist } [y, F(x)] = \inf \{\|y - u\| : u \in F(x)\}$$

(the infimum over the empty set is $+\infty$). The *regularity modulus* of F at (x_0, y_0) is defined by

$$\text{reg } F(x_0, y_0) = \inf \{M > 0 : (22) \text{ holds } \}.$$

The *radius of metric regularity* of F at (x_0, y_0) is defined by

$$\begin{aligned} &\text{rad } F(x_0, y_0) \\ &= \inf \{\|G\| : G \in L(P, Q), F + G \text{ is not metrically regular at } [x_0, y_0 + G(x_0)]\}, \end{aligned} \tag{23}$$

where $L(P, Q)$ denotes the space of all linear bounded operators acting between P and Q.

We consider proper functions

$$f : B(0, L) \to R \cup \{+\infty\}$$

such that (8) holds. Then the arg min map corresponding to the optimization problem $(B(0, L), f)$ assigns to each $p \in E^*$ with $\|p\| \leq r$ sufficiently small the unique minimizer $m(p) = m(f, p)$. Thus

$$\text{arg min } = m : B(0, r) \to E$$

with inverse

$$m^{-1} : B(0, L) \subset E \to B(0, r) \subset E^*.$$

Given f fulfilling (8), the problem $(B(0, L), f)$ will be called *well-conditioned* if $c_2(f) < +\infty$.

The main result of this section characterizes the condition number $c_2(f)$ as the modulus of metric regularity of the inverse to the arg min map corresponding to the objective function f. We emphasize that no smoothness assumption is required about f (moreover constraints may be present).

THEOREM 6.1. *Let $f : B(0, L) \to R \cup \{+\infty\}$ fulfill (6) and (8). If m^{-1} is metrically regular at $(0, 0)$, then*

$$c_2(f) = \text{reg} m^{-1}(0, 0).$$

Proof. By (8), $c_2(f)$ is the Lipschitz modulus of the single-valued map m at 0. By a known result (see [5, formula (1.4) with $F = m^{-1}$]), $\text{reg} m^{-1}(0, 0)$ agrees with the infimum of those $k > 0$ which are Lipschitz constants of m at 0. Hence the Lipschitz modulus of m at 0 agrees with the regularity modulus of m^{-1} at $(0, 0)$ (because of (6)), whence the conclusion. □

COROLLARY 6.1. *If f fulfills (6) and (8), then $(B(0, L), f)$ is well-conditioned iff m^{-1} is metrically regular at $(0, 0)$.*

COROLLARY 6.2. *Let f fulfill (6) and (8). If m^{-1} is metrically regular at $(0, 0)$ with locally closed graph there, then*

$$\text{rad } m^{-1}(0, 0) \geq \frac{1}{c_2(f)}.$$

If, moreover, E is finite-dimensional, then

$$\operatorname{rad} m^{-1}(0,0) = \frac{1}{c_2(f)}.$$

Proof. Both conclusions come from [5, Theorem 1.5]. □

As we see from (23) and Corollary 6.1, $\operatorname{rad} m^{-1}(0,0)$ can be considered as a measure of the distance to ill-conditioning of the optimization problem $(B(0,L), f)$. Therefore the first conclusion of Corollary 6.2 is one half of the condition number theorem, similar to Theorem 3.1 under rather different assumptions. The second conclusion is a form of the condition number theorem in the finite-dimensional setting. However, the distance to ill-conditioning defined by rad $m^{-1}(0,0)$ is a not necessarily variational notion, since adding linear continuous transformations to m^{-1} does not correspond in general to perturbing additively the objective function, even in the setting of sections 3 and 4, where $m(p) = Df^{-1}(p)$ for all sufficiently small p (due to (12) or (17)).

Results from [5] show that the same notion of distance to ill-conditioning can be obtained by considering a number of different perturbation classes (not only the class $L(E, E^*)$ in the definition (23)).

Acknowledgments. Thanks to A. Ioffe, who pointed out the link between metric regularity and the distance theorem of [16], and to A. Dontchev for bringing [4] to my attention.

REFERENCES

[1] J. P. Dedieu, *Approximate Solutions of Numerical Problems, Condition Number Analysis and Condition Number Theorem*, Lectures in Appl. Math. 32, AMS, Providence, RI, 1996, pp. 285–293.
[2] J. W. Demmel, *On condition numbers and the distance to the nearest ill-posed problem*, Numer. Math., 51 (1987), pp. 251–289.
[3] J. W. Demmel, *The geometry of ill-conditioning*, J. Complexity, 3 (1987), pp. 201–229.
[4] A. L. Dontchev, W. W Hager, K. Malanowski, and V. M. Veliov, *On quantitative stability in optimization and optimal control*, Set-Valued Anal., 8 (2000), pp. 31–50.
[5] A. L. Dontchev, A. S. Lewis, and R. T. Rockafellar, *The radius of metric regularity*, Trans. Amer. Math. Soc., 355 (2003), pp. 493–517.
[6] C. Eckart and G. Young, *The approximation of one matrix by another of lower rank*, Psychometrika 1 (1936), pp. 211 - 218.
[7] P. Gill, W. Murray, and M. Wright, *Numerical Linear Algebra and Optimization*, Vol. 1, Addison-Wesley, Redwood City, CA, 1991.
[8] A. D. Ioffe, *Metric regularity and subdifferential calculus*, Russian Math. Surveys, 55 (2000), pp. 501–558.
[9] A. S. Lewis, *Ill-conditioned convex processes and conic linear systems*, Math. Oper. Res., 24 (1999), pp. 829–834.
[10] R. A. Poliquin and R. T. Rockafellar, *Tilt stability of a local minimum*, SIAM J. Optim., 8 (1998), pp. 287–299.
[11] J. Renegar, *Is it possible to know a problem instance is ill-posed?*, J. Complexity, 10 (1994), pp. 1–56.
[12] J. Renegar, *Incorporating condition measures into the complexity theory of linear programming*, SIAM J. Optim., 5 (1995), pp. 506–524.
[13] J. Renegar, *Linear programming, complexity theory and elementary functional analysis*, Math. Programming, 70 (1995), pp. 279–351.
[14] J. Schwartz, *Nonlinear Functional Analysis*, Gordon and Breach, New York, 1969.
[15] T. Zolezzi, *Well-posedness and conditioning of optimization problems*, Pliska Stud. Math. Bulgar., 12 (1998), pp. 267–280.
[16] T. Zolezzi, *On the distance theorem in quadratic optimization*, J. Convex Anal., 9 (2002), pp. 693–700.

SIAM J. OPTIM.
Vol. 14, No. 2, pp. 517–533

SHARP ESTIMATES FOR HOFFMAN'S CONSTANT FOR SYSTEMS OF LINEAR INEQUALITIES AND EQUALITIES*

C. ZĂLINESCU†

Abstract. We extend the formulae given by Azé and Corvellec, Belousov and Andronov, and Ng and Zheng for the Hoffman constant associated to the system of linear inequalities $Ax \leq b$ and relate them to that established by Li.

Key words. convex function, Hoffman constant, linear inequality, Lipschitz constant, multifunction

AMS subject classifications. 90C25, 90C48

DOI. S1052623402403505

1. Introduction. Consider X a normed vector space, X^* its topological dual, and $A : X \to \mathbb{R}^m$, $C : X \to \mathbb{R}^l$ continuous linear operators with $m, l \in \mathbb{N}$, $m + l \geq 1$. There exist $a_1, \ldots, a_m, c_{m+1}, \ldots, c_{m+l} \in X^*$ (uniquely determined) such that $Ax = (\langle x, a_1\rangle, \ldots, \langle x, a_m\rangle)$ and $Cx = (\langle x, c_{m+1}\rangle, \ldots, \langle x, c_{m+l}\rangle)$ for every $x \in X$. We are interested in the system of linear inequalities and equalities

$$\langle x, a_i\rangle \leq b_i,\ \langle x, c_j\rangle = d_j, \quad i \in I := \{1, \ldots, m\},\ j \in J := \{m+1, \ldots, m+l\}. \tag{1.1}$$

(Of course, $I = \emptyset$ if $m = 0$ and $J = \emptyset$ if $l = 0$.) Considering the cone $\mathbb{R}^m_+ := \{y \in \mathbb{R}^m \mid y_i \geq 0\ \forall\, i \in I\}$ and the order $\leq$ on $\mathbb{R}^m$ induced by this cone (i.e., $y \leq y'$ if and only if $y' - y \in \mathbb{R}^m_+$), the system (1.1) becomes

$$Ax \leq b, \quad Cx = d \tag{1.2}$$

for $b = (b_1, \ldots, b_m)$, $d = (d_{m+1}, \ldots, d_{m+l})$. Denote by $F(b, d)$ the solution set of (1.2); in this way we get a multifunction $F : \mathbb{R}^{m+l} \rightrightarrows X$ whose domain is $\operatorname{dom} F = \operatorname{Im} \mathcal{A} + \mathbb{R}^m_+ \times \{0\}$, where $\mathcal{A} : X \to \mathbb{R}^{m+l}$, $\mathcal{A}(x) := (Ax, Cx)$. Hoffman showed in his celebrated paper [4] that for every $(b, d) \in \operatorname{dom} F$ there exists $\tau > 0$ such that

$$\tau \cdot d\big(x, F(b, d)\big) \leq \|(Ax - b)_+\| + \|Cx - d\| \quad \forall\, x \in X,$$

where, for $\gamma \in \mathbb{R}$, $\gamma_+ := \max\{0, \gamma\}$, while for $y \in \mathbb{R}^m$, $y_+ := \big((y_1)_+, \ldots, (y_m)_+\big)$. In fact, because the space $\mathbb{R}^{m+l}$ can appear as a whole, we are interested in an estimate of the form

$$\tau \cdot d\big(x, F(b, d)\big) \leq \|((Ax - b)_+, Cx - d)\| \quad \forall\, x \in X, \tag{1.3}$$

where $\|\cdot\|$ is a norm on $\mathbb{R}^{m+l}$; of course, the existence of $\tau > 0$ satisfying (1.3) follows from Hoffman's statement because all norms on $\mathbb{R}^{m+l}$ are equivalent.

Remark 1. Let $(b, d) \in \operatorname{dom} F$; then there exist $x_0 \in X$ and $b_0 \in \mathbb{R}^m_+$ such that $b = Ax_0 + b_0$ and $Cx_0 = d$. It is obvious that $F(b, d) = x_0 + F(b_0, 0)$, and so (1.3) holds if and only if (1.3) holds for b replaced by b_0 and d replaced by 0. So, when estimating

*Received by the editors March 5, 2002; accepted for publication (in revised form) June 8, 2003; published electronically November 6, 2003.
http://www.siam.org/journals/siopt/14-2/40350.html
†University "Al. I. Cuza" Iaşi, Faculty of Mathematics, Bd. Carol I, Nr. 11, 700506 Iaşi, Romania (zalinesc@uaic.ro).

τ in (1.3) we may take only $b \in \mathbb{R}^m_+$. In particular, when $A|_{\ker C}$ is surjective one can take $b_0 = 0$, and so τ is independent on $(b,d) \in \operatorname{dom} F = \mathbb{R}^m \times \operatorname{Im} C$.

Remark 2. We could take $C = 0$ because the equality $Cx = d$ may be replaced by the system of inequalities $Cx \le d$, $(-C)x \le -d$. This is not a very good procedure because one must change the space $\mathbb{R}^{m+l}$ with the space $\mathbb{R}^{m+l+l}$; in such a situation one must decide what norm to choose on the last space. As we shall see later we need to impose some supplementary conditions on the behavior of the norm $\|(y,z)\|$ in the variables y_i with $i \in I$ which are not needed for variables z_j with $j \in J$; replacing an equality by two inequalities, we have to impose such conditions on all the variables of the norm on $\mathbb{R}^{m+l+l}$. Another reason is furnished by the preceding remark; if $A = 0$, we observe that τ does not depend on $d \in \operatorname{dom} F$, but the system of inequalities depends on the elements in the domain of F.

Note that one can assume, at least theoretically, that X is finite-dimensional, or reflexive. Indeed, considering $\widehat{X} := X/\ker \mathcal{A}$ endowed with the quotient norm, $\widehat{\mathcal{A}} : \widehat{X} \to \mathbb{R}^{m+l}$ defined by $\widehat{\mathcal{A}}\widehat{x} := \big(\widehat{A}\widehat{x}, \widehat{C}\widehat{x}\big) := (Ax, Cx)$ ($\widehat{x}$ being the class $x + \ker \mathcal{A}$ of x) and $\widehat{F}(b,d) := \{\widehat{x} \in \widehat{X} \mid \widehat{A}\widehat{x} \le b,\ \widehat{C}\widehat{x} = d\}$, one has, as observed by Ng and Zheng [11] in the case $l = 0$, that $d\big(\widehat{x}, \widehat{F}(b,d)\big) = d\big(x, F(b,d)\big)$ for every $x \in X$, and so (1.3) holds if and only if

$$\tau \cdot d\big(\widehat{x}, \widehat{F}(b)\big) \le \big\|\big((\widehat{A}\widehat{x} - b)_+, \widehat{C}\widehat{x} - d\big)\big\| \quad \forall \widehat{x} \in \widehat{X}.$$

The consideration of equalities in the system (1.2) is inspired by Li's articles [8], [9]; in these articles the author is interested by Lipschitz constants for the feasible multifunction F, as well as for the solution multifunction S of a linear programming problem whose feasible set is given by (1.2). We say that $\gamma \ge 0$ is a Lipschitz constant for F at $(b,d) \in \operatorname{dom} F$ if

$$e\left(F(b',d'), F(b,d)\right) \le \gamma \cdot \|(b',d') - (b,d)\| \quad \forall\, (b',d') \in \mathbb{R}^{m+l},$$

where $e(D,E) := \sup_{x \in D} d(x,E)$ is the Hausdorff–Pompeiu excess of D over E with $e(\emptyset, E) := 0$, the distance $d(x,E)$ from x to E being defined by $d(x,E) := \inf\{\|x - y\| \mid y \in E\}$ with $d(x,\emptyset) := +\infty$. In fact there exists a deep relationship between the Hoffman and Lipschitz constants of F at (b,d), as we shall see in what follows (see also Belousov and Andronov's article [2]).

As the existence of $\tau > 0$ satisfying (1.3) is ensured by Hoffman's theorem, an important problem is to give computable estimates of τ, or even formulae for the sharp τ. Of course, such estimates will depend on the norms on X and $\mathbb{R}^{m+l}$. Recently, in the case $l = 0$ (and so $C = 0$, $d = 0$), Azé and Corvellec [1] obtained a formula and Ng and Zheng [11] obtained an estimate (which in fact is a formula, as we shall see later on) for the sharp Hoffman constant at $(b,d) \in \operatorname{dom} F$,

$$\delta_{b,d} := \inf_{x \in X \setminus F(b,d)} \frac{\|((Ax - b)_+, Cx - d)\|}{d\big(x, F(b,d)\big)}, \tag{1.4}$$

when $\mathbb{R}^m$ is endowed with the box norm $\|\cdot\|_\infty$ (the norm on X being arbitrary), while Belousov and Andronov [2] obtained a formula for the sharp uniform Hoffman constant

$$\delta = \inf\{\delta_{b,d} \mid (b,d) \in \operatorname{dom} F\} \tag{1.5}$$

when $X = \mathbb{R}^k$ is endowed with the Euclidean norm and $\mathbb{R}^m$ is endowed with a (pseudo)norm $\|\cdot\|$ satisfying the condition

$$\|y\| \le \|z\| \quad \forall\, y, z \in \mathbb{R}^m,\ 0 \le y \le z. \tag{1.6}$$

Because not only these pairs of norms are useful, our aim is to give formulae for $\delta_{b,d}$ and δ also for other pairs of norms. Note that the estimates for the Lipschitz constant of Bergthaller and Singer [3] for the inequality system $Ax \le b$ are established for the norm $\|\cdot\|_\infty$ on $\mathbb{R}^m$ and an arbitrary norm on X (although the authors say their method works for the norm $\|\cdot\|_p$ on $\mathbb{R}^m$) while the sharp global Lipschitz constant established by Li [8], [9] are for C surjective and arbitrary norms on $X = \mathbb{R}^k$ and $\mathbb{R}^{m+l}$.

In order to obtain such formulae we use a result established in [14].

Since the case $\mathcal{A} = 0$ is trivial, in what follows we assume that $\mathcal{A} \ne 0$; in this case $F(b,d) \ne X$ for every $(b,d) \in \mathbb{R}^{m+l}$. (We could even assume that $a_i \ne 0$ for every $i \in I$ and $c_j \ne 0$ for every $j \in J$.)

2. Preliminary notions and results. Throughout this paper X is a real normed vector space whose norm is $\|\cdot\|$; its topological dual is denoted by X^* and the dual norm is denoted by $\|\cdot\|_*$. The value of $x^* \in X^*$ at $x \in X$ is denoted, as usual, by $\langle x, x^*\rangle$. The duality mapping of X is the multifunction $\Phi_X : X \rightrightarrows X^*$ defined by

$$\Phi_X(x) := \{x^* \in X^* \mid \langle x, x^*\rangle = \|x\|^2 = \|x^*\|_*^2\};$$

$\Phi_X(x)$ is nothing else but the (Fenchel) subdifferential of the function $\frac{1}{2}\|\cdot\|^2$ at x.

Let $C \subset X$ be a nonempty closed convex set. For every $x \in X$ we set $P_C(x) := \{c \in C \mid d(x,C) = \|x - c\|\}$. It is known that $\overline{x} \in P_C(x)$ if and only if $\overline{x} \in C$ and $\Phi_X(x - \overline{x}) \cap N(C, \overline{x}) \ne \emptyset$, where $N(C,\overline{x})$ is the normal cone of C at $\overline{x}$ defined by

$$N(C,\overline{x}) := \{x^* \in X^* \mid \langle c - \overline{x}, x^*\rangle \le 0\ \forall\, c \in C\}.$$

It follows that for $\overline{x} \in C$ and $u \in \Phi_X^{-1}\big(N\big(C,\overline{x}\big)\big) \cap S_X$ we have $\overline{x} \in P_C(\overline{x} + tu)$ and $d(\overline{x} + tu, C) = t$ for every $t \ge 0$; as usual, $S_X := \{u \in X \mid \|u\| = 1\}$. When X is reflexive $P_C(x)$ is nonempty for every $x \in X$. Recall now the following result which is stated in [14, Prop. 3.5] (see also [15, Thm. 3.10.7]), where, as usual, for the proper convex function $f : X \to \overline{\mathbb{R}}$, $\operatorname{dom} f := \{x \in X \mid f(x) < \infty\}$ is the domain of f, $\partial f(x) := \{x^* \in X^* \mid \langle x' - x, x^*\rangle \le f(x') - f(x)\ \forall\, x' \in X\}$ is the subdifferential of f at $x \in \operatorname{dom} f$ ($\partial f(x) := \emptyset$ for $x \in X \setminus \operatorname{dom} f$), $f'(x,u) := \lim_{t \to 0_+} t^{-1}\big(f(x+tu) - f(x)\big)$ is the directional derivative of f at $x \in \operatorname{dom} f$ in the direction $u \in X$, and $[f \le t] := \{x \in X \mid f(x) \le t\}$ and $[f = t] := \{x \in X \mid f(x) = t\}$ are the sublevel and the level sets of f at height $t \in \mathbb{R}$, respectively; $\mathcal{C}(A,x) := \operatorname{cl}\big(\operatorname{cone}(A - x)\big)$ is the closed conic hull of $A - x$ for $A \subset X$ and $x \in X$.

Proposition 2.1. *Let X be a Banach space and $f : X \to \overline{\mathbb{R}}$ be a proper lower semicontinuous convex function. Assume that $t \in [\inf f, \infty[$ is such that $[f \le t] \ne \emptyset$. Then*

$$l_f(t) := \inf_{x \in \operatorname{dom} f \setminus [f \le t]} \frac{f(x) - t}{d\,(x, [f \le t])} = d\big(0, \partial f(X \setminus [f \le t])\big). \tag{2.1}$$

Assume now that X is reflexive. Then

$$\begin{aligned} l_f(t) &= \inf\left\{ f'\left(y, \frac{x-y}{\|x-y\|}\right) \,\middle|\, x \in \operatorname{dom} f \setminus [f \le t],\ y \in P_{[f\le t]}(x)\right\} \\ &= \inf\left\{f'(y,u) \mid y \in [f=t],\ u \in S_X \cap \Phi_X^{-1}\big(N([f\le t], y)\big)\right\} \\ &= \inf\left\{ \frac{f'(y,u)}{d\big(u, \mathcal{C}([f\le t], y)\big)} \,\middle|\, y \in [f=t],\ u \in X \setminus \mathcal{C}([f\le t], y)\right\}. \end{aligned} \tag{2.2}$$

Moreover, if $t > \inf f$, then

$$l_f(t) = \inf\left\{ f'\left(y, \frac{u}{\|u\|}\right) \,\middle|\, y \in [f=t],\ u \in \Phi_X^{-1}\big(\partial f(y)\big)\right\}.$$

When $t = 0$, the positivity of $l_f(0)$ defined above is equivalent to the existence of a global error bound for the inequality system $f(x) \le 0$. When X is finite-dimensional and the function f is convex there are many results concerning error bounds (see the recent papers [7], [5] and the references therein). For X infinite-dimensional the above result is one of the most general to our knowledge. Lemaire [6, Prop. 7.1], Azé and Corvellec [1, Thm. 2.2], and Wu and Ye [12] stated formula (2.1) in the present form, while Ng and Zheng [11, Thm. 3.3] stated it for X a reflexive Banach space and f a finite-valued continuous convex function.

We mention that throughout this paper the space $\mathbb{R}^k$ (k will be m, l, or $m+l$) is endowed with an (arbitrary) norm denoted also by $\|\cdot\|$; when needed, we shall specify supplementary conditions on $\|\cdot\|$. We identify the topological dual of the normed space $\big(\mathbb{R}^k, \|\cdot\|\big)$ with $\mathbb{R}^k$ by the pairing

$$\langle y, \mu\rangle := y_1\mu_1 + \cdots + y_m\mu_m.$$

So the dual norm $\|\cdot\|_*$ on $\mathbb{R}^k$ is defined by $\|\mu\|_* := \sup\{\langle y,\mu\rangle \mid \|y\| \le 1\}$. The (positive) dual cone of $E \subset \mathbb{R}^m$ is $E^+ := \{\mu \in \mathbb{R}^m \mid \langle y,\mu\rangle \ge 0\ \forall\, y \in E\}$; it is obvious that $(\mathbb{R}^m_+)^+ = \mathbb{R}^m_+$.

It is obvious that $\delta_{b,d}$ defined in (1.4) is exactly $l_f(0)$ for

$$f : X \to \mathbb{R}, \quad f(x) := \|((Ax-b)_+, Cx-d)\|. \tag{2.3}$$

In order to apply the preceding result we need the convexity of f, which is ensured by the convexity of the positively homogeneous function

$$h : \mathbb{R}^{m+l} \to \mathbb{R}, \quad h(y,z) := \|(y_+, z)\|. \tag{2.4}$$

In this sense the next result is useful.

LEMMA 2.2. *Let h be defined by* (2.4). *Then the following hold:*

(i) *h is sublinear if and only if the norm $\|\cdot\|$ satisfies the condition*

$$\|(y,z)\| \le \|(y',z)\| \quad \forall\, z \in \mathbb{R}^l,\ \forall\, y, y' \in \mathbb{R}^m,\ 0 \le y \le y'. \tag{2.5}$$

(ii) *Assume that $\|\cdot\|$ satisfies condition* (2.5). *Then*

$$\partial h(0,0) \subset \mathbb{R}^m_+ \times \mathbb{R}^l,$$
$$\partial h(y,z) = \{(\mu,\zeta) \in \partial h(0,0) \mid \langle y,\mu\rangle + \langle z,\zeta\rangle = \|(y_+,z)\|\} \quad \forall\, y \in \mathbb{R}^m,\ \forall\, z \in \mathbb{R}^l.$$

Moreover, $\partial h(0,0) \subset \{(\mu,\zeta) \in \mathbb{R}^{m+l} \mid \|(\mu,\zeta)\|_* \le 1\}$ *if and only if*

$$\|(y_+,z)\| \le \|(y,z)\| \quad \forall\, y \in \mathbb{R}^m,\ \forall\, z \in \mathbb{R}^l. \tag{2.6}$$

(iii) *Assume that* $\|\cdot\|$ *satisfies conditions* (2.5) *and* (2.6). *Then*

$$\partial h(0,0) = \{(\mu,\zeta) \in \mathbb{R}^m_+ \times \mathbb{R}^l \mid \|(\mu,\zeta)\|_* \le 1\}. \tag{2.7}$$

Moreover, if $(\mu,\zeta) \in \partial h(y,z)$, *then* $\langle y_-,\mu\rangle = 0$, *and so* $\mu_i = 0$ *whenever* $y_i < 0$, *where* $y_- := (-y)_+$. *In particular, if* $(y_+,z) \ne (0,0)$ *and* $(\mu,\zeta) \in \partial h(y,z)$, *then* $\|(\mu,\zeta)\|_* = 1$.

Proof. (i) Assume that $\|\cdot\|$ satisfies condition (2.5) and take $y, y' \in \mathbb{R}^m$, $z, z' \in \mathbb{R}^l$. Then $0 \le (y+y')_+ \le y_+ + y'_+$, and so

$$\begin{aligned} h\big((y,z)+(y',z')\big) &= \|((y+y')_+, z+z')\| \le \big\|\big(y_+ + y'_+, z+z'\big)\big\| \\ &\le \|(y_+,z)\| + \big\|(y'_+,z')\big\| = h(y,z) + h(y',z'). \end{aligned}$$

As h is obviously positively homogeneous, h is sublinear. Conversely, assume that h is sublinear and take $y, y' \in \mathbb{R}^m$ with $0 \le y \le y'$ and $z \in \mathbb{R}^l$. Then $y = y' + (-v)$ for some $v \ge 0$, and so $h(y,z) \le h(y',z) + h(-v,0) = h(y',z)$. It follows that $\|(y,z)\| = h(y,z) \le h(y',z) = \|(y',z)\|$.

(ii) Let $\|\cdot\|$ satisfy condition (2.5). Consider $(\mu,\zeta) \in \partial h(0,0)$. Then for every $y \ge 0$ we have that $\langle -y,\mu\rangle + \langle 0,\zeta\rangle \le \|((-y)_+,0)\| = 0$, and so $\mu \in \big(\mathbb{R}^m_+\big)^+ = \mathbb{R}^m_+$. The formula for $\partial h(y,z)$ for arbitrary $y \in \mathbb{R}^m$ and $z \in \mathbb{R}^l$ follows by a well-known result for sublinear functions (see, for example, [15, Thm. 2.4.14(iii)]).

Assume that (2.6) holds. Then for $(\mu,\zeta) \in \partial h(0,0)$ and $y \in \mathbb{R}^m$, $z \in \mathbb{R}^l$ we have that $\langle y,\mu\rangle + \langle z,\zeta\rangle \le \|(y_+,z)\| \le \|(y,z)\|$, and so $\|(\mu,\zeta)\|_* \le 1$. Conversely, assume that $\partial h(0,0) \subset \{(\mu,\zeta) \in \mathbb{R}^{m+l} \mid \|(\mu,\zeta)\|_* \le 1\}$ and take $y \in \mathbb{R}^m$, $z \in \mathbb{R}^l$. Then $h(y,z) = \langle y,\overline{\mu}\rangle + \langle z,\overline{\zeta}\rangle$ for some $\big(\overline{\mu},\overline{\zeta}\big) \in \partial h(0,0)$, and so

$$\|(y_+,z)\| = h(y,z) \le \max\{\langle y,\mu\rangle + \langle z,\zeta\rangle \mid \|(\mu,\zeta)\|_* \le 1\} = \|(y,z)\|.$$

(iii) Assume that $\|\cdot\|$ satisfies (2.5) and (2.6). The inclusion $\partial h(0,0) \subset \{(\mu,\zeta) \in \mathbb{R}^m_+ \times \mathbb{R}^l \mid \|(\mu,\zeta)\|_* \le 1\}$ is immediate from (i) and (ii). Let $\mu \in \mathbb{R}^m_+$ and $\zeta \in \mathbb{R}^l$ be such that $\|(\mu,\zeta)\|_* \le 1$. Then for $y \in \mathbb{R}^m$ and $z \in \mathbb{R}^l$ we have that $\langle y,\mu\rangle + \langle z,\zeta\rangle \le \langle y_+,\mu\rangle + \langle z,\zeta\rangle \le \|(y_+,z)\| \cdot \|(\mu,\zeta)\|_* \le h(y,z)$, which means that $(\mu,\zeta) \in \partial h(0,0)$. Therefore (2.7) holds.

Now let $(\mu,\zeta) \in \partial h(y,z)$ $(\subset \partial h(0,0))$. Then

$$\begin{aligned} \|(y_+,z)\| &= \langle y,\mu\rangle + \langle z,\zeta\rangle = \langle y_+ - y_-,\mu\rangle + \langle z,\zeta\rangle = \langle y_+,\mu\rangle - \langle y_-,\mu\rangle + \langle z,\zeta\rangle \\ &\le \langle y_+,\mu\rangle + \langle z,\zeta\rangle \le \|(y_+,z)\| \cdot \|(\mu,\zeta)\|_* \le \|(y_+,z)\|, \end{aligned}$$

whence $\langle y_-,\mu\rangle = 0$ and $\langle y_+,\mu\rangle + \langle z,\zeta\rangle = \|(y_+,z)\|$. Since $\|(\mu,\zeta)\|_* \le 1$, when $(y_+,z) \ne 0$, from the last equality, we get $\|(\mu,\zeta)\|_* = 1$. □

Note that the norm $\|\cdot\|$ on $\mathbb{R}^{m+l}$ satisfies conditions (2.5) and (2.6) if and only if

$$\|(y_+,z)\| \le \|(y+y',z)\| \quad \forall\, y \in \mathbb{R}^m,\ \forall\, y' \in \mathbb{R}^m_+,\ \forall\, z \in \mathbb{R}^l, \tag{2.8}$$

or, equivalently,

$$d\big((0,0),(y,z)+\mathbb{R}^m_+ \times \{0\}\big) = \|(y_+,z)\| \quad \forall\, y \in \mathbb{R}^m,\ \forall\, z \in \mathbb{R}^l. \tag{2.9}$$

A sufficient condition for (2.5) and (2.6) to hold is

$$\|(y,z)\| = \|(|y|,z)\| \quad \forall\, y \in \mathbb{R}^m,\ \forall\, z \in \mathbb{R}^l, \tag{2.10}$$

where $|(y_1,\dots,y_m)| := (|y_1|,\dots,|y_m|)$.

Indeed, if (2.10) holds, then the mapping $t \mapsto \|(t,y_2,\dots,y_m,z)\|$ from $\mathbb{R}$ into $\mathbb{R}$ is an even convex function; hence it attains its infimum at 0 and is nondecreasing on $\mathbb{R}_+$. The same is true for every variable y_i. Hence (2.5) holds; (2.6) holds too because $0 \le y_+ \le |y|$.

Of course, the norm $\|\cdot\|$ on $\mathbb{R}^{m+l}$ satisfies (2.10) whenever the condition below holds:

$$\|v\| = \|\,|v|\,\| \quad \forall\, v \in \mathbb{R}^{m+l}. \tag{2.11}$$

Note that the usual norm $\|\cdot\|_p$ on $\mathbb{R}^{m+l}$ ($\|v\|_p := (\sum_{i=1}^{m+l} |v_i|^p)^{1/p}$ for $p \in [1,\infty[$, $\|v\|_\infty := \max\{|v_i| \mid 1 \le i \le m+l\}$) verifies condition (2.11), and so it verifies conditions (2.5) and (2.6), too. Recall that the dual norm of $\|\cdot\|_p$ is $\|\cdot\|_q$ with $q \in [1,\infty]$, $1/p + 1/q = 1$.

LEMMA 2.3. *Assume that the norm $\|\cdot\|$ on $\mathbb{R}^{m+l}$ satisfies condition* (2.10) *and consider $\Phi := \Phi_{\mathbb{R}^{m+l}}$ the duality mapping of $\mathbb{R}^{m+l}$. Then*

(i) *the norm $\|\cdot\|_*$ on $\mathbb{R}^{m+l}$ satisfies* (2.10)*;*

(ii) $\forall\,(y,z) \in \mathbb{R}^{m+l}$, $\forall\,(\mu,\zeta) \in \Phi(y,z)$, $\forall\, i \in I\ :\ y_i\mu_i \ge 0$;

(iii) $\forall\,(y,z) \in \mathbb{R}^m_+ \times \mathbb{R}^l$, $\exists\,(\mu,\zeta) \in \Phi(y,z) \cap (\mathbb{R}^m_+ \times \mathbb{R}^l)$, $\forall\, i \in I\ :\ y_i = 0 \Rightarrow \mu_i = 0$.

(iv) *Moreover, if the norm $\|\cdot\|$ on $\mathbb{R}^{m+l}$ satisfies* (2.11)*, then*

$$\forall\,(y,z) \in \mathbb{R}^m_+ \times \mathbb{R}^l,\ \exists\,(\mu,\zeta) \in \Phi(y,z) \cap (\mathbb{R}^m_+ \times \mathbb{R}^l),$$
$$\forall\, i \in I,\ \forall\, j \in J\ :\ y_i = 0 \Rightarrow \mu_i = 0,\ z_j = 0 \Rightarrow \zeta_j = 0.$$

Proof. (i) Let $\mu, \overline{\mu} \in \mathbb{R}^m$ be such that $|\mu_i| = |\overline{\mu}_i|$ for every $i \in I$, and $\zeta \in \mathbb{R}^l$. There exists $(y,z) \in \mathbb{R}^{m+l}$ with $\|(y,z)\| = 1$ such that $\|(\mu,\zeta)\|_* = \sum_{i\in I} y_i\mu_i + \sum_{j\in J} z_j\zeta_j$. Take $\overline{y}_i := y_i$ if $\mu_i\overline{\mu}_i \ge 0$ and $\overline{y}_i := -y_i$ if $\mu_i\overline{\mu}_i < 0$. Because the norm $\|\cdot\|$ satisfies (2.10) we have that $\|(y,z)\| = \|(\overline{y},z)\| = 1$. But $\sum_{i\in I} y_i\mu_i + \sum_{j\in J} z_j\zeta_j = \sum_{i\in I} \overline{y}_i\overline{\mu}_i + \sum_{j\in J} z_j\zeta_j \le \|(\overline{y},z)\| \cdot \|(\overline{\mu},\zeta)\|_* = \|(\overline{\mu},\zeta)\|_*$, whence $\|(\mu,\zeta)\|_* \le \|(\overline{\mu},\zeta)\|_*$. Hence $\|(\mu,\zeta)\|_* = \|(\overline{\mu},\zeta)\|_*$, which implies that $\|(\mu,\zeta)\|_* = \|(|\mu|,\zeta)\|$.

(ii) Let $(y,z) \in \mathbb{R}^{m+l}$ and $(\mu,\zeta) \in \Phi(y,z)$. Then $\|(\mu,\zeta)\|_*^2 = \|(y,z)\|^2 = \sum_{i\in I} y_i\mu_i + \sum_{j\in J} z_j\zeta_j$. Assume that $y_{i_0}\mu_{i_0} < 0$ for some $i_0 \in I$. Taking $\overline{y}_i := y_i$ for $i \in I \setminus \{i_0\}$ and $\overline{y}_{i_0} := -y_{i_0}$, we have that $\|(\overline{y},z)\| = \|(y,z)\|$ and so we get the contradiction

$$\begin{aligned}\|(\mu,\zeta)\|_*^2 &= \sum_{i\in I} y_i\mu_i + \sum_{j\in J} z_j\zeta_j < \sum_{i\in I} \overline{y}_i\mu_i + \sum_{j\in J} z_j\zeta_j \\ &\le \|(\overline{y},z)\| \cdot \|(\mu,\zeta)\|_* = \|(\mu,\zeta)\|_*^2.\end{aligned}$$

(iii) Let $(y,z) \in \mathbb{R}^m_+ \times \mathbb{R}^l$ and take $(\overline{\mu},\overline{\zeta}) \in \Phi(y,z)$. Set $I_0 := \{i \in I \mid y_i > 0\}$. Consider $\mu_i := \overline{\mu}_i$ for $i \in I_0$ and $\mu_i := 0$ for $i \in I \setminus I_0$. Then $\mu := (\mu_1,\dots,\mu_m) \in \mathbb{R}^m_+$ and $\mu_i = 0$ whenever $y_i = 0$. Because $|\mu_i| \le |\overline{\mu}_i|$ for all $i \in I$, by (i) we have that $\|(\mu,\overline{\zeta})\|_* \le \|(\overline{\mu},\overline{\zeta})\|_* = \|(y,z)\|$. On the other hand,

$$\begin{aligned}\|(\overline{\mu},\overline{\zeta})\|_*^2 &= \sum_{i\in I} y_i\overline{\mu}_i + \sum_{j\in J} z_j\overline{\zeta}_j = \sum_{i\in I} y_i\mu_i + \sum_{j\in J} z_j\overline{\zeta}_j \le \|(y,z)\| \cdot \|(\mu,\overline{\zeta})\|_* \\ &= \|(\overline{\mu},\overline{\zeta})\|_* \cdot \|(\mu,\overline{\zeta})\|_*,\end{aligned}$$

whence $\|(\overline{\mu},\overline{\zeta})\|_* \le \|(\mu,\overline{\zeta})\|_*$. Hence $\|(\overline{\mu},\overline{\zeta})\|_* = \|(\mu,\overline{\zeta})\|_* = \|(y,z)\|$; from the equality $\sum_{i\in I} y_i\overline{\mu}_i + \sum_{j\in J} z_j\overline{\zeta}_j = \sum_{i\in I} y_i\mu_i + \sum_{j\in J} z_j\overline{\zeta}_j$ we get $(\mu,\overline{\zeta}) \in \Phi(y,z)$.

(iv) In the proof of (iii) we consider also $J_0 := \{j \in J \mid z_j \neq 0\}$ and take $\zeta_j := \overline{\zeta}_j$ for $j \in J_0$ and $\zeta_j := 0$ for $j \in J \setminus J_0$. Proceeding as in the proof of (iii) we obtain that $(\mu,\zeta) \in \Phi(y,z)$. □

Take $\|\cdot\| = \|\cdot\|_p$ with $p \in [1,\infty]$ and $(y,z) \in \mathbb{R}^{m+l}$ with $(y_+,z) \neq (0,0)$. Then for $p = 1$,

$$\partial h(y,z) = \big\{(\mu,\zeta) \in [0,1]^m \times [-1,1]^l \mid y_i < 0 \Rightarrow \mu_i = 0, \\ y_i > 0 \Rightarrow \mu_i = 1,\ z_j \neq 0 \Rightarrow \zeta_j = \operatorname{sgn} z_j\big\},$$

where $\operatorname{sgn}\alpha := \alpha/|\alpha|$ for $\alpha \in \mathbb{R}\setminus\{0\}$, $\operatorname{sgn} 0 := 0$; for $1 < p < \infty$,

$$\partial h(y,z) = \left\{\|(y_+,z)\|_p^{1-p}\left((y_1)_+^{p-1},\dots,(y_m)_+^{p-1},|z_1|^{p-1}\operatorname{sgn} z_1,\dots,|z_l|^{p-1}\operatorname{sgn} z_l\right)\right\},$$

and for $p = \infty$,

$$\partial h(y,z) = \left\{(\mu,\zeta) \in \mathbb{R}^m_+ \times \mathbb{R}^l \;\middle|\; \sum_{i\in I}\mu_i + \sum_{j\in J}|\zeta_j| = 1,\ y_i < \|(y_+,z)\|_\infty \Rightarrow \mu_i = 0,\right.$$

$$\left. |z_j| < \|(y_+,z)\|_\infty \Rightarrow \zeta_j = 0,\ z_j\zeta_j \ge 0\ \forall\, j \in J\right\}. \tag{2.12}$$

As noted in the introduction, there are strong relationships between the Hoffman and Lipschitz constants for the multifunction F. The next result was observed in [13].

LEMMA 2.4. *Consider the multifunction $\Gamma : X \rightrightarrows Y$, $x_0 \in \operatorname{dom}\Gamma$, and $\gamma \in [0,\infty[$, where $(Y,\|\cdot\|)$ is another normed vector space. Then*

$$d\big(y,\Gamma(x_0)\big) \le \gamma d\left(x_0,\Gamma^{-1}(y)\right)\ \ \forall y \in Y \tag{2.13}$$

if and only if

$$e\big(\Gamma(x),\Gamma(x_0)\big) \le \gamma\,\|x - x_0\|\ \ \forall\, x \in X. \tag{2.14}$$

Condition (2.13) means that the multifunction Γ has a global error bound at x_0 as introduced by Li and Singer [10].

Note that for the multifunction $F : \mathbb{R}^{m+l} \rightrightarrows X$ defined in the introduction we have that $F^{-1}(x) = (Ax,Cx) + \mathbb{R}^m_+ \times \{0\}$, and so relation (2.13) becomes

$$d\big(x,F(b,d)\big) \le \gamma d\left((b,d),(Ax,Cx) + \mathbb{R}^m_+ \times \{0\}\right).$$

But

$$d\big((b,d),(Ax,Cx) + \mathbb{R}^m_+ \times \{0\}\big) = d\left((0,0),(Ax-b,Cx-d) + \mathbb{R}^m_+ \times \{0\}\right) \\ \le \|((Ax-b)_+,Cx-d)\|,$$

with equality if $\|\cdot\|$ satisfies conditions (2.5) and (2.6) (or, equivalently, (2.9)).

COROLLARY 2.5. *Let $(b,d) \in \operatorname{dom} F$. If*

$$e\left(F(b',d'),F(b,d)\right) \le \gamma\,\|(b',d') - (b,d)\|\ \ \forall\,(b',d') \in \operatorname{dom} F,$$

then

$$d\big(x, F(b,d)\big) \le \gamma \,\|((Ax-b)_+, Cx-d)\| \;\; \forall\, x \in X.$$

Moreover, if the norm $\|\cdot\|$ *on* $\mathbb{R}^{m+l}$ *satisfies conditions* (2.5) *and* (2.6)*, then the converse holds.*

Proof. Just note that in (2.14) one can take only $x \in \operatorname{dom}\Gamma$. Then apply the preceding lemma and the above discussion. □

In fact, the preceding corollary is valid when A is replaced by an arbitrary function $f : X \to \mathbb{R}^m$ (or even defined on a subset of X). In such a case Corollary 2.5 (for $l = 0$) was established by Belousov and Andronov [2] with the norm $\|\cdot\|$ replaced by a function $g : \mathbb{R}^m \to \mathbb{R}_+$ satisfying similar conditions to (2.5) and (2.6).

3. Extensions of the Belousov–Andronov formula. Throughout this section X is a reflexive Banach space; of course, $\Phi_{X^*} = (\Phi_X)^{-1}$ in this case.

Let $A : X \to \mathbb{R}^m$, $C : X \to \mathbb{R}^l$, $b \in \mathbb{R}^m$, $d \in \mathbb{R}^l$, and $F : \mathbb{R}^{m+l} \rightrightarrows X$ be as in the introduction. The adjoint C^* of C is given by $C^*\zeta = \sum_{j\in J} \zeta_j c_j$, and similarly A^*. For $K \in \mathcal{P}(I) := \{L \mid L \subset I\}$ we set $C_\emptyset := \{0\}$ and

$$C_K := \operatorname{cone}\{a_i \mid i \in K\} = \left\{ \sum_{i\in K} \mu_i a_i \;\middle|\; \mu_i \ge 0 \;\forall\, i \in K \right\}$$

when $K \ne \emptyset$. For $u \in X$ and $K \in \mathcal{P}(I)$ we consider the element $\eta_K^u \in \mathbb{R}^m$ having the components $(\eta_K^u)_i := (\langle u, a_i\rangle)_+$ for $i \in K$ and $(\eta_K^u)_i := 0$ for $i \in I \setminus K$. If $u \in S_X$ and $\Phi_X(u) \cap C_K \ne \emptyset$, then $(\eta_K^u)_i > 0$ for some $i \in K$; indeed, taking $x^* \in \Phi_X(u) \cap C_K$ we have that $1 = \langle u, x^*\rangle = \sum_{i\in K} \mu_i \langle u, a_i\rangle$ with $\mu_i \ge 0$. Let $(b,d) \in \operatorname{dom} F$ be fixed and consider $x \in F(b,d)$; then $I_b(x) := \{i \in I \mid \langle x, a_i\rangle = b_i\} \in \mathcal{P}(I)$ and the normal cone of $F(b,d)$ at x is

$$\begin{aligned} N\big(F(b,d),x\big) &= \left\{ \sum_{i\in I_b(x)} \mu_i a_i + \sum_{j\in J} \zeta_j c_j \;\middle|\; \mu_i \in \mathbb{R}_+ \;\forall\, i \in I_b(x),\; \zeta_j \in \mathbb{R} \;\forall\, j \in J \right\} \\ &= C_{I_b(x)} + \operatorname{Im} C^*. \end{aligned} \tag{3.1}$$

THEOREM 3.1. *Assume that the norm* $\|\cdot\|$ *on* $\mathbb{R}^{m+l}$ *satisfies condition* (2.5)*. Then for every* $(b,d) \in \operatorname{dom} F$

$$\delta_{b,d} = \inf\left\{\|(\eta_K^u, Cu)\| \mid K \in \mathcal{I}_{b,d},\; u \in S_X,\; \Phi_X(u) \cap (C_K + \operatorname{Im} C^*) \ne \emptyset\right\}, \tag{3.2}$$

where $\mathcal{I}_{b,d} := \{I_b(x) \mid x \in F(b,d)\}$*, and*

$$\delta = \inf\left\{\|(\eta_K^u, Cu)\| \mid K \in \mathcal{P}(I),\; u \in S_X,\; \Phi_X(u) \cap (C_K + \operatorname{Im} C^*) \ne \emptyset\right\}, \tag{3.3}$$

both infima being attained when X *is finite-dimensional.*

Proof. Let $(b,d) \in \operatorname{dom} F$ be fixed. Consider the function f defined by (2.3); by Lemma 2.2(i), f is convex. Using relation (2.2) in Proposition 2.1, we obtain that

$$\begin{aligned} \delta_{b,d} &= \inf\left\{f'(x,u) \mid x \in F(b,d),\; u \in S_X \cap \Phi_X^{-1}\big(N(F(b,d),x)\big)\right\} \\ &= \inf\left\{f'(x,u) \mid x \in F(b,d),\; u \in S_X \cap \Phi_X^{-1}\left(C_{I_b(x)} + \operatorname{Im} C^*\right)\right\}. \end{aligned}$$

But for $x \in F(b,d)$ and $u \in X$ we have that $f'(x,u) = \lim_{t\to 0^+} t^{-1}\|((A(x+tu)-b)_+, tCu)\|$. For $i \in I_b(x)$ and $t > 0$ we have that $(\langle x+tu, a_i\rangle - b_i)_+ = t\,(\langle u, a_i\rangle)_+ =$

$t\big(\eta^u_{I_b(x)}\big)_i$. Because $\langle x, a_i\rangle - b_i < 0$ for $i \in I \setminus I_b(x)$, there exists $\varepsilon > 0$ such that $\langle x + tu, a_i\rangle - b_i < 0$ for all $i \in I \setminus I_b(x)$ and $t \in]0, \varepsilon]$. Hence, for such i and t we have that $(\langle x + tu, a_i\rangle - b_i)_+ = 0 = t\big(\eta^u_{I_b(x)}\big)_i$. It follows that $\big(A(x + tu) - b\big)_+ = t \cdot \eta^u_{I_b(x)}$ for $t \in]0, \varepsilon]$, whence $f'(x, u) = \big\|(\eta^u_{I_b(x)}, Cu)\big\|$. From the above expression of $\delta_{b,d}$ we obtain (3.2).

Taking into consideration that (3.2) holds for every $(b, d) \in \operatorname{dom} F$, the inequality $\geq$ holds in (3.3). Consider $K \in \mathcal{P}(I)$ and $u \in S_X$ such that $\Phi_X(u) \cap (C_K + \operatorname{Im} C^*) \neq \emptyset$. Fix an $\overline{x} \in X$ and take $b_i := \langle \overline{x}, a_i\rangle$ for $i \in K$, $b_i := \langle \overline{x}, a_i\rangle + 1$ for $i \in I \setminus K$, and $d := C\overline{x}$. It is obvious that $K = I_b(\overline{x})$ for $b := (b_1, \ldots, b_m)$. Then $\|(\eta^u_K, Cu)\| \geq \delta_{b,d} \geq \delta$. Therefore (3.3) holds.

Now assume $\dim X < \infty$. Let $(b, d) \in \operatorname{dom} F$ be fixed and consider $\big(\big\|(\eta^{u_n}_{K_n}, Cu_n)\big\|\big) \to \delta_{b,d}$ with $K_n \in \mathcal{I}_{b,d}$ and $u_n \in S_X$ such that $\Phi_X(u_n) \cap (C_{K_n} + \operatorname{Im} C^*) \neq \emptyset$ for every n. Since $\mathcal{I}_{b,d}$ is finite and $\dim X < \infty$ we may assume that $K_n = K$ for every n and $(u_n) \to u \in S_X$. Because the graph of Φ_X is closed and $C_K + \operatorname{Im} C^*$ is also closed we have that $\Phi_X(u) \cap (C_K + \operatorname{Im} C^*) \neq \emptyset$. The definition of $\eta^{u_n}_K$ shows that $(\eta^{u_n}_K) \to \eta^u_K$, and so $\delta_{b,d} = \|(\eta^u_K, Cu)\|$. Hence the infimum in (3.2) is attained. A similar argument shows that the infimum in (3.3) is also attained. □

The formula (3.3) was stated by Belousov and Andronov [2] for $l = 0$, for $X = \mathbb{R}^k$ endowed with the Euclidean norm, and for the norm on $\mathbb{R}^m$ replaced by a pseudonorm verifying condition (1.6).

4. Extensions of Ng–Zheng and Azé–Corvellec formulae. In this section we are interested in estimates or formulae for $\delta_{b,d}$ of types similar to those established by Ng and Zheng [11] or Azé and Corvellec [1]. Taking into consideration Corollary 2.5, these estimates are related to those of Bergthaller and Singer [3]. Using formula (2.1) for the function f defined by relation (2.3) we obtain the following result, where, as above, $\mathcal{A}x = (Ax, Cx)$.

THEOREM 4.1. *Assume that the norm $\|\cdot\|$ on $\mathbb{R}^{m+l}$ satisfies conditions* (2.5) *and* (2.6). *Then for every $(b, d) \in \operatorname{dom} F$ one has*

(4.1)
$$\delta_{b,d} = \inf\{\|x^*\|_* \mid \exists\, x \in X \ :\ ((Ax - b)_+, Cx - d) \neq 0,\ x^* \in \mathcal{A}^*\,(\partial h(\mathcal{A}x - (b, d)))\}$$

(4.2)
$$\begin{aligned} &= \inf\{\|A^*\mu + C^*\zeta\|_* \mid \exists\, x \in X \ :\ ((Ax - b)_+, Cx - d) \neq 0,\\ &\qquad\qquad (\mu, \zeta) \in \partial h(Ax - b, Cx - d)\}\\ &= \inf\big\{\, \|A^*\mu + C^*\zeta\|_* \mid x \in X,\ (\mu, \zeta) \in \mathbb{R}^m_+ \times \mathbb{R}^l,\ \|(\mu, \zeta)\|_* = 1,\\ &\qquad\qquad \langle Ax - b, \mu\rangle + \langle Cx - d, \zeta\rangle = \|((Ax - b)_+, Cx - d)\| > 0\big\}. \end{aligned}$$

Moreover, if X is a reflexive Banach space, then

$$\begin{aligned} \delta_{b,d} = \inf\{&\|A^*\mu + C^*\zeta\|_* \mid x \in X,\ (\mu, \zeta) \in \mathbb{R}^m_+ \times \mathbb{R}^l,\ \langle (Ax - b)_-, \mu\rangle = 0,\\ &\langle Ax - b, \mu\rangle + \langle Cx - d, \zeta\rangle = \|((Ax - b)_+, Cx - d)\| > 0\}. \end{aligned} \tag{4.3}$$

Proof. We apply Proposition 2.1 for f defined by (2.3) and $t = 0$. Then $f(x) = h(\mathcal{A}x - (b, d))$ for every $x \in X$. Since h is a continuous sublinear function by Lemma 2.2(i), we have that f is a continuous convex function, and so $\partial f(x) = \mathcal{A}^*\big(\partial h(\mathcal{A}x - (b, d))\big)$ for $x \in X$. But $\mathcal{A}^*(\mu, \zeta) = A^*\mu + C^*\zeta$, and so the first part of the conclusion follows applying again Lemma 2.2.

Since for $x \in X$ and $(\mu,\zeta) \in \mathbb{R}^m_+ \times \mathbb{R}^l$ with $\langle Ax-b,\mu\rangle + \langle Cx-d,\zeta\rangle = \|((Ax-b)_+, Cx-d)\|$ and $\|(\mu,\zeta)\|_* = 1$ we have that $\langle (Ax-b)_+,\mu\rangle + \langle Cx-d,\zeta\rangle = \langle Ax-b,\mu\rangle + \langle Cx-d,\zeta\rangle$ and $\langle (Ax-b)_-,\mu\rangle = 0$, the inequality $\geq$ in (4.3) is obvious.

Assume that X is reflexive. Let $x \in X$ and $(\mu,\zeta) \in \mathbb{R}^m_+ \times \mathbb{R}^l$ be such that $\langle (Ax-b)_+,\mu\rangle + \langle Cx-d,\zeta\rangle = \langle Ax-b,\mu\rangle + \langle Cx-d,\zeta\rangle = \|((Ax-b)_+, Cx-d)\| > 0$. Of course, $x \notin F(b,d)$. Consider $\overline{x} \in F(b,d)$ such that $d(x,F(b,d)) = \|x-\overline{x}\|$. Because $A\overline{x} \leq b$ and $\mu \geq 0$, we have that

$$\begin{aligned}\|((Ax-b)_+, Cx-d)\| &= \langle Ax-b,\mu\rangle + \langle Cx-d,\zeta\rangle \leq \langle Ax-A\overline{x},\mu\rangle + \langle Cx-C\overline{x},\zeta\rangle \\ &= \langle x-\overline{x}, A^*\mu + C^*\zeta\rangle \leq \|x-\overline{x}\| \cdot \|A^*\mu + C^*\zeta\|,\end{aligned}$$

and so $\delta_{b,d} \leq \|A^*\mu + C^*\zeta\|$. The conclusion follows. □

In order to obtain other formulae or estimates for $\delta_{b,d}$ let us introduce other notation. We shall deal with pairs (K,L) and triples (K,L^+,L^-) of sets with $K \subset I$ and $L, L^+, L^- \subset J$; we assume always that $L^+ \cap L^- = \emptyset$. By $(K,L) \subset (K',L')$ and $(K,L^+,L^-) \subset (K',L'^+,L'^-)$ we mean $K \subset K'$, $L \subset L'$ and $K \subset K'$, $L^+ \subset L'^+$, $L^- \subset L'^-$, respectively. For such pairs and triples we consider the compact sets

(4.4)
$$M_{K,L} := \left\{(\mu,\zeta) \in \mathbb{R}^m_+ \times \mathbb{R}^l \mid \|(\mu,\zeta)\|_* = 1,\ i \in I \setminus K \Rightarrow \mu_i = 0,\ j \in J \setminus L \Rightarrow \zeta_j = 0\right\},$$

$$M_{K,L^+,L^-} := \left\{(\mu,\zeta) \in M_{K,L^+\cup L^-} \mid j \in L^+ \Rightarrow \zeta_j \geq 0,\ j \in L^- \Rightarrow \zeta_j \leq 0\right\} \tag{4.5}$$

and the numbers

$$\tau_{K,L} := \min\left\{\|A^*\mu + C^*\zeta\|_* \mid (\mu,\zeta) \in M_{K,L}\right\}, \tag{4.6}$$

$$\tau_{K,L^+,L^-} := \min\left\{\|A^*\mu + C^*\zeta\|_* \mid (\mu,\zeta) \in M_{K,L^+,L^-}\right\}. \tag{4.7}$$

It is obvious that $\tau_{K,L} \geq \tau_{K',L'}$ if $(K,L) \subset (K',L')$ and $\tau_{K,L^+,L^-} \geq \tau_{K',L'^+,L'^-}$ if $(K,L^+,L^-) \subset (K',L'^+,L'^-)$. Because $M_{K,L} = \bigcup\left\{M_{K,L^+,L^-} \mid L = L^+ \cup L^-\right\}$,

$$\tau_{K,L} = \min\left\{\tau_{K,L^+,L^-} \mid L = L^+ \cup L^-\right\}. \tag{4.8}$$

Inspired by the notions introduced by Ng and Zheng in [11], we consider the classes of *regular pairs* and *regular triples*

$$\begin{aligned}\mathcal{R}(I,J) &:= \{(K,L) \mid (K,L) \subset (I,J),\ K \cup L \neq \emptyset,\ \tau_{K,L} > 0\},\\ \mathcal{R}'(I,J) &:= \{(K,L^+,L^-) \mid (K,L^+\cup L^-) \subset (I,J),\ K \cup L^+ \cup L^- \neq \emptyset,\ \tau_{K,L^+,L^-} > 0\}.\end{aligned}$$

Since $\mathcal{R}(I,J)$ and $\mathcal{R}'(I,J)$ are nonempty (as $\mathcal{A} \neq 0$) and finite, we have that

$$\rho := \min\{\tau_{K,L} \mid (K,L) \in \mathcal{R}(I,J)\} \in]0,\infty[, \tag{4.9}$$

$$\rho' := \min\{\tau_{K,L^+,L^-} \mid (K,L^+,L^-) \in \mathcal{R}'(I,J)\} \in]0,\infty[. \tag{4.10}$$

Taking into consideration (4.8) we have that $\rho \geq \rho'$. We shall see below that $\rho = \rho'$, but this is not obvious because we could have $(K,L^+,L^-) \in \mathcal{R}'(I,J)$ with $(K,L^+\cup L^-) \notin \mathcal{R}(I,J)$.

We say that the pair $(K,L) \subset (I,J)$ is *full* if $\operatorname{lin}(\{a_i \mid i \in K\} \cup \{c_j \mid j \in L\}) = \operatorname{Im}\mathcal{A}^*$; we denote by $\mathcal{R}_f(I,J)$ the class of full regular pairs. Similarly, the triple (K,L^+,L^-) is *full* if $(K,L^+\cup L^-)$ is full, and we denote by $\mathcal{R}'_f(I,J)$ the class of

full regular triples. The notation $\mathcal{R}(I)$, $\mathcal{R}(J)$, $\mathcal{R}_f(I)$, $\mathcal{R}_f(J)$ is now self-explanatory. Consider also

$$\mathcal{L}(I,J) := \big\{(K,L)\,|\,(K,L)\subset(I,J), \{a_i\,|\,i\in K\}\cup\{c_j\,|\,j\in L\} \text{ is linearly independent}\big\},$$
$$\mathcal{L}'(I,J) := \{(K,L^+,L^-)\mid (K,L^+\cup L^-)\in\mathcal{L}(I,J)\},$$

and similarly $\mathcal{L}(I)$, $\mathcal{L}(J)$, $\mathcal{L}_f(I,J)$, $\mathcal{L}'_f(I,J)$, $\mathcal{L}_f(I)$, $\mathcal{L}_f(J)$; so $K\in\mathcal{L}_f(I)$ if and only if $\{a_i\mid i\in K\}$ is a basis for $\operatorname{Im}A^*$. It is obvious that $\tau_{K,L}>0$ if and only if $L\in\mathcal{L}(J)$, $\tau_{K,\emptyset}>0$, and $C_K\cap\operatorname{lin}\{c_j\mid j\in L\}=\{0\}$. In particular $\mathcal{L}(I,J)\subset\mathcal{R}(I,J)$.

LEMMA 4.2. *For every $(K,L)\in\mathcal{R}(I,J)$ there exists $(K',L')\in\mathcal{R}_f(I,J)$ such that $(K,L)\subset(K',L')$; similarly, for any $(K,L)\in\mathcal{L}(I,J)$ there exists $(K',L')\in\mathcal{L}_f(I,J)$ such that $(K,L)\subset(K',L')$, and for every $(K,L^+,L^-)\in\mathcal{R}'(I,J)$ there exists $(K,L'^+,L'^-)\in\mathcal{R}'_f(I,J)$ such that $(K,L^+,L^-)\subset(K,L'^+,L'^-)$. In particular, any maximal regular pair and any maximal regular triple (with respect to inclusion) is full.*

Proof. Indeed, let $(K,L)\in\mathcal{R}(I,J)$. Assume that $Z_0:=\operatorname{lin}(\{a_i\mid i\in K\}\cup\{c_j\mid j\in L\})\neq\operatorname{Im}\mathcal{A}^*$. Then there exists $i'\in I\setminus K$ such that $a_{i'}\notin Z_0$ or $j'\in J\setminus K$ such that $c_{j'}\notin Z_0$. Consider the first case, the second one being treated similarly. It follows that $(K',L)\in\mathcal{R}(I,J)$, where $K':=K\cup\{i'\}$; otherwise $0=\sum_{i\in I}\mu_i a_i+\sum_{j\in J}\zeta_j c_j$ for some $(\mu,\zeta)\in M_{K',L}$. If $\mu_{i'}=0$, we get the contradiction $0=\tau_{K,L}$, because $(\mu,\zeta)\in M_{K,L}$ in this case; if $\mu_{i'}\neq 0$, we get the contradiction $a_{i'}\in Z_0$. Thus, there exists $(K_1,L_1)\in\mathcal{R}(I,J)$ such that $(K,L)\subset(K_1,L_1)$ and $Z_0\subset Z_1:=\operatorname{lin}(\{a_i\mid i\in K_1\}\cup\{c_j\mid j\in L_1\})$ with $Z_0\neq Z_1$. Continuing in this way we obtain a pair $(K_h,L_h)\in\mathcal{R}_f(I,J)$ with $(K,L)\subset(K_h,L_h)$ in a finite number of steps. The proof for triples is similar.

When $(K,L)\in\mathcal{L}(I,J)$, just complete $\{a_i\mid i\in K\}\cup\{c_j\mid j\in L\}$ to a basis of Y with elements of $\{a_i\mid i\in I\}\cup\{c_j\mid j\in J\}$. ∎

Another result in the same spirit, but with a more involved proof, is the following.

LEMMA 4.3. *Let $(K,L)\in\mathcal{R}(I,J)$. Then there exists $K_0\subset K$ such that $(K_0,L)\in\mathcal{L}(I,J)$ and $\tau_{K_0,L}=\tau_{K,L}$. Similarly, if $(K,L^+,L^-)\in\mathcal{R}'(I,J)$, then there exists $(K_0,L_0^+,L_0^-)\subset(K,L^+,L^-)$ such that $(K_0,L_0^+,L_0^-)\in\mathcal{L}'(I,J)$ and $\tau_{K_0,L_0^+,L_0^-}=\tau_{K,L^+,L^-}$. In particular, if $0\notin\operatorname{co}\{a_i\mid i\in K\}$ for some $\emptyset\neq K\subset I$, then there exists $K_0\in\mathcal{L}(K)$ such that $d\,(0,\operatorname{co}\{a_i\mid i\in K\})=d\,(0,\operatorname{co}\{a_i\mid i\in K_0\})$.*

Proof. Let $(K,L)\in\mathcal{R}(I,J)$; as observed above, $L\in\mathcal{L}(J)$. Consider

$$\mathcal{I}:=\{K'\subset K\mid\tau_{K',L}=\tau_{K,L}\}$$

and take $K_0\in\mathcal{I}$ such that $\operatorname{card}K_0\leq\operatorname{card}K'$ for every $K'\in\mathcal{I}$. If $K_0=\emptyset$, then $(K_0,L)\in\mathcal{L}(I,J)$.

Let $K_0\neq\emptyset$ and take $(\eta,\xi)\in M_{K_0,L}$ such that $\tau_{K_0,L}=\|\overline{x}^*\|$ where $\overline{x}^*=A^*\eta+C^*\xi$. By the choice of K_0 we have that $\eta_i>0$ for every $i\in K_0$. Assuming that $(K_0,L)\notin\mathcal{L}(I,J)$, there exists $(\lambda,\nu)\in\mathbb{R}^{m+l}\setminus\{(0,0)\}$ such that $0=A^*\lambda+C^*\nu$, $\lambda_i=0$ for $i\in I\setminus K_0$, and $\nu_j=0$ for $j\in J\setminus L$. There exists some $t\in\mathbb{R}$ such that $\eta_i+t\lambda_i\geq 0$ for all $i\in K_0$ and $\eta_{i_0}+t\lambda_{i_0}=0$ for some $i_0\in K_0$. Let $K_0':=K_0\setminus\{i_0\}\subset K_0$. On the one hand we have that $\tau_{K_0',L}\geq\tau_{K_0,L}$ and $(\eta',\xi'):=(\eta,\xi)+t(\lambda,\nu)\in M_{K_0',L}$. On the other hand $\overline{x}^*=A^*(\eta+t\lambda)+C^*(\xi+t\nu)=A^*\eta'+C^*\xi'$, and so $\tau_{K,L}=\|\overline{x}^*\|\geq\tau_{K_0',L}$. Hence $(K_0',L)\in\mathcal{I}$, contradicting the choice of K_0. Therefore $(K_0,L)\in\mathcal{L}(I,J)$.

The proof for the case when $(K,L^+,L^-)\in\mathcal{R}'(I,J)$ is similar. Consider

$$\mathcal{IJ}:=\{(K',L'^+,L'^-)\mid(K',L'^+,L'^-)\subset(K,L^+,L^-)\ :\ \tau_{K',L'^+,L'^-}=\tau_{K,L^+,L^-}\}$$

and take $(K_0, L_0^+, L_0^-) \in \mathcal{IJ}$ such that $\operatorname{card}(K_0 \cup L_0^+ \cup L_0^-) \leq \operatorname{card}(K' \cup L'^+ \cup L'^-)$ for every $(K', L'^+, L'^-) \in \mathcal{IJ}$. Let $(\eta, \xi) \in M_{K_0, L_0^+, L_0^-}$ be such that $\tau_{K_0, L_0^+, L_0^-} = \|\overline{x}^*\|$, where $\overline{x}^* = A^*\eta + C^*\xi$. Then $\eta_i > 0$ for $i \in K_0$, $\zeta_j > 0$ for $j \in L_0^+$, and $\zeta_j < 0$ for $j \in L_0^-$. As above, assuming that $(K_0, L_0^+, L_0^-) \notin \mathcal{L}'(I, J)$ we get a contradiction with the choice of (K_0, L_0^+, L_0^-).

Taking $l = 0$ and $\|\cdot\| = \|\cdot\|_\infty$ on $\mathbb{R}^m$ one obtains the last conclusion. □

The last part of the above lemma was obtained by Azé and Corvellec [1, Lem. 3.1]. From Lemmas 4.2 and 4.3 and relation (4.8) applied for $(K, L) \in \mathcal{L}(I, J)$ we get immediately the following corollary.

COROLLARY 4.4. *The following relations hold:*

$$\begin{aligned}
\rho &= \min\{\tau_{K,L} \mid (K, L) \text{ maximal in } \mathcal{R}(I, J)\} \\
&= \min\{\tau_{K,L} \mid (K, L) \in \mathcal{R}_f(I, J)\} \\
&= \min\{\tau_{K,L} \mid (K, L) \in \mathcal{L}(I, J)\} \\
&= \min\{\tau_{K,L} \mid (K, L) \in \mathcal{L}_f(I, J)\}, \\
\rho' &= \min\{\tau_{K,L^+,L^-} \mid (K, L^+, L^-) \text{ maximal in } \mathcal{R}'(I, J)\} \\
&= \min\{\tau_{K,L^+,L^-} \mid (K, L^+, L^-) \in \mathcal{R}'_f(I, J)\} \\
&= \min\{\tau_{K,L^+,L^-} \mid (K, L^+, L^-) \in \mathcal{L}'(I, J)\} \\
&= \min\{\tau_{K,L^+,L^-} \mid (K, L^+, L^-) \in \mathcal{L}'_f(I, J)\},
\end{aligned}$$

and $\rho = \rho'$.

When the norm on $\mathbb{R}^{m+l}$ is $\|\cdot\|_\infty$ we can also extend the notion of peak set introduced by Ng and Zheng [11]. So, we say that (K, L^+, L^-) is a *peak triple* at $(b, d) \in \operatorname{dom} F$ if there exists $x \in X$ such that $s := f(x) = \|((Ax - b)_+, Cx - d)\|_\infty > 0$ and $\langle x, a_i\rangle - b_i = s$ for $i \in K$, $\langle x, c_j\rangle - d_j = s$ for $j \in L^+$, $\langle x, c_j\rangle - d_j = -s$ for $j \in L^-$, $\langle x, a_i\rangle - b_i < s$ for $i \in I \setminus K$, $|\langle x, c_j\rangle - d_j| < s$ for $j \in J \setminus (L^+ \cup L^-)$. Of course, $K \cup L^+ \cup L^- \neq \emptyset$ if (K, L^+, L^-) is a peak triple. Because $f(x) > 0 = \inf f$, we have that $0 \notin \partial f(x)$; since $\partial f(x) = \{A^*\mu + C^*\zeta \mid (\mu, \zeta) \in M_{K,L^+,L^-}\}$ (see relation (2.12)), we have that $(K, L^+, L^-) \in \mathcal{R}'(I, J)$ in this case. Denote by $\mathcal{P}_{b,d}(I, J)$ the class of peak triples and by $\mathcal{F}_{b,d}(I, J)$ (resp., $\mathcal{M}_{b,d}(I, J)$) the class of full (resp., maximal) peak triples at (b, d).

THEOREM 4.5. *Let* $\mathbb{R}^{m+l}$ *be endowed with the box norm* $\|\cdot\|_\infty$ *and* $(b, d) \in \operatorname{dom} F$. *Then*

$$\delta_{b,d} = \min\{\tau_{K,L^+,L^-} \mid (K, L^+, L^-) \in \mathcal{M}_{b,d}(I, J)\} \tag{4.11}$$

$$= \min\{\tau_{K,L^+,L^-} \mid (K, L^+, L^-) \in \mathcal{F}_{b,d}(I, J)\} \tag{4.12}$$

$$= \min\{\tau_{K,L^+,L^-} \mid \mathcal{L}'(I, J) \ni (K, L^+, L^-) \subset (K', L'^+, L'^-) \in \mathcal{P}_{b,d}(I, J)\}. \tag{4.13}$$

Proof. By relation (4.2) of Theorem 4.1 and relation (2.12) we have $\delta_{b,d} = \min\{\tau_{K,L^+,L^-} \mid (K, L^+, L^-) \in \mathcal{P}_{b,d}(I, J)\}$, and so $\delta_{b,d}$ is less than or equal to the two quantities appearing in (4.11) and (4.12). Let us show that for any peak triple (K, L^+, L^-) there exists $(K', L'^+, L'^-) \in \mathcal{F}_{b,d}(I, J)$ such that $(K, L^+, L^-) \subset (K', L'^+, L'^-)$. Fix $(K, L^+, L^-) \in \mathcal{P}_{b,d}(I, J)$ and consider $(K_0, L_0^+, L_0^-) \in \mathcal{P}_{b,d}(I, J)$ such that $(K, L^+, L^-) \subset (K_0, L_0^+, L_0^-)$ and $\operatorname{card}(K_0 \cup L_0^+ \cup L_0^-) \geq \operatorname{card}(K' \cup L'^+ \cup L'^-)$ for any $(K', L'^+, L'^-) \in \mathcal{P}_{b,d}(I, J)$ with $(K, L^+, L^-) \subset (K', L'^+, L'^-)$; the existence of (K_0, L_0^+, L_0^-) is ensured by the finiteness of the class of peak triples. It is obvious that (K_0, L_0^+, L_0^-) is in $\mathcal{M}_{b,d}(I, J)$, and so equality holds in (4.11). Assume that $(K_0, L_0^+, L_0^-) \notin \mathcal{F}_{b,d}(I, J)$, and so $Z_0 := \operatorname{lin}\big(\{a_i \mid i \in K_0\} \cup \{c_j \mid j \in L_0^+ \cup L_0^-\}\big) \neq$

$\operatorname{Im}\mathcal{A}^*$. Hence there exists $i' \in I \setminus K_0$ or $j' \in J \setminus (L_0^+ \cup L_0^-)$ such that $a_{i'} \notin Z_0$ or $c_{j'} \notin Z_0$. It follows that there exists $x' \in X$ such that $\langle x', a_i\rangle = 0$ for $i \in K_0$, $\langle x', c_j\rangle = 0$ for $j \in L_0^+ \cup L_0^-$, and $\langle x', a_{i'}\rangle = 1$ or $\langle x', c_{j'}\rangle = 1$. Taking $x_0 \in X$, which corresponds to (K_0, L_0^+, L_0^-), we have that $\langle x_0 + tx', a_i\rangle - b_i = s_0 := f(x_0)$ for $i \in K_0$, $\langle x_0 + tx', c_j\rangle - d_j = s_0$ for $j \in L_0^+$, and $\langle x_0 + tx', c_j\rangle - d_j = -s_0$ for $j \in L_0^-$, for every $t \in \mathbb{R}$. Moreover, $\langle x_0, a_i\rangle - b_i < s_0$ for $i \in I \setminus K_0$, and $|\langle x_0, c_j\rangle - d_j| < s_0$ for $j \in J \setminus L_0$. Take the greatest $t_0 > 0$ such that $\langle x_0 + tx', a_i\rangle - b_i \le s_0$ for $i \in I \setminus K_0$, and $|\langle x_0 + tx', c_j\rangle - d_j| \le s_0$ for $j \in J \setminus (L_0^+ \cup L_0^-)$. Then at least one of these inequalities becomes an equality. Let $x_1 := x_0 + t_0 x'$; then $s_1 := f(x_1) = f(x_0)$, $K_1 := \{i \in I \mid \langle x_1, a_i\rangle - b_i = s_1\} \supset K_0$, $L_1^+ := \{j \in J \mid \langle x, c_j\rangle - d_j = s_1\} \supset L_0^+$, $L_1^- := \{j \in J \mid \langle x, c_j\rangle - d_j = -s_1\} \supset L_0^-$, and at least one of these inclusions is strict. Because, obviously, (K_1, L_1^+, L_1^-) is a peak triple, we have a contradiction. Therefore $(K_0, L_0^+, L_0^-) \in \mathcal{F}_{b,d}(I, J)$. Hence the equality holds in (4.12), too.

For (4.13) just use Lemma 4.3. □

The proof above shows that $\mathcal{M}_{b,d}(I, J) \subset \mathcal{F}_{b,d}(I, J) \subset \mathcal{P}_{b,d}(I, J)$. When $l = 0$ the sets J, L, L^+, L^- are empty, and so we omit them in the above notation. So we write M_K, τ_K, $\mathcal{P}_b(I)$, $\mathcal{F}_b(I)$, and $\mathcal{M}_b(I)$ instead of $M_{K,L}, \dots, \mathcal{M}_{b,d}(I, J)$; we also write δ_b instead of $\delta_{b,d}$. In this case ($l = 0$) relation (4.13) is proved by Azé and Corvellec in [1, Thm. 3.1], while (4.12) strengthens Theorem 4.4 in [11], where it is shown that $\delta_b \ge \min\{\tau_K \mid K \in \mathcal{F}_b(I)\}$.

Unfortunately, for an arbitrary norm $\|\cdot\|$ on $\mathbb{R}^{m+l}$ (however, satisfying conditions (2.5) and (2.6)) we have only the formulae for $\delta_{b,d}$ which are provided by Theorem 4.1. In the next result we provide an estimate for $\delta_{b,d}$ in the general case. This estimate practically follows from (the proof of) Li's Theorem 3.4 in [8].

PROPOSITION 4.6. *Assume that X is a reflexive Banach space. Then*

$$\delta_{b,d} \ge \max_{L \in \mathcal{L}_f(J)} \alpha_{b,d}^L, \tag{4.14}$$

where

$$\alpha_{b,d}^L := \min\{\tau_{K,L} \mid x \in F(b,d),\ K \subset I_b(x),\ (K, L) \in \mathcal{L}(I, J)\}.$$

Proof. We use Corollary 2.5. Fix $L \in \mathcal{L}_f(J)$. We consider $(b', d') \in \operatorname{dom} F$ and $x' \in F(b', d') \setminus F(b, d)$. Let $x \in F(b, d)$ be such that $\|x' - x\| = d(x', F(b, d))$. Then $\Phi_X(x' - x) \cap N(F(b, d), x) \ne \emptyset$; let x^* be an element of this set. Using formula (3.1), the set

$$M_{x^*} := \big\{(\eta, \xi) \in \mathbb{R}^m_+ \times \mathbb{R}^l \mid x^* = A^*\eta + C^*\xi,\ \eta_i > 0 \Rightarrow i \in I_b(x)\big\}$$

is nonempty. Of course $(\eta, \xi) \ne (0, 0)$ for $(\eta, \xi) \in M_{x^*}$; this is due to the fact that $x' \ne x$, and so $x^* \ne 0$. Proceeding as in the proof of Lemma 3.1 in [8], there exist $(\mu, \zeta) \in M_{x^*}$ and $K \subset I_b(x)$ such that $(K, L) \in \mathcal{L}(I, J)$ and $\mu_i = 0$ for $i \in I \setminus K$, $\zeta_j = 0$ for $j \in J \setminus L$. Indeed, for $(\eta, \xi) \in M_{x^*}$ let $K_{\eta,\xi} := \{i \in I \mid \eta_i > 0\}$. Of course, $K_{\eta,\xi} \subset I_b(x)$. Let $(\mu, \zeta) \in M_{x^*}$ be such that $\operatorname{card} K \le \operatorname{card} K_{\eta,\xi}$ for every $(\eta, \xi) \in M_{x^*}$, where $K := K_{\mu,\zeta}$. Let $\lambda \in \mathbb{R}^m$ be such that $\lambda_i \ne 0 \Rightarrow i \in K$ and $u^* := A^*\lambda \in \operatorname{Im} C^*$, i.e., $u^* = C^*\nu$ for some $\nu \in \mathbb{R}^l$. Assume that $\lambda_{i_0} \ne 0$ for some $i_0 \in K$; we (may) even assume that $\lambda_{i_0} > 0$. Taking $\bar{t} := \min\{\lambda_i^{-1}\mu_i \mid \lambda_i > 0\}$, we have that $(\mu', \zeta') := (\mu, \zeta) - \bar{t}(\lambda, -\nu) \in M_{x^*}$ and $\operatorname{card} K_{\mu',\zeta'} < \operatorname{card} K$. Hence $K \in \mathcal{L}(I)$ and $\operatorname{lin}\{a_i \mid i \in K\} \cap \operatorname{Im} C^* = \{0\}$. Since $x^* - A^*\mu \in \operatorname{Im} C^*$ and $L \in \mathcal{L}_f(J)$ we may assume that $\zeta_j \ne 0 \Rightarrow j \in L$. Then (μ, ζ) is the desired element of M_{x^*}.

It follows that $(\overline{\mu},\overline{\zeta}) := \|(\mu,\zeta)\|_*^{-1}(\mu,\zeta) \in M_{K,L}$. Because $x^* \in \Phi_X(x'-x)$ and $\langle Ax,\mu\rangle = \langle b,\mu\rangle$, we have that

$$\begin{aligned}\|x^*\|_* \cdot \|x'-x\| &= \langle x'-x,x^*\rangle = \langle Ax'-Ax,\mu\rangle + \langle Cx'-Cx,\zeta\rangle \le \langle b'-b,\mu\rangle + \langle d'-d,\zeta\rangle \\ &\le \|(b',d')-(b,d)\| \cdot \|(\mu,\zeta)\|_*,\end{aligned}$$

whence

$$\begin{aligned}\alpha_{b,d}^L \cdot d\left(x',F(b,d)\right) &\le \left\|A^*\overline{\mu}+C^*\overline{\zeta}\right\|_* \cdot \|x'-x\| = \|(\mu,\zeta)\|_*^{-1} \cdot \|x^*\|_* \cdot \|x'-x\| \\ &\le \|(b',d')-(b,d)\|.\end{aligned}$$

Hence

$$\alpha_{b,d}^L \cdot e\left(F(b',d'),F(b,d)\right) \le \|(b',d')-(b,d)\| \quad \forall\,(b',d') \in \operatorname{dom} F. \tag{4.15}$$

The conclusion follows using Corollary 2.5. □

When $l=0$, the norm on $\mathbb{R}^m$ is $\|\cdot\|_\infty$, and $b \in \operatorname{dom} F$ (we omit the second component in this case), then the constant C defined in relation (1.31) of Bergthaller and Singer's paper [3] is nothing else but α_b^{-1}; we omit $d\,(=0)$ and $L\,(=J=\emptyset)$ in this case. As proven by Azé and Corvellec in their Example 3.1 from [1] one can have strict inequality in (4.14). The inequality may be strict also in the case $m=0$. Consider $X:=\mathbb{R}$, $l=2$, and $C:\mathbb{R}\to\mathbb{R}^2$ defined by $Cx:=(x,x)$. Consider the Euclidean norm on $\mathbb{R}$ and $\mathbb{R}^2$. Then $\delta_{0,0}=\min\{\|Cx\| \mid |x|=1\}=\sqrt{2}$, while $\alpha_{0,0}^{\{1\}}=\alpha_{0,0}^{\{2\}}=1$.

5. Estimates for the global Hoffman constant. In this section we are interested in formulae and/or estimates for the Hoffman constant δ of F. The first result is based on Proposition 4.6.

PROPOSITION 5.1. *Assume that X is a reflexive Banach space and the norm $\|\cdot\|$ on $\mathbb{R}^{m+l}$ satisfies the conditions* (2.5) *and* (2.6). *Then*

$$\delta \ge \max_{L\in\mathcal{L}_f(J)} \min_{(K,L)\in\mathcal{L}(I,J)} \tau_{K,L}. \tag{5.1}$$

Moreover, if the norm $\|\cdot\|$ on $\mathbb{R}^{m+l}$ satisfies condition (2.10) *and $\{c_j \mid j\in J\}$ is linearly independent, then*

$$\delta = \min\{\tau_{K,J} \mid (K,J)\in\mathcal{L}(I,J)\}.$$

Proof. Let $L\in\mathcal{L}_f(J)$ and $(b,d)\in\operatorname{dom} F$. From the expression of $\alpha_{b,d}^L$ defined in Proposition 4.6, we have that $\alpha_{b,d}^L \ge \alpha^L := \min\{\tau_{K,L} \mid (K,L)\in\mathcal{L}(I,J)\}$. Using Proposition 4.6, we have that $\delta=\inf\{\delta_{b,d} \mid (b,d)\in\operatorname{dom} F\} \ge \max\{\alpha^L \mid L\in\mathcal{L}_f(J)\}$. Hence (5.1) holds.

Assume now that the norm $\|\cdot\|$ on $\mathbb{R}^{m+l}$ satisfies condition (2.10) and $\{c_j \mid j\in J\}$ is linearly independent. Take $K\subset I$ such that $(K,J)\in\mathcal{L}(I,J)$ and $(\mu,\zeta)\in M_{K,J}$. Applying Lemma 2.3(iii) for the duality mapping on $(\mathbb{R}^{m+l},\|\cdot\|_*)$, there exists $(u,v)\in\mathbb{R}_+^m\times\mathbb{R}^l$ such that $\|(u,v)\|=\langle u,\mu\rangle+\langle v,\zeta\rangle=1$ and $u_i=0$ whenever $\mu_i=0$. Since $\{a_i \mid i\in K\}\cup\{c_j \mid j\in J\}$ is linearly independent, there exists $x\in X$ such that $\langle x,a_i\rangle=u_i$ for every $i\in K$ and $\langle x,c_j\rangle=v_j$ for every $j\in J$. Consider $b_i:=0$ for $i\in K$ and $b_i:=\max\{\langle x,a_i\rangle+1,0\}$ for $i\in I\setminus K$. Then $b:=(b_1,\dots,b_m)\in\mathbb{R}_+^m$ and $((Ax-b)_+,Cx)=(u,v)\ne(0,0)$. Since $\langle Ax-b,\mu\rangle+\langle Cx,\zeta\rangle=\|((Ax-b)_+,Cx)\|>0$, by Theorem 4.1, we have that $\delta\le\delta_{b,0}\le\|A^*\mu+C^*\zeta\|_*$, and so $\delta\le\tau_{K,J}$. The conclusion follows. □

In fact the condition that X is a reflexive Banach space in the statement of the preceding proposition is not essential because, as mentioned in the introduction, we can suppose that X is finite-dimensional.

The above expression of δ when $\{c_j \mid j \in J\}$ is linearly independent is obtained by Li in relation (5.10) of [8] for the norm $\|\cdot\|$ on $\mathbb{R}^{m+l}$ satisfying condition (2.11); note that Li's proof is very different. Note also that the inequality in (5.1) can be strict if $\{c_j \mid j \in J\}$ is not linearly independent, as the example given at the end of the preceding section shows.

Taking into consideration Corollary 4.4 we have that $\delta \geq \rho$ (where ρ is defined by relation (4.9)); this inequality can be strict even if $\{c_j \mid j \in J\}$ is linearly independent, as shown by Li in [8, Prop. 5.1]. Using Corollary 4.4 and the preceding proposition, in the case $l = 0$ there are several formulae for δ.

COROLLARY 5.2. *Assume that the norm $\|\cdot\|$ on $\mathbb{R}^m$ satisfies condition* (2.10). *Then*

$$\begin{aligned}\delta &= \min\{\tau_K \mid K \in \mathcal{R}(I)\} = \min\{\tau_K \mid K \in \mathcal{R}_f(I)\} = \min\{\tau_K \mid K \text{ is maximal in } \mathcal{R}(I)\}\\ &= \min\{\tau_K \mid K \in \mathcal{L}(I)\} = \min\{\tau_K \mid K \in \mathcal{L}_f(I)\}.\end{aligned}$$

Proposition 5.1 can be used for deriving a formula for δ established by Belousov and Andronov [2] when the spaces are endowed with Euclidean norms. Assume that X is a Hilbert space and $\emptyset \neq K \subset I$; by G^K we denote the Gram matrix $(\langle a_i, a_{i'}\rangle)_{(i,i')\in K\times K}$. When $(K, L) \subset (I, J)$, the Gram matrix $G^{K,L}$ is defined similarly. Moreover, for $y \in \mathbb{R}^K$, $y > 0$ means that $y_i > 0$ for every $i \in K$.

COROLLARY 5.3. *Assume that X is a Hilbert space, $\{c_j \mid j \in J\}$ is linearly independent, and $\mathbb{R}^{m+l}$ is endowed with the Euclidean norm $\|\cdot\|_2$. Then*

$$\delta = \sqrt{\overline{\lambda}},$$

where $\overline{\lambda}$ is the smallest among all eigenvalues of Gram matrices $M^{K,J}$ which correspond to eigenvectors $(y, z) \in \mathbb{R}^K \times \mathbb{R}^J$ with $y > 0$ for $K \subset I$ such that $(K, J) \in \mathcal{L}(I, J)$.

Proof. Let $(K, J) \in \mathcal{L}(I, J)$ be such that $\lambda \in \mathbb{R}$ is an eigenvalue corresponding to the eigenvector $(\overline{y}, \overline{z}) \in \mathbb{R}^K \times \mathbb{R}^J$ with $\overline{y} > 0$. We may assume that $\|(\overline{y}, \overline{z})\|_2 = 1$. Consider $\mu \in \mathbb{R}^m$ such that $\mu_i := \overline{y}_i$ for $i \in K$, $\mu_i := 0$ for $i \in I \setminus K$, and $\zeta := \overline{z} \in \mathbb{R}^l$. Then

$$\|A^*\mu + C^*\zeta\|^2 = \left\langle (\overline{y}, \overline{z}), M^{K,J}(\overline{y}, \overline{z})^T \right\rangle = \left\langle (\overline{y}, \overline{z}), \lambda(\overline{y}, \overline{z})^T \right\rangle = \lambda, \tag{5.2}$$

whence $\tau_{K,J} \leq \sqrt{\lambda}$. Because in our case the second part of Proposition 5.1 applies, we obtain that $\delta \leq \sqrt{\overline{\lambda}}$. Let us prove now the reverse inequality. Using again Proposition 5.1, there exists $K \subset I$ such that $(K, J) \in \mathcal{L}(I, J)$ and $\delta = \tau_{K,J}$. By the definition of $\tau_{K,J}$, there exists $(\mu, \zeta) \in M^{K,J}$ such that $\tau_{K,J} = \|A^*\mu + C^*\zeta\|$. Taking $K_0 := \{i \in K \mid \mu_i > 0\}$, we have that $(\mu, \zeta) \in M^{K_0,J}$. It follows that $\tau_{K,J} = \tau_{K_0,J}$. Replacing eventually K by K_0, we may assume that $\mu_i > 0$ for every $i \in K$. Let $D := \{(y, z) \in \mathbb{R}^K \times \mathbb{R}^J \mid y > 0\}$, $\phi, \psi : D \to \mathbb{R}$ be defined by $\phi(y, z) := \frac{1}{2}\left\langle (y, z), M^{K,J}(y, z)^T \right\rangle$, and $\psi(y, z) := \frac{1}{2}\|(y, z)\|^2$. It is obvious that D is an open set, ϕ, ψ are $\mathcal{C}^1$ functions, and $\nabla\phi(y, z) = M^{K,J}(y, z)^T$, $\nabla\psi(y, z) = (y, z)^T$. Let $\overline{y}_i := \mu_i$ for $i \in K$ and $\overline{z} := \zeta$. It follows that $(\overline{y}, \overline{z})$ is an optimal solution of the minimization problem: min $\phi(y, z)$ subject to $\psi(y, z) = \frac{1}{2}$. Since $\nabla\psi(\overline{y}, \overline{z})$ is onto, there exists $\lambda \in \mathbb{R}$ such that

$$0 = \nabla\left(\phi - \lambda\psi\right)(\overline{y}, \overline{z}) = M^{K,J}(\overline{y}, \overline{z})^T - \lambda(\overline{y}, \overline{z})^T, \tag{5.3}$$

which means that $(\overline{y},\overline{z})$ is an eigenvector of $M^{K,J}$ and λ is an eigenvalue which corresponds to $(\overline{y},\overline{z})$. From (5.3) and (5.2) we obtain that $\delta^2=\lambda$, and so $\delta\geq\sqrt{\overline{\lambda}}$. The proof is complete. □

6. Concluding remarks. For deriving the formulae and estimates for the sharp Hoffman constant we used results on global error bounds for convex inequality systems. A similar approach was used by Azé and Corvellec [1] and Ng and Zheng [11].

• We established a formula for the sharp Hoffman constant $\delta_{b,d}$ at $(b,d)\in\operatorname{dom}F$ and a formula for the global sharp Hoffman constant δ in the spirit of those established by Belousov and Andronov [2]; taking $l=0$ and $X=\mathbb{R}^k$ endowed with the Euclidean norm, the formula for δ reduces to that given in [2].

• We showed that the Hoffman constant found by Ng and Zheng is in fact the sharp Hoffman constant.

• We showed that the sharp Hoffman constant established by Li [8] is valid for more general norms on $\mathbb{R}^{m+l}$; a similar formula for the sharp Hoffman constant at $(b,d)\in\operatorname{dom}F$ is furnished. Also, using Li's formula, we deduced a formula for the sharp Hoffman constant by using eigenvalues of Gram matrices in the case when all spaces are Hilbert spaces; this reduces to the Belousov and Andronov formula when $C=0$.

• We gave properties and characterized several types of monotonicity for norms on $\mathbb{R}^{m+l}$.

Acknowledgments. We thank the referees for their attentive reading of the manuscript and for their remarks, which improved the presentation of the paper. The current shorter proof of Lemma 4.3 is based on an idea of one of the referees.

REFERENCES

[1] D. Azé and J.-N. Corvellec, *On the sensitivity analysis of Hoffman constants for systems of linear inequalities*, SIAM J. Optim., 12 (2002), pp. 913–927.

[2] E. G. Belousov and V. G. Andronov, *On exact Lipschitz and Hoffman constants for systems of linear inequalities*, Vestnik Moskov. Univ. Ser. XV Vychisl. Mat. Kibernet., 47 (1999), pp. 28–32.

[3] C. Bergthaller and I. Singer, *The distance to a polyhedron*, Linear Algebra Appl., 169 (1992), pp. 111–129.

[4] A. J. Hoffman, *On approximate solutions of systems of linear inequalities*, J. Research Nat. Bur. Standards, 49 (1952), pp. 263–265.

[5] D. Klatte and W. Li, *Asymptotic constraint qualifications and global error bounds for convex inequalities*, Math. Program., 84 (1999), pp. 137–160.

[6] B. Lemaire, *Well-posedness, conditioning and regularization of minimization, inclusion and fixed-point problems*, Pliska Stud. Math. Bulgar., 12 (1998), pp. 71–84.

[7] A. S. Lewis and J.-S. Pang, *Error bounds for convex inequality systems*, in Proceedings of the 5th International Symposium on Generalized Convexity (Luminy, 1996), J.-P. Crouzeix, J.-E. Martinez-Legaz, and M. Volle, eds., Kluwer Academic, Dordrecht, The Netherlands, 1998, pp. 75–110.

[8] W. Li, *The sharp Lipschitz constants for feasible and optimal solutions of a perturbed linear program*, Linear Algebra Appl., 187 (1993), pp. 15–40.

[9] W. Li, *Sharp Lipschitz constants for basic optimal solutions and basic feasible solutions of linear programs*, SIAM J. Control Optim., 32 (1994), pp. 140–153.

[10] W. Li and I. Singer, *Global error bounds for convex multifunctions and applications*, Math. Oper. Res., 23 (1998), pp. 443–462.

[11] K. F. Ng and X. Y. Zheng, *Error bounds for lower semicontinuous functions in normed spaces*, SIAM J. Optim., 12 (2001), pp. 1–17.

[12] Z. Wu and J. J. Ye, *On error bounds for lower semicontinuous functions*, Math. Program., 92 (2002), pp. 301–314.

[13] C. Zălinescu, *A nonlinear extension of Hoffman's error bounds for linear inequalities*, Math. Oper. Res., 28 (2003), pp. 524–532.

[14] C. Zălinescu, *Weak sharp minima, well-behaving functions and global error bounds for convex inequalities in Banach spaces*, in Proceedings of the 12th Baikal International Conference on Optimization Methods and their Applications, V. Bulatov and V. Baturin, eds., Institute of System Dynamics and Control Theory, Siberian Branch of the Russian Academy of Sciences, Irkutsk, 2001, pp. 272–284.

[15] C. Zălinescu, *Convex Analysis in General Vector Spaces*, World Scientific, River Edge, NJ, 2002.

SIAM J. OPTIM.
Vol. 14, No. 2, pp. 534–547

NEW SEQUENTIAL LAGRANGE MULTIPLIER CONDITIONS CHARACTERIZING OPTIMALITY WITHOUT CONSTRAINT QUALIFICATION FOR CONVEX PROGRAMS*

V. JEYAKUMAR†, G. M. LEE‡, AND N. DINH§

Abstract. In this paper a new sequential Lagrange multiplier condition characterizing optimality without a constraint qualification for an abstract nonsmooth convex program is presented in terms of the subdifferentials and the ϵ-subdifferentials. A sequential condition involving only the subdifferentials, but at nearby points to the minimizer for constraints, is also derived. For a smooth convex program, the sequential condition yields a limiting Kuhn–Tucker condition at nearby points without a constraint qualification. It is shown how the sequential conditions are related to the standard Lagrange multiplier condition. Applications to semidefinite programs, semi-infinite programs, and semiconvex programs are given. Several numerical examples are discussed to illustrate the significance of the sequential conditions.

Key words. ϵ-subdifferential, sequential ϵ-subgradient optimality conditions, necessary and sufficient conditions

AMS subject classifications. 90C25, 52A41, 26E15

DOI. S1052623402417699

1. Introduction.

Consider the convex programming model problem

$$\text{(P)} \quad \begin{aligned} &\text{Minimize} \quad f(x) \\ &\text{subject to} \quad g(x) \in -S, \end{aligned}$$

where X is a reflexive Banach space, Z is a locally convex (Hausdorff) space, S is a closed convex cone in Z, which does not necessarily have nonempty interior, $f : X \to \mathbb{R}$ is a continuous convex function, and $g : X \to Z$ is a continuous and S-convex function. It is well known that for the convex programming problem (P) the Lagrange multiplier condition that

$$(\exists \lambda \in S^+) \;\; 0 \in \partial f(a) + \partial(\lambda g)(a), \;\; \lambda g(a) = 0, \tag{1}$$

is sufficient for optimality. However, this condition requires a constraint qualification to completely characterize optimality. The constraint qualifications do not always hold for finite-dimensional convex programs and frequently fail for infinite-dimensional convex programs. Over the years a great deal of attention has been focused on the characterizations of optimality which avoid a constraint qualification. As a result, various modified Lagrange multiplier conditions without a constraint qualification have been given in the literature (see [2, 3, 13, 17] and the references therein). More

*Received by the editors November 11, 2002; accepted for publication (in revised form) June 17, 2003; published electronically November 6, 2003.
http://www.siam.org/journals/siopt/14-2/41769.html

†Department of Applied Mathematics, University of New South Wales, Sydney 2052, Australia (jeya@maths.unsw.edu.au).

‡Department of Applied Mathematics, Pukyong National University, Pusan 608-737, Korea (gmlee@pknu.ac.kr). The work of this author was partially carried out while he was visiting the University of New South Wales.

§Department of Mathematics-Informatics, Pedagogical Institute of Ho Chi Minh City, HCM City, Vietnam. The work of this author was carried out while he was at the Pukyong National University, Korea, and was supported by a KOSEF-APEC Postdoctoral Fellowship.

recently, Thibault [20] gave an elegant sequential form of the Lagrange multiplier condition for (P) characterizing optimality without a constraint qualification in the case where S is a closed convex *normal* cone. The sequential condition involving the subdifferentials at nearby points to a minimizer was derived using the sequential subdifferential calculus of convex functions (see [1, 10, 20, 21]).

The aim of this paper is threefold: First, we establish a new sequential form of the Lagrange multiplier condition that is expressed at the minimizer rather than at nearby points. This is achieved by employing both the subdifferential and the ϵ-subdifferential [7, 8, 9] for deriving the sequential condition. The key to the derivation is the simple description of the epigraph of a conjugate function in terms of the ϵ-subdifferentials and the direct application of the Hahn–Banach separation theorem [11]. Second, we show how a sequential Lagrange multiplier condition involving only the subdifferentials at nearby points to a minimizer can be derived from the new sequential form. We derive such a condition as an application of the Brondsted–Rockafellar theorem [4, 20], which paves the way for describing an ϵ-subgradient at a point in terms of the subgradients at nearby points. These results show in particular that the absence of a constraint qualification for a smooth convex program means that the Lagrange multipliers are weakened to satisfy the Kuhn–Tucker conditions in the limit by a sequence of Lagrange multipliers at nearby points for constraints. Third, we show how our sequential condition is related to the Lagrange multiplier condition (1) under a constraint qualification. We establish that the new sequential condition collapses to the Lagrange multiplier condition (1) under a simple, but more general, closed cone constraint qualification which is implied by well-known constraint qualifications such as the generalized Slater condition [15, 16] or the Robinson regularity condition [17].

The rest of the paper is organized as follows. Section 2 explains some basic results on convex sets and functions and points out important properties of conjugate functions and ϵ-subdifferentials that will be used later in the paper. Section 3 presents sequential Lagrange multiplier conditions for the convex programming model problem (P). Section 4 describes a simple closed cone condition as a constraint qualification which ensures that the Lagrange multiplier condition (1) holds. Finally, section 5 derives sequential Lagrange multiplier conditions for semidefinite programs, semi-infinite programs, and semiconvex programs and illustrates the significance of the sequential conditions by numerical examples. The main results are presented in reflexive Banach spaces X in order to avoid the use of nets.

2. Preliminaries: Conjugacy and ϵ-subdifferentials. Let us first recall some notation and preliminary results which will be used throughout the paper. In what follows we will always assume that the feasible set $A := g^{-1}(-S) = \{x \in X \mid g(x) \in -S\}$ is nonempty. The continuous dual space of X will be denoted X' and will be endowed with the weak* topology. For a set $D \subset$ X, the *closure* and *convex hull* of D will be denoted $\operatorname{cl} D$ and $\operatorname{co} D$, respectively. The *support function* σ_D is defined by $\sigma_D(u) = \sup_{x\in D} u(x)$, and the *cone* generated by D will be denoted $\operatorname{cone} D := \bigcup_{\alpha\geq 0} \alpha D$. The *core* of D is defined by

$$\operatorname{core} D = \{d \in D \mid (\forall x \in X)(\exists \epsilon > 0)(\forall \lambda \in [-\epsilon, \epsilon])\ d + \lambda x \in D\}.$$

The (positive) polar of the cone $S \subseteq Z$ is the cone $S^+ = \{\theta \in Z' \mid \theta(k) \geq 0\ \forall k \in S\}$. Let $f : X \to \mathbb{R} \cup \{+\infty\}$ be a proper lower semicontinuous convex function. Then the *conjugate* function $f^* : X' \to \mathbb{R} \cup \{+\infty\}$ is defined by

$$f^*(v) = \sup\{v(x) - f(x) \mid x \in \operatorname{dom} f\},$$

where the domain of f, dom f, is given by

$$\text{dom } f = \{x \in X \mid f(x) < +\infty\}.$$

The epigraph of f, epi f, is defined by

$$\text{epi } f = \{(x, r) \in X \times \mathbb{R} \mid x \in \text{ dom } f,\ f(x) \le r\}.$$

Recall that, for $\epsilon \ge 0$, the *ϵ-subdifferential* of f at $a \in \text{dom } f$ is defined as the nonempty weak* closed convex set

$$\partial_\epsilon f(a) = \{v \in X' \mid f(x) - f(a) \ge v(x - a) - \epsilon \ \forall x \in \text{ dom } f\}.$$

For a detailed discussion on the ϵ-subdifferential and its properties, see [7, 8, 9]. Note that $\bigcap_{\epsilon>0} \partial_\epsilon f(a) = \partial f(a)$, where the latter set denotes the usual convex subdifferential of f at a. If $\tilde{f}(x) = f(x) - k$, $x \in X$, $k \in \mathbb{R}$, then epi $\tilde{f}^* =$ epi $f^* + (0, k)$. The following proposition, which describes the relationship between the epigraph of a conjugate function and the ϵ-subdifferential and which plays a key role in deriving the main results, was recently given in [13]. For the sake of completeness we give a short proof here.

PROPOSITION 2.1. *Let $f : \mathrm{X} \to \mathbb{R} \cup \{+\infty\}$ be a lower semicontinuous (proper) convex function and let $a \in$ dom f. Then*

$$\text{epi } f^* = \bigcup_{\epsilon \ge 0} \{(v,\ \epsilon + v(a) - f(a)) : v \in \partial_\epsilon f(a)\}. \tag{2}$$

Proof. Let $(u, r) \in \text{epi } f^*$. Then $f^*(u) \le r$. From the definition of conjugate function, for each $x \in X$, $f^*(u) \ge u(x) - f(x)$; thus, for each $u \in X$, $u(x) - f(x) \le r$. Let $\epsilon_0 = r + f(a) - u(a) \ge 0$. So, $r = \epsilon_0 - f(a) + u(a)$. Now, for each $x \in X$,

$$f(x) - f(a) \ge u(x) - r - f(a) = u(x - a) - \epsilon_0;$$

thus, $u \in \partial_{\epsilon_0} f(a)$. Hence,

$$\text{epi } f^* \subset K := \bigcup_{\epsilon \ge 0} \{(v,\ \epsilon + v(a) - f(a)) : v \in \partial_\epsilon f(a)\}.$$

Conversely, let $(u, r) \in K$. Then there exists $\epsilon_0 \ge 0$ such that $u \in \partial_{\epsilon_0} f(a)$ and $r = -f(a) + u(a) + \epsilon_0$. This gives us $f^*(u) + f(a) - u(a) \le \epsilon_0$, which means that $f^*(u) \le \epsilon_0 + u(a) - f(a)$; thus, $f^*(u) \le r$ and so, $(u,r) \in \text{epi } f^*$. ∎

The mapping $g : X \to Z$ is called *S-convex* if for every u, $v \in X$ and every $t \in [0, 1]$,

$$g(tu + (1 - t)v) - tg(u) - (1 - t)g(v) \in -S.$$

For a continuous S-convex mapping g, it is easy to show that the set

$$\bigcup_{\lambda \in S^+} \text{epi } (\lambda g)^*$$

is a convex cone [15]. We conclude this section by recalling a version of the Brondsted–Rockafellar theorem which was established in [20].

PROPOSITION 2.2 (Brondsted–Rockafellar Theorem [4, 20]). *Let X be a Banach space and let $f : X \to \mathbb{R} \cup \{+\infty\}$ be a proper lower semicontinuous convex function. Then for any real number $\epsilon > 0$ and any $x^* \in \partial_\epsilon f(\bar{x})$ there exist $x_\epsilon \in X$, $x^*_\epsilon \in \partial f(x_\epsilon)$ such that*

$$\|x_\epsilon - \bar{x}\| \le \sqrt{\epsilon}, \quad \|x^*_\epsilon - x^*\| \le \sqrt{\epsilon}, \quad \textit{and} \quad |f(x_\epsilon) - x^*_\epsilon(x_\epsilon - \bar{x}) - f(\bar{x})| \le 2\epsilon.$$

3. Sequential Lagrange multiplier conditions. In this section, we present, without any constraint qualification, a necessary and sufficient optimality condition for (P) in the form of a sequential Lagrange multiplier rule expressed in terms of the subdifferentials and the ϵ-subdifferentials of the functions involved at the minimizer. We then derive corresponding conditions characterizing optimality involving only subdifferentials at nearby points for constraint functions. We begin by establishing a dual condition characterizing the feasibility of the problem (P). This is then used to derive the sequential condition. For convenience, we shall denote composition of mappings by juxtaposition, i.e., $\lambda \circ g$ as λg, where $\lambda \in Z'$ and $g : X \to Z$.

LEMMA 3.1. *Let $g : X \to Z$ be a continuous and S-convex mapping. Then $g^{-1}(-S) \neq \emptyset$ if and only if $(0,-1) \notin \operatorname{cl}\left(\bigcup_{\lambda\in S^+} \operatorname{epi}(\lambda g)^*\right)$.*

Proof. Let $\lambda \in S^+$ and let $x^* \in X'$. Since $-\lambda g(x) \geq 0$ for all $x \in g^{-1}(-S)$, we have

$$\begin{aligned}(\lambda g)^*(x^*) &= \sup_{x\in X}[x^*(x) - \lambda g(x)]\\ &\geq \sup_{x\in A}[x^*(x) - \lambda g(x)]\\ &\geq \sup_{x\in A} x^*(x) = \sigma_A(x^*),\end{aligned}$$

where $A = g^{-1}(-S)$. This inequality, together with the fact that epi σ_A is weak* closed, gives us that

$$\operatorname{cl}\left(\bigcup_{\lambda\in S^+} \operatorname{epi}\,(\lambda g)^*\right) \subset \operatorname{epi}\,\sigma_A.$$

If $g^{-1}(-S) \neq \emptyset$, then clearly $(0,-1) \notin \operatorname{epi}\,\sigma_A$, and so from the above inclusion $(0,-1) \notin \operatorname{cl}\left(\bigcup_{\lambda\in S^+} \operatorname{epi}\,(\lambda g)^*\right)$.

Conversely if $(0,-1) \notin \operatorname{cl}\left(\bigcup_{\lambda\in S^+} \operatorname{epi}\,(\lambda g)^*\right)$, then by the Hahn–Banach separation theorem [11] there is $(x,\alpha) \in X \times \mathbb{R}$, $(x,\alpha) \neq (0,0)$ such that $-\alpha < 0$ and

$$v(x) + \gamma\alpha \geq 0 \;\; \forall (v,\gamma) \in \operatorname{cl}\left(\bigcup_{\lambda\in S^+} \operatorname{epi}\,(\lambda g)^*\right).$$

Let $\bar{x} = \frac{x}{\alpha}$. Then we have

$$v(-\bar{x}) - \gamma \leq 0 \quad \forall (v,\gamma) \in \operatorname{cl}\left(\bigcup_{\lambda\in S^+} \operatorname{epi}\,(\lambda g)^*\right).$$

So, for each $\lambda \in S^+$ and for each $v \in \operatorname{dom}\,(\lambda g)^*$, $v(-\bar{x}) - (\lambda g)^*(v) \leq 0$. Since λg is continuous,

$$(\lambda g)(-\bar{x}) = (\lambda g)^{**}(-\bar{x}) = \sup_v [v(-\bar{x}) - (\lambda g)^*(v)] \leq 0.$$

This implies that $g(-\bar{x}) \in -S$, and hence $g^{-1}(-S) \neq \emptyset$. ∎

The corresponding result to Lemma 3.1 in the case of real-valued functions g was used in [12] to study dual characterizations of the containment of a convex set in another convex set. We now derive the sequential necessary and sufficient condition for optimality for (P). Note that the weak* convergence of the sequence $\{w_n\}$ of X' to w will be denoted by $w_n \to_* w$.

THEOREM 3.1. *For the convex program* (P)*, let $a \in A$. Then the point a is a minimizer of* (P) *if and only if there exist $u \in \partial f(a)$, $\{\epsilon_n\} \subset \mathbb{R}_+$, $\{\lambda_n\} \subset S^+$, and*

$\{v_n\} \subset X'$ such that $v_n \in \partial_{\epsilon_n}(\lambda_n g)(a)$, $u + v_n \to_* 0$, $\lambda_n g(a) \to 0$, and $\epsilon_n \to 0$ as $n \to \infty$.

Proof. We first note that $a \in A$ is a minimizer of (P) if and only if there exists $u \in \partial f(a)$ such that $u(x) \geq u(a)$ for each $x \in g^{-1}(-S)$. We now show that this condition is equivalent to the statement that there exists $u \in \partial f(a)$ such that

$$(3) \qquad \big(\ -u, -u(a)\ \big) \in \mathrm{cl}\left(\bigcup_{\lambda \in S^+} \mathrm{epi}\ (\lambda g)^*\right).$$

Indeed, if (3) holds, then there exist $\{\mu_n\} \subset S^+$, $\{u_n\} \subset X'$, and $\{r_n\} \subset \mathbb{R}$ such that $u_n \to_* -u$, $r_n \to -u(a)$, and $(\mu_n g)^*(u_n) \leq r_n$ for each n. Then, for each $x \in g^{-1}(-S)$, $u_n(x) \leq r_n + \mu_n g(x) \leq r_n$. Letting $n \to \infty$, we get $-u(x) \leq -u(a)$.

Conversely, assume that there exists $u \in \partial f(a)$ such that $u(x) \geq u(a)$ for each $x \in g^{-1}(-S)$. Suppose to the contrary that $(-u, -u(a)) \notin \mathrm{cl}\ \big(\bigcup_{\lambda \in S^+} \mathrm{epi}\ (\lambda g)^*\big)$. By Lemma 3.1, $(0, -1) \notin \mathrm{cl}\,\big(\bigcup_{\lambda \in S^+} \mathrm{epi}\ (\lambda g)^*\big)$, since $g^{-1}(-S) \neq \emptyset$. Therefore,

$$B \cap \left(\mathrm{cl}\left(\bigcup_{\lambda \in S^+} \mathrm{epi}\ (\lambda g)^*\right)\right) = \emptyset,$$

where

$$B := \{\delta(-u, -u(a)) + (1 - \delta)(0, -1) \in X' \times \mathbb{R} \mid \delta \in [0, 1]\}$$

is the segment connecting the points $(-u, -u(a))$ and $(0, -1)$. Otherwise, there is $\delta \in (0, 1)$ such that

$$\delta(-u, -u(a)) + (1 - \delta)(0, -1) \in \mathrm{cl}\left(\bigcup_{\lambda \in S^+} \mathrm{epi}\ (\lambda g)^*\right);$$

thus, $(-\delta u, -\delta u(a) - (1 - \delta)) \in \mathrm{cl}\,\big(\bigcup_{\lambda \in S^+} \mathrm{epi}(\lambda g)^*\big)$.

Also $\{0\} \times \mathbb{R}_+ \subset \mathrm{cl}\,\big(\bigcup_{\lambda \in S^+} \mathrm{epi}(\lambda g)^*\big)$, since $0 \in S^+$ and $\mathrm{epi}(0g)^* = \{0\} \times \mathbb{R}_+$. So we have

$$(-\delta u, -\delta u(a)) = (-\delta u, -\delta u(a) - (1 - \delta)) + (0, 1 - \delta) \in \mathrm{cl}\left(\bigcup_{\lambda \in S^+} \mathrm{epi}\ (\lambda g)^*\right),$$

which implies that

$$(-u, -u(a)) = \frac{1}{\delta}(-\delta u, -\delta u(a)) \in \mathrm{cl}\left(\bigcup_{\lambda \in S^+} \mathrm{epi}\ (\lambda g)^*\right),$$

a contradiction. Now, by the Hahn–Banach separation theorem, there is $(x, \beta) \in X \times \mathbb{R}$, $(x, \beta) \neq (0, 0)$, such that

$$[\delta(-u, -u(a)) + (1 - \delta)(0, -1)](x, \beta) < 0 \ \ \forall \delta \in [0, 1]$$

and

$$v(x) + \gamma\beta \geq 0 \quad \forall (v, \gamma) \in \mathrm{cl}\left(\bigcup_{\lambda \in S^+} \mathrm{epi}\ (\lambda g)^*\right).$$

By letting $\delta = 0$ we get $\beta > 0$, and by letting $\delta = 1$ we obtain $u(x) + u(a)\beta > 0$; thus, $u(\frac{-x}{\beta}) < u(a)$. On the other hand, for each $\lambda \in S^+$,

$$v\left(\frac{-x}{\beta}\right) - \gamma \leq 0 \quad \forall (v, \gamma) \in \text{ epi } (\lambda g)^*.$$

This gives us

$$v\left(\frac{-x}{\beta}\right) - (\lambda g)^*(v) \leq 0 \quad \forall v \in \text{ dom } (\lambda g)^*.$$

Hence, for each $\lambda \in S^+$,

$$(\lambda g)\left(\frac{-x}{\beta}\right) = (\lambda g)^{**}\left(\frac{-x}{\beta}\right) = \sup_v \left[v\left(\frac{-x}{\beta}\right) - (\lambda g)^*(v)\right] \leq 0.$$

From this we deduce that $\frac{-x}{\beta} \in g^{-1}(-S)$. This is a contradiction, as $u(\frac{-x}{\beta}) < u(a)$.

Now (2) shows that (3) is equivalent to the condition that

$$(-u, -u(a)) \in \text{cl}\left(\bigcup_{\lambda \in S^+} \bigcup_{\epsilon \geq 0} \left\{\left(w,\ w(a) + \epsilon - \lambda g(a)\right) \mid w \in \partial_\epsilon \lambda g(a)\right\}\right). \tag{4}$$

From (4), we see that there exist $\{\epsilon_n\} \subset \mathbb{R}_+$, $\{\lambda_n\} \subset S^+$, and $v_n \in \partial_{\epsilon_n}(\lambda_n g)(a)$ such that

$$(-u, -u(a)) = \lim_n \ (v_n,\ v_n(a) + \epsilon_n - \lambda_n g(a)).$$

Hence

$$u = -\lim_n v_n,\ u \in \partial f(a), \tag{5}$$

$$-u(a) = \lim_n [v_n(a) + \epsilon_n - \lambda_n g(a)]. \tag{6}$$

From (5) and (6), we obtain

$$\lim_n v_n(a) = -u(a) = \lim_n [v_n(a) + \epsilon_n - \lambda_n g(a)],$$

which implies that $0 = \lim_n [\epsilon_n - \lambda_n g(a)]$. Since $\epsilon_n \geq 0$ and $-\lambda_n g(a) \geq 0$ for all $n \in \mathbb{N}$, $\lim_n \epsilon_n = 0$ and $\lim_n \lambda_n g(a) = 0$. Therefore, if a is a minimizer of (P), then there exist $u \in \partial f(a)$, $\{\epsilon_n\} \subset \mathbb{R}_+$, $\{\lambda_n\} \subset S^+$, and $\{v_n\} \subset X'$ such that $v_n \in \partial_{\epsilon_n}(\lambda_n g)(a)$, $u + v_n \to_* 0$, $\lambda_n g(a) \to 0$, and $\epsilon_n \to 0$ as $n \to \infty$.

Conversely, if this sequential condition holds, then

$$-u(a) = \lim_n v_n(a) = \lim_n [v_n(a) + \epsilon_n - \lambda_n g(a)].$$

This, together with the condition that $v_n \to_* -u$ as $n \to \infty$, gives us (4), which in turn implies that a is a minimizer of (P). □

The following example illustrates that the ϵ-subdifferentials in the description of the sequential Lagrange multiplier condition in Theorem 3.1 are essential and they cannot be replaced by the subdifferentials for the constraint functions.

EXAMPLE 3.1. *Consider the problem*

$$\begin{array}{ll} \textit{Minimize} & x \\ \textit{subject to} & (x^2+y^2)^{\frac{1}{2}} - y \le 0. \end{array}$$

Let $f : \mathbb{R}^2 \to \mathbb{R}$ *and* $g : \mathbb{R}^2 \to \mathbb{R}$ *be given by* $f(x,y) = x$, $g(x,y) = (x^2+y^2)^{\frac{1}{2}} - y$, $(x,y) \in \mathbb{R}^2$, $S = \mathbb{R}_+$. *The feasible set is* $A = \{(x,y) \mid x = 0,\ y \ge 0\}$, $a = (0,1)$ *is a minimizer, and* $\partial f(a) = \{(1,0)\}$. *Observe that the Slater condition does not hold and that for any* $\lambda > 0$ *and* $\epsilon > 0$,

$$\partial_\epsilon(\lambda g)(a) = \{(v_1, v_2) \mid v_1^2 + (v_2+\lambda)^2 \le \lambda^2,\ v_2 \ge -\epsilon\}.$$

For each $n \in \mathbb{N}$, *if we take* $\epsilon_n = \frac{1}{n}$, $\lambda_n = \frac{1}{2}(n + \frac{2}{n}) + 1$, *and* $v_n = (-1 - \frac{1}{n}, -\frac{1}{n})$, *then* $v_n \in \partial_{\epsilon_n}(\lambda_n g)(a)$, $u + v_n \to 0$, $\epsilon_n \to 0$, *and* $\lambda_n g(a) = 0$. *Hence the sequential Lagrange multiplier condition holds. It is also worth noting that for each* $\lambda \in \mathbb{R}_+$, *we have* $\partial(\lambda g)(a) = \{(0,0)\}$; *hence,*

$$-u = (-1,0) \notin \operatorname{cl}\left(\bigcup_{\lambda \in \mathbb{R}_+} \partial(\lambda g)(a)\right).$$

We now derive from Theorem 3.1 a sequential condition similar to the one in Thibault [20] solely in terms of the subdifferentials of the functions involved.

THEOREM 3.2. *For the convex program* (P), *let* $a \in A$. *Then the point* a *is a minimizer of* (P) *if and only if there exist* $u \in \partial f(a)$, $\{\lambda_n\} \subset S^+$, $\{x_n\} \subset X$, *and* $\{v_n\} \subset X'$ *such that* $v_n \in \partial(\lambda_n g)(x_n)$, $u + v_n \to_* 0$, $\|x_n - a\| \to 0$, *and* $\lambda_n g(x_n) \to 0$ *as* $n \to \infty$.

Proof. Suppose that a is a minimizer of (P). By Theorem 3.1 there exist $u \in \partial f(a), \{\lambda_n\} \subset S^+, \{\epsilon_n\} \subset \mathbb{R}_+$, and $w_n \in \partial_{\epsilon_n}(\lambda_n g)(a)$ such that $u + w_n \to_* 0$, $\lambda_n g(a) \to 0$, *and* $\epsilon_n \to 0$ as $n \to \infty$. Since $w_n \in \partial_{\epsilon_n}(\lambda_n g)(a)$, it follows from Proposition 2.2 that there exist $x_n \in X$ and $v_n \in \partial(\lambda_n g)(x_n)$ satisfying

$$\|x_n - a\| \le \sqrt{\epsilon_n}, \quad \|w_n - v_n\| \le \sqrt{\epsilon_n}, \quad |\lambda_n g(x_n) - v_n(x_n - a) - \lambda_n g(a)| \le 2\epsilon_n.$$

Since $\epsilon_n \to 0$, we have $\|x_n - a\| \to 0$, $\|w_n - v_n\| \to 0$. This implies that $u + v_n \to_* 0$ and $v_n(x_n - a) \to 0$; hence $\lambda_n g(x_n) \to 0$.

Conversely, suppose that there exist $u \in \partial f(a)$, $\{\lambda_n\} \subset S^+$, $\{x_n\} \subset X$, and $v_n \in \partial(\lambda_n g)(x_n)$ such that $u + v_n \to_* 0$, $\|x_n - a\| \to 0$, and $\lambda_n g(x_n) \to 0$ as $n \to \infty$. On one hand, since $v_n \in \partial(\lambda_n g)(x_n)$, we have $(\lambda_n g)^*(v_n) = v_n(x_n) - \lambda_n g(x_n)$. So,

$$(v_n,\ v_n(x_n) - \lambda_n g(x_n)) \in \operatorname{epi}\ (\lambda_n g)^*. \tag{7}$$

On the other hand, since $\|x_n - a\| \to 0$, $v_n \to_* -u$, and $\lambda_n g(x_n) \to 0$ as $n \to \infty$, we get

$$\lim_n [v_n(x_n) - \lambda_n g(x_n)] = -u(a).$$

Combining this with (7) we get

$$(-u,\ -u(a)) \in \operatorname{cl}\left(\bigcup_{\lambda \in S^+} \operatorname{epi}\ (\lambda g)^*\right),$$

which proves that a is a minimizer of (P). □

COROLLARY 3.1. *For the convex program* (P)*, assume further that f and g are Fréchet differentiable on X and that $a \in A$. Then the point a is a minimizer of* (P) *if and only if there exist $\{\lambda_n\} \subset S^+$ and $\{x_n\} \subset X$ such that $\nabla f(a) + \nabla(\lambda_n g)(x_n) \to 0$, $\|x_n - a\| \to 0$, and $\lambda_n g(x_n) \to 0$ as $n \to \infty$.*

Proof. The conclusion follows from the previous theorem as in this case $\partial(\lambda_n g)(x_n) = \{\nabla(\lambda_n g)(x_n)\}$ and $\partial f(a) = \{\nabla f(a)\}$. □

The following simple example illustrates the significance of the sequential Lagrange multiplier condition for differentiable convex programs.

EXAMPLE 3.2. *Consider the problem*

$$\begin{array}{ll} \textit{Minimize} & -x \\ \textit{subject to} & [\max\{0, x\}]^2 \le 0. \end{array}$$

Let $f(x) := -x$, $g(x) := [\max\{0, x\}]^2$. Then the feasible set is $A = (-\infty, 0]$, and $a = 0$ is a minimizer. Let $\{\lambda_n\} \subset \mathbb{R}_+$ be any sequence such that $\lambda_n > 0$ for all n and $\lambda_n \to +\infty$, and let $x_n := \frac{1}{2\lambda_n}$. Then

$$\nabla f(a) + \lim_{n\to\infty} \nabla(\lambda_n g)(x_n) = -1 + \lim_{n\to\infty} \frac{2}{2\lambda_n}\lambda_n = 0, \text{ and}$$

$$\lim_{n\to\infty} \lambda_n g(x_n) = \lim_{n\to\infty} \lambda_n \frac{1}{(2\lambda_n)^2} = 0.$$

Hence, $\{\lambda_n\}$ is a sequential Lagrange multiplier for the problem at the minimizer $a = 0$. However, the Kuhn–Tucker condition for the problem does not hold at a.

4. A simple constraint qualification. In this section we derive the Lagrange multiplier condition (1) under a general *closed cone constraint qualification* which requires that the convex cone $\bigcup_{\lambda\in S^+} \operatorname{epi}(\lambda g)^*$ is weak* closed. This constraint qualification holds under the Robinson regularity condition, that is, $0 \in \operatorname{core}(g(X) + S)$ (see [17]), in the case where Z is a Banach space, or the generalized Slater condition that $\operatorname{int} S$ is nonempty and $-g(x_0) \in \operatorname{int} S$ (see [15]). Let us first see the relationship between the Lagrange multiplier condition and the convex cone $\bigcup_{\lambda\in S^+} \operatorname{epi}(\lambda g)^*$.

LEMMA 4.1. *For the problem* (P)*, let $a \in g^{-1}(-S)$. Then the following statements are equivalent:*

(i) $(\exists\ u \in \partial f(a))\quad (-u, -u(a)) \in \bigcup_{\lambda\in S^+} \operatorname{epi}(\lambda g)^*$

(ii) $(\exists \lambda \in S^+)\ \ 0 \in \partial f(a) + \partial(\lambda g)(a)$ and $\lambda g(a) = 0$.

Proof. (ii) ⇒ (i) Assume that (ii) holds. Then there exist $u \in \partial f(a)$ and $\lambda \in S^+$ such that $-u \in \partial(\lambda g)(a)$ and $\lambda g(a) = 0$. So, $(\lambda g)^*(-u) \le -u(a)$, and $(-u, -u(a)) \in \operatorname{epi}(\lambda g)^*$. Hence (i) holds.

(i) ⇒ (ii) If (i) holds, then there exist $u \in \partial f(a)$ and $\lambda \in S^+$ such that $(-u, -u(a)) \in \operatorname{epi}(\lambda g)^*$. Since $\operatorname{epi}(\lambda g)^*$ can be expressed in the form

$$\operatorname{epi}(\lambda g)^* = \bigcup_{\epsilon\ge 0} \{(w,\ w(a) + \epsilon - \lambda g(a))\ \mid w \in \partial_\epsilon(\lambda g)(a)\},$$

we deduce that $-u \in \partial_\epsilon(\lambda g)(a)$ and $-u(a) = -u(a) + \epsilon - \lambda g(a)$ for some $\epsilon \ge 0$. Since $a \in A$ and $\lambda \in S^+$, the last equality gives $\lambda g(a) = \epsilon = 0$. Thus $-u \in \partial(\lambda g)(a)$, and so we have $0 \in \partial f(a) + \partial(\lambda g)(a)$ and (ii) holds. □

THEOREM 4.1. *For the program* (P)*, assume that the closed cone constraint qualification holds and that* $a \in A$*. Then* a *is a minimizer of* (P) *if and only if there exists* $\lambda \in S^+$ *such that* $0 \in \partial f(a) + \partial(\lambda g)(a)$ *and* $\lambda g(a) = 0$.

Proof. Since the convex cone $\bigcup_{\lambda \in S^+} \text{epi}(\lambda g)^*$ is weak* closed, it follows from the first part of the proof of Theorem 3.1 (see (3)) that a is a minimizer of (P) if and only if there exists $u \in \partial f(a)$ such that $(-u, -u(a)) \in \bigcup_{\lambda \in S^+} \text{epi}\ (\lambda g)^*$. Applying Lemma 4.1 we obtain the desired conclusion. □

5. Applications and examples. In this section we apply the results in section 3 to some special classes of problems such as the classes of convex semidefinite programs [22] and convex semi-infinite programs [6, 13]. We also extend the main results to semiconvex programs [18].

5.1. Semidefinite programs. Consider the convex semidefinite programming model problem

$$\begin{aligned} \text{(SDP)} \quad & \text{Minimize} \quad f(x) \\ & \text{subject to} \quad F_0 + \sum_{i=1}^{m} x_i F_i \succeq 0, \end{aligned}$$

where $f : \mathbb{R}^m \to \mathbb{R}$ is a convex function, and for $i = 0, 1, \ldots, m$, $F_i \in S_n$, the space of $(n \times n)$ symmetric matrices. The space S_n is partially ordered by the Löwner order; that is, for $M, N \in S_n$, $M \succeq N$ if and only if $M - N$ is positive semidefinite. The inner product in S_n is defined by $(M, N) = \text{Tr}[MN]$, where $\text{Tr}[\cdot]$ is the trace operation. Let $S := \{M \in S_n \mid M \succeq 0\}$. Then

$$S^+ = \{\theta \in S_n \mid (\theta, Z) \geq 0 \ \forall Z \in S\} = S.$$

Let $F(x) := F_0 + \sum_{i=1}^m x_i F_i$, $\hat{F}(x) = \sum_{i=1}^m x_i F_i$, $x = (x_1, \ldots, x_m) \in \mathbb{R}^m$. Then $\hat{F}$ is a linear operator from $\mathbb{R}^m$ to S_n and its dual is defined by $\hat{F}^*(Z) = (\text{Tr}[F_1 Z], \ldots, \text{Tr}[F_m Z])$ for any $Z \in S_n$. Clearly, $A := \{x \in \mathbb{R}^m \mid F(x) \in S\}$ is the feasible set of (SDP).

There are classes of semidefinite programs where well-known constraint qualifications fail (see [19, 22]). So it is of interest to examine sequential Lagrange multiplier conditions for (SDP) without a constraint qualification. Here we obtain such a condition as an easy consequence of Theorem 3.1.

THEOREM 5.1. *For the problem* (SDP)*, let* $a \in A$*. Then the point* a *is a minimizer of* (SDP) *if and only if there exist* $u \in \partial f(a)$ *and a sequence* $\{Z_k\} \subset S$ *such that*

$$u - \hat{F}^*(Z_k) \to 0 \quad \textit{and} \quad \text{Tr}[Z_k F(a)] \to 0 \ \textit{as}\ k \to \infty.$$

Proof. Observe that F is affine and that for each $\epsilon \geq 0$ and for each $Z \in S$, $\partial_\epsilon(ZF)(x) = \hat{F}^*(Z)$. The conclusion follows from Theorem 3.1 by setting $g(x) = -F(x)$. □

When $f(x) = c^T x$, for some fixed $c \in \mathbb{R}^m$, the previous asymptotic optimality condition reduces to the following simple form given in [19]: There exists $\{Z_k\} \subset S$ such that $\hat{F}^*(Z_k) \to c$ and $\text{Tr}[Z_k F(a)] \to 0$ as $k \to \infty$. For (SDP), the closed cone constraint qualification reduces to the condition that the cone

$$D := \bigcup_{(Z,\ \delta) \in S \times \mathbb{R}_+} \left(-\hat{F}^*(Z),\ \text{Tr}[ZF_0] + \delta\right)$$

is closed. Under this condition, the following Lagrange multiplier rule holds (see [14]):

$$\hat{F}^*(Z) \in \partial f(a) \quad \text{and} \quad \operatorname{Tr}(ZF(a)) = 0.$$

The example below shows that for a semidefinite program the sequential Lagrange multiplier condition holds, whereas the nonasymptotic Lagrange multiplier condition fails to hold at the minimizer.

EXAMPLE 5.1. *Consider the problem*

$$\begin{array}{ll} \textit{Minimize} & x_1 + |x_2| \\ \textit{subject to} & \begin{pmatrix} 0 & x_1 & 0 \\ x_1 & x_2+1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \succeq 0. \end{array}$$

Let $f(x_1, x_2) := x_1 + |x_2|$ *and let*

$$F_0 := \begin{pmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \quad F_1 := \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \quad F_2 := \begin{pmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix},$$

and $F(x) := F_0 + x_1F_1 + x_2F_2,\ x = (x_1, x_2) \in \mathbb{R}^2$. *The feasible set of the problem is* $A = \{x \in \mathbb{R}^2 \mid F(x) \succeq 0\} = \{(0, x_2) \in \mathbb{R}^2 \mid x_2 \geq -1\}$ *and* $a = (0, 0)$ *is a minimizer. Define* $\hat{Z} := \{Z_n\}$, *where*

$$Z_n = \begin{pmatrix} n & \frac{1}{2} & 0 \\ \frac{1}{2} & \frac{1}{n} & 0 \\ 0 & 0 & 0 \end{pmatrix}.$$

Clearly, $\partial f(0, 0) = \{(1, \xi) \mid \xi \in [-1, 1]\}$, $Z_n \succeq 0$ *for all* $n \in \mathbb{N}$, *and* $\operatorname{Tr}[Z_nF_0] = \frac{1}{n},\ \operatorname{Tr}[Z_nF_1] = 1,\ \operatorname{Tr}[Z_nF_2] = \frac{1}{n}$. *Thus*

$$\lim_{n\to\infty} \hat{F}^* Z_n = \lim_{n\to\infty} (\operatorname{Tr}[Z_nF_1], \operatorname{Tr}[Z_nF_2]) = \lim_{n\to\infty} \left(1, \frac{1}{n}\right) = (1, 0) \in \partial f(0, 0),$$

and $\lim_{n\to\infty} \operatorname{Tr}[Z_nF(a)] = \lim_{n\to\infty} \operatorname{Tr}[Z_nF_0] = 0$. *Hence* $\hat{Z}$ *is a sequential Lagrange multiplier for the problem at* $a = (0, 0)$. *It is easy to see that the generalized Slater constraint qualification does not hold for this problem.*

5.2. Semi-infinite programs. Consider the following convex semi-infinite programming model problem

$$\begin{array}{lll} \text{(SIP)} & \text{Minimize} & f(x) \\ & \text{subject to} & g_i(x) \leq 0, \quad i \in I, \end{array}$$

where $X = \mathbb{R}^n$, I is an index set with cardinality possibly infinite, and $f,\ g_i : X \to \mathbb{R},\ i \in I$, are convex functions. Let $Z = \prod_I \mathbb{R}$ denote the product space with the product topology. Then Z' is the generalized finite sequence space consisting of all functionals $v : I \to \mathbb{R}$ with finite support (see [13]). Let $S = \prod_I \mathbb{R}_+$; then the (positive) dual cone S^+ of the cone S is

$$\Lambda := \{\lambda = (\lambda_i) \in Z' \mid \lambda_i \geq 0 \quad \forall\, i \in I,\ \lambda_i = 0 \text{ for all but a finite number of } i \in I\}.$$

Define $g := (g_i)$. Then $g : X \to Z$ is continuous and S-convex and the problem (SIP) can be rewritten in the form of (P). The next lemma is useful in deriving sequential Lagrange multiplier conditions for (SIP).

LEMMA 5.1. *Let $X = \mathbb{R}^n$ and let $g_i : X \to \mathbb{R},\ i \in I$ be convex functions. Then*

$$\bigcup_{\lambda\in\Lambda}\bigcup_{\epsilon\geq 0}\{(v,\ v(a)+\epsilon-\lambda g(a))\ \mid\ v\in\partial_\epsilon(\lambda g)(a)\}$$
$$= \text{co cone}\left(\bigcup_{i\in I}\bigcup_{\epsilon\geq 0}\{(v_i,\ v_i(a)+\epsilon-g_i(a))\ \mid\ v_i\in\partial_\epsilon g_i(a)\}\right).$$

Proof. Let

$$C := \bigcup_{\lambda\in\Lambda}\bigcup_{\epsilon\geq 0}\{(v,\ v(a)+\epsilon-\lambda g(a))\ \mid\ v\in\partial_\epsilon(\lambda g)(a)\}.$$

Then

$$(u,\alpha)\in C \iff \begin{pmatrix} \exists\lambda=(\lambda_i)\in\Lambda,\ \exists\epsilon\geq 0 \text{ such that} \\ u\in\partial_\epsilon(\lambda g)(a)=\partial_\epsilon(\textstyle\sum_i\lambda_i g_i)(a) \\ \alpha = v(a)+\epsilon-\lambda g(a)\end{pmatrix}.$$

By Theorem 2.1 in [7], there exist $\epsilon_i \geq 0$ with $\sum_i \epsilon_i = \epsilon$ and $v_i \in \partial_{\epsilon_i}(\lambda_i g_i)(a)$ such that $u = \sum_i v_i$, and

$$\alpha = \sum_i v_i(a) + \sum_i \epsilon_i - \sum_i \lambda_i g_i(a) = \sum_{\lambda_i>0}\lambda_i[v_i'(a)+\epsilon_i'-g_i(a)].$$

Let $v_i' = \frac{v_i}{\lambda_i}$ and $\epsilon_i' = \frac{\epsilon_i}{\lambda_i}$ for i with $\lambda_i > 0$. Then it is easy to see that $v_i' \in \partial_{\epsilon_i'} g_i(a)$. Hence, $(u,\alpha)\in C$ if and only if

$$u = \sum_{\lambda_i>0}\lambda_i v_i' \quad\text{and}\quad \alpha = \sum_{\lambda_i>0}\lambda_i[v_i'(a)+\epsilon_i'-g_i(a)]$$

or, equivalently,

$$(u,\ \alpha)\in\text{co cone}\left(\bigcup_{i\in I}\bigcup_{\epsilon\geq 0}\{(v_i,\ v_i(a)+\epsilon-g_i(a))\ \mid\ v_i\in\partial_\epsilon g_i(a)\}\right). \qquad \square$$

THEOREM 5.2. *For the problem* (SIP)*, let $a \in A$. Then the following statements are equivalent*:

(i) *a is a minimizer of* (SIP).

(ii) *There exists $u \in \partial f(a)$ such that*

$$(-u,-u(a))\in\text{cl}\left(\text{co cone}\bigcup_{i\in I}\bigcup_{\epsilon\geq 0}\{(v_i,\ v_i(a)+\epsilon-g_i(a))\ \mid\ v_i\in\partial_\epsilon g_i(a)\}\right).$$

(iii) *There exist $u \in \partial f(a)$, $\{\lambda_k\} \subset \Lambda$, $\{\epsilon_k\} \subset \mathbb{R}_+$, and $\{v_k\} \subset \mathbb{R}^n$ such that $v_k \in \partial_{\epsilon_k}(\lambda_k g)(a)$, $u + v_k \to_* 0$, $\lambda_k g(a) \to 0$, and $\epsilon_k \to 0$ as $k \to \infty$.*

(iv) *There exist $u \in \partial f(a)$, $\{x_k\} \subset X$, $\{\lambda_k\} \subset \Lambda$, and $\{v_k\} \subset \mathbb{R}^n$ such that $v_k \in \partial(\lambda_k g)(x_k)$, $u + v_k \to_* 0$, $\|x_k - a\| \to 0$, and $\lambda_k g(x_k) \to 0$ as $k \to \infty$.*

Proof. The equivalence between (i) and (iii) (between (i) and (iv)) is a direct consequence of Theorem 3.1 (Theorem 3.2, respectively) while the equivalence between (i) and (ii) follows from Lemma 5.1 and (2). □

For the corresponding dual characterizations of set containments involving semi-infinite convex constraints, see [12]. The following example illustrates the significance of the sequential Lagrange multipliers for semi-infinite programming problems.

EXAMPLE 5.2. *Consider the problem*

$$\begin{array}{ll} \textit{Minimize} & x^2+y \\ \textit{subject to} & x \le 0 \\ & y \le 0, \\ & \frac{x}{i} - y \le 0,\ i = 3,4,5,\dots . \end{array}$$

Let $f(x,y) := x^2+y,\ g_1(x,y) := x,\ g_2(x,y) := y,\ g_i(x,y) = \frac{x}{i} - y$ *for all* $i = 3,4,5,\dots$, *and let* $g := (g_i)_{i\in\mathbb{N}}$. *The feasible set of the problem is*

$$A = \{(x,y) \mid g_i(x,y) \le 0\ \forall i \in \mathbb{N}\} = \{(x,y) \in \mathbb{R}^2 \mid x \le 0, y = 0\}$$

and $a = (0,0)$ *is a minimizer of the problem. Note that* $\partial f(0,0) = \{(0,1)\}$ *and that there is no* $(x_0,y_0) \in \mathbb{R}^2$ *satisfying* $g_i(x_0,y_0) < 0$ *for all* $i \in \mathbb{N}$. *Then for each* $\epsilon_n \ge 0$,

$$\partial_{\epsilon_n}(\bar\lambda_n g)(a) = \lambda_1^n(1,0) + \lambda_2^n(0,1) + \sum_{i\ge 3} \lambda_i^n \left(\frac{1}{i}, -1\right).$$

Consequently,

$$v_n = (v_1^n, v_2^n) \in \partial_{\epsilon_n}(\bar\lambda_n g)(a) \iff \begin{cases} v_1^n = \lambda_1^n + \sum_{i\ge 3} \lambda_i^n \frac{1}{i}, \\ v_2^n = \lambda_2^n - \sum_{i\ge 3} \lambda_i^n. \end{cases}$$

Let us take $\bar\lambda_1 = \bar\lambda_2 = (0,0,\dots,0,\dots)$, $\bar\lambda_n = (0,0,\dots,0,1+\frac{1}{n},0,\dots),\ n \ge 3$, *where the only nonzero component* $1+\frac{1}{n}$ *is at the* n*th position. Then for each* $\{\epsilon_n\} \subset \mathbb{R}_+,\ \epsilon_n \to 0$,

$$v_n \in \partial_{\epsilon_n}(\bar\lambda_n g)(0,0) \iff v_n = \left(\frac{1+\frac{1}{n}}{n}, -1 - \frac{1}{n}\right).$$

Hence $-v_n \to (0,1) \in \partial f(0,0)$. *Clearly,* $\bar\lambda_n g(a) = 0$ *for all* $n \in \mathbb{N}$.

5.3. Semiconvex programs. We now see that the approach developed in the previous sections can easily be extended to a larger class of nonsmooth problems with convex constraints. Let us now assume that $f : X \to \mathbb{R}$ is locally Lipschitz and $a \in X$. The function f is said to be *semiconvex* at a (see [18]) if, for any $u \in \partial^c f(a)$, $u(x-a) \ge 0$ implies that $f(x) \ge f(a)$, where $\partial^c f(a)$ stands for the Clarke subdifferential [5] of f at a. Observe that the fractional functions of the form $f(x) = \frac{p(x)}{q(x)}$, where p, q are locally Lipschitz, $p \ge 0$, convex, and $q > 0$ and concave, satisfy the semiconvexity property. Moreover, the notion of semiconvexity collapses to the notion of pseudoconvexity when f is continuously Fréchet differentiable.

THEOREM 5.3. *For the problem* (P), *assume that* f *is locally Lipschitz and semiconvex at a point* $a \in A$ *and* g *is continuous and* S*-convex. Then* a *is a minimizer for* (P) *if and only if there exist* $u \in \partial^c f(a)$, $\{\epsilon_n\} \subset \mathbb{R}_+$, $\{\lambda_n\} \subset S^+$, *and* $\{v_n\} \subset X'$ *such that* $v_n \in \partial_{\epsilon_n}(\lambda_n g)(a))$, $u + v_n \to_* 0$, $\lambda_n g(a) \to 0$, *and* $\epsilon_n \to 0$ *as* $n \to \infty$.

Moreover, if the closed cone constraint qualification holds, then a is a minimizer for (P) *if and only if there is* $\lambda \in S^+$ *such that* $0 \in \partial^c f(a) + \partial(\lambda g)(a)$ *and* $\lambda g(a) = 0$.

Proof. Since f is locally Lipschitz, it follows from a theorem of Clarke (see [5], p. 52) that if a is a minimizer of f over A, then $0 \in \partial^c f(a) + N_A(a)$, where $N_A(a)$ is the Clarke normal cone at a which coincides with the normal cone of A at a in the sense of convex analysis since A is closed and convex. This means that if the point a is a minimizer for (P), then there exists $u \in \partial^c f(a)$ such that

$$x \in A \quad \Longrightarrow \quad u(x - a) \geq 0.$$

Under the semiconvexity of f, this is also a sufficient condition for optimality. In other words, a is a minimizer of (P) if and only if there exists $u \in \partial^c f(a)$ such that

$$x \in g^{-1}(-S) \Rightarrow u(x) \geq u(a).$$

The rest of the proof follows by using the same argument as in the proofs of Theorems 3.1 and 4.1 and so is omitted here. ∎

Acknowledgment. The authors are grateful to the referees for their constructive suggestions and are thankful to Professor Nguyen Dong Yen for his comments on the preliminary version of the paper.

REFERENCES

[1] H. Attouch, J.-B. Baillon, and M. Thera, *Variational sum of monotone operators*, J. Convex Anal., 1 (1994), pp. 1–29.
[2] J. M. Borwein and H. Wolkowicz, *Characterizations of optimality without constraint qualification for the abstract convex program*, Math. Programming Stud., 19 (1982), pp. 77–100.
[3] J. M. Borwein and H. Wolkowicz, *Characterizations of optimality for the abstract convex program with finite dimensional range*, J. Austral. Math. Soc. Ser. A, 30 (1981), pp. 390–411.
[4] A. Brondsted and R. T. Rockafellar, *On the subdifferential of convex functions*, Proc. Amer. Math. Soc., 16 (1965), pp. 605–611.
[5] F. H. Clarke, *Optimization and Nonsmooth Analysis*, John Wiley & Sons, New York, 1983.
[6] M. A. Goberna and M. A. Lopez, *Linear Semi-infinite Optimization*, Wiley Ser. Math. Methods Pract. 2, John Wiley & Sons, Chichester, 1998.
[7] J. B. Hiriart-Urruty, *ϵ-subdifferential*, in Convex Analysis and Optimization, J. P. Aubin and R. Vinter, eds., Pitman, London, 1982, pp. 43–92.
[8] J. B. Hiriart-Urruty and C. Lemarechal, *Convex Analysis and Minimization Algorithms*, Vol. I, Springer-Verlag, Berlin, Heidelberg, 1993.
[9] J. B. Hiriart-Urruty and C. Lemarechal, *Convex Analysis and Minimization Algorithms*, Vol. II, Springer-Verlag, Berlin, Heidelberg, 1993.
[10] J. B. Hiriart-Urruty and R. R. Phelps, *Subdifferential calculus using ϵ-subdifferentials*, J. Funct. Anal., 18 (1993), pp. 154–166.
[11] R. B. Holmes, *Geometric Functional Analysis*, Springer-Verlag, Berlin, 1975.
[12] V. Jeyakumar, *Characterizing set containments involving infinite convex constraints and reverse-convex constraints*, SIAM J. Optim., 13 (2003), pp. 947–959.
[13] V. Jeyakumar, *Asymptotic dual conditions characterizing optimality for convex programs*, J. Optim. Theory Appl., 93 (1997), pp. 153–165.
[14] V. Jeyakumar and M. Nealon, *Complete dual characterizations of optimality for convex semidefinite programming*, in Constructive, Experimental, and Nonlinear Analysis (Limoges, 1999), CMS Conf. Proc. 27, AMS, Providence, RI, 2000, pp. 165–173.
[15] V. Jeyakumar, A. M. Rubinov, B. M. Glover, and Y. Ishizuka, *Inequality systems and global optimization*, J. Math. Anal. Appl., 202 (1996) pp. 900–919.
[16] V. Jeyakumar and H. Wolkowicz, *Generalizations of Slater's constraint qualification for infinite convex programs*, Math. Programming, 57 (1992), pp. 85–102.
[17] V. Jeyakumar and A. Zaffaroni, *Asymptotic conditions for weak and proper optimality in infinite dimensional convex vector optimization*, Numer. Funct. Anal. Optim., 17 (1996), pp. 323–343.

[18] R. Mifflin, *Semismooth and semiconvex functions in constrained optimization*, SIAM J. Control Optim., 15 (1977), pp. 959–972.

[19] M. V. Ramana, L. Tunçel, and H. Wolkowicz, *Strong duality for semidefinite programming*, SIAM J. Optim., 7 (1997), pp. 641–662.

[20] L. Thibault, *Sequential convex subdifferential calculus and sequential Lagrange multipliers*, SIAM. J. Control Optim., 35 (1997), pp. 1434–1444.

[21] L. Thibault, *Limiting convex subdifferential calculus with applications to integration and maximal monotonicity of subdifferential*, in Constructive, Experimental, and Nonlinear Analysis (Limoges, 1999), CMS Conf. Proc. 27, AMS, Providence, RI, 2000, pp. 279–289.

[22] H. Wolkowicz, R. Saigal, and L. Vandenberghe, *Handbook of Semi-definite Programming*, Internat. Ser. Oper. Res. Management Sci. 27, Kluwer Academic, Dordrecht, The Netherlands, 2000.

SIAM J. OPTIM.
Vol. 14, No. 2, pp. 548–566

OPTIMIZATION WITH STOCHASTIC DOMINANCE CONSTRAINTS*

DARINKA DENTCHEVA† AND ANDRZEJ RUSZCZYŃSKI‡

Abstract. We introduce stochastic optimization problems involving stochastic dominance constraints. We develop necessary and sufficient conditions of optimality and duality theory for these models and show that the Lagrange multipliers corresponding to dominance constraints are concave nondecreasing utility functions. The models and results are illustrated on a portfolio optimization problem.

Key words. stochastic programming, stochastic dominance, partial orders, optimality conditions, duality

AMS subject classifications. Primary, 90C15, 90C46, 90C48; Secondary, 46N10, 60E15, 91B06

DOI. S1052623402420528

1. Introduction.

The relation of *stochastic dominance* is a fundamental concept of decision theory and economics (see [11, 12, 26, 33]). A random variable X *dominates* another random variable Y in the second order, which we write as $X \succeq_{(2)} Y$, if $\mathbb{E}[u(X)] \geq \mathbb{E}[u(Y)]$ for every concave nondecreasing function $u(\cdot)$, for which these expected values are finite. We refer the reader to the monograph [20] for a modern view on the stochastic dominance relation and other comparison methods for random outcomes.

The main objective of this paper is to introduce a new stochastic optimization model involving dominance relations as constraints and to analyze its properties. We concentrate on the second-order dominance constraints, as they are the most important from the theoretical point of view. Further we extend our analysis to dominance relations of higher orders.

A basic model of stochastic optimization can be formulated as follows:

$$\max_{z \in Z} \mathbb{E}[\varphi(z, \omega)]. \tag{1.1}$$

In this formulation ω denotes an elementary event in a probability space $(\Omega, \mathcal{F}, P)$, z is a decision vector in an appropriate space $\mathcal{Z}$, and $\varphi : \mathcal{Z} \times \Omega \to \mathbb{R}$. The set $Z \subset \mathcal{Z}$ is defined either explicitly, or via some constraints that may involve the elementary event ω and must hold with some prescribed probability.

The first stochastic optimization models with expected values were introduced in [1, 6]. Mathematical theory of expectation models involving two-stage and multistage decisions has been developed in [37, 38] and in [30, 31, 32]. A comprehensive treatment of the theory and numerical methods for expectation models can be found in [3].

Models involving constraints on probability were introduced [5, 18, 24]. The book [25] discusses in detail the theory and numerical methods for linear models with one

*Received by the editors December 28, 2002; accepted for publication (in revised form) June 16, 2003; published electronically November 6, 2003. This work was supported by National Science Foundation awards 0303545 and 0303728.

http://www.siam.org/journals/siopt/14-2/42052.html

†Stevens Institute of Technology, Department of Mathematical Sciences, Castle Point on Hudson, Hoboken, NJ 07030 (ddentche@stevens-tech.edu).

‡Rutgers University, Department of Management Science and Information Systems and RUTCOR, 94 Rockafeller Rd., Piscataway, NJ 08854 (rusz@rutcor.rutgers.edu).

probabilistic constraint on finitely many inequalities.

Another way to look at problem (1.1) is to consider the set C of random variables X such that, for some $z \in Z$, one has $X(\omega) \leq \varphi(z, \omega)$ a.s. Then we can write the model as

$$\max_{X \in C} \mathbb{E}[X].$$

Von Neumann and Morgenstern, in their book [36], introduced the *expected utility hypothesis*: for every rational decision maker there exists a utility function $u(\cdot)$ such that she prefers outcome X over outcome Y if and only if $\mathbb{E}[u(X)] > \mathbb{E}[u(Y)]$. Therefore the decision maker solves the following optimization problem:

$$\max_{X \in C} \mathbb{E}[u(X)].$$

In practice, however, it is almost impossible to elicit the utility function of a decision maker explicitly. Additional difficulties arise when there is a group of decision makers with different utility functions who have to come to a consensus.

In some applications a reference outcome Y in $\mathcal{L}^1(\Omega, \mathcal{F}, P)$ is available. It may have the form $Y(\omega) = \varphi(\bar{z}, \omega)$, $\omega \in \Omega$, for some policy $\bar{z}$. Our intention is to have the new outcome, X, preferable over Y. Therefore, we introduce the following optimization problem:

$$\max\ f(X) \tag{1.2}$$

$$\text{subject to } X \succeq_{(2)} Y, \tag{1.3}$$

$$X \in C. \tag{1.4}$$

Here Y is a random variable in $\mathcal{L}^1(\Omega, \mathcal{F}, P)$, the set $C \subset \mathcal{L}^1(\Omega, \mathcal{F}, P)$ is convex and closed, and $f : C \to \mathbb{R}$ is a concave continuous functional. Constraint (1.3) guarantees that for any decision maker, whose utility function $u(\cdot)$ is concave and nondecreasing, the solution X of the problem will satisfy the relation $\mathbb{E}[u(X)] \geq \mathbb{E}[u(Y)]$.

Another class of models that recently attracted much attention are mean-risk models. In our notation they take the form

$$\max_{X \in C} \Big\{ \mathbb{E}[X] - \lambda \rho(X) \Big\}.$$

In this problem $\lambda > 0$ and $\rho(\cdot)$ is a risk functional which depends on the entire distribution of X and assigns to it a scalar measure of its variability. For example, the expected shortfall below the mean,

$$\rho(X) = \mathbb{E}\Big[(\mathbb{E}[X] - X)_+ \Big],$$

may be used as the risk functional. Here $(X)_+ = \max(0, X)$. Mean-risk models are also closely related to stochastic dominance relations. If we use an appropriate risk measure ρ and the parameter λ is within a certain range, then the optimal outcome $\hat{X}$ is not stochastically dominated by any other feasible outcome (see [21, 22, 23]). Other stochastic optimization models involving general risk functionals were considered by [35, 13, 27, 29].

Our model (1.2)–(1.4) is a new way to formulate a stochastic optimization problem.

Example 1. Let $R_1, \dots, R_N \in \mathcal{L}^1(\Omega, \mathcal{F}, P)$ be random returns of assets $1, \dots, N$. Our aim is to invest our capital in these assets in order to obtain some desirable characteristics of the total return on the investment. Denoting by $z_1, \dots, z_N$ the fractions of the initial capital invested in assets $1, \dots, N$, we can easily derive the formula for the total return:

$$X = R_1 z_1 + \cdots + R_N z_N.$$

Let $Z = \{z \in \mathbb{R}^N : z_1 + \cdots + z_N = 1,\ z_n \geq 0,\ n = 1, \dots, N\}$ be the set of possible asset allocations. Clearly, the set C of portfolio returns is just the convex hull of $R_1, \dots, R_N$. The problem of maximizing

$$f(X) = \mathbb{E}X$$

in C has a trivial and meaningless solution: invest everything in asset(s) having the highest expected return. Our model (1.2)–(1.4) approaches the problem of selecting the most preferred portfolio in a new way: we have some *reference* random return Y and require our outcome to dominate Y. For example, the reference return may be the return of an existing portfolio or a market index. In this way we guarantee that no risk-averse decision maker will prefer Y over the optimal solution of (1.2)–(1.4). We thus provide an alternative approach to mean-risk portfolio models (see, e.g., [16, 17, 34]).

Problem (1.2)–(1.4) is interesting from the mathematical point of view. It involves a new form of constraint which has not been explored in the stochastic optimization theory. Our analysis will shed more light on the place of this model in the general optimization theory.

Moreover, our model is relevant for economics. We show that the Lagrange multiplier associated with the dominance constraint can be identified with a certain concave and nondecreasing utility function.

In section 2 we formally define the stochastic dominance relations and analyze properties of the set defined by this relation. In section 3 we consider equivalent formulations of dominance-constrained optimization problems. Section 4 is devoted to the derivation of necessary and sufficient conditions of optimality. In section 5 we formulate duality relations.

In sections 6 and 7 we extend our theory to multiple dominance constraints and to dominance constraints of higher orders. Finally, we provide a numerical illustration in section 8.

2. The dominance constraint. Consider a random variable $X \in \mathcal{L}^1(\Omega, \mathcal{F}, P)$ and its distribution function

$$F(X; \eta) = P[X \leq \eta] \quad \text{for } \eta \in \mathbb{R}.$$

Define the function $F_2(X; \cdot)$ as

$$F_2(X; \eta) = \int_{-\infty}^{\eta} F(X; \alpha)\, d\alpha \quad \text{for } \eta \in \mathbb{R}. \tag{2.1}$$

As an integral of a nondecreasing function, it is a convex function of η.

Furthermore, for $X \in \mathcal{L}^m(\Omega, \mathcal{F}, P)$ we can define recursively the functions

$$F_k(X; \eta) = \int_{-\infty}^{\eta} F_{k-1}(X; \alpha)\, d\alpha \quad \text{for } \eta \in \mathbb{R}, \quad k = 3, \dots, m+1. \tag{2.2}$$

They are also convex and nondecreasing functions of the second argument.

DEFINITION 2.1. *We say that a random variable* $X \in \mathcal{L}^{k-1}(\Omega, \mathcal{F}, P)$ dominates *in the* k*th order another random variable* $Y \in \mathcal{L}^{k-1}(\Omega, \mathcal{F}, P)$ *if*

$$F_k(X;\eta) \le F_k(Y;\eta) \text{ for all } \eta \in \mathbb{R}. \tag{2.3}$$

We shall denote relation (2.3) as

$$X \succeq_{(k)} Y \tag{2.4}$$

and the set of X satisfying this relation as

$$A_k(Y) = \{X \in \mathcal{L}^{k-1}(\Omega, \mathcal{F}, P) : X \succeq_{(k)} Y\}. \tag{2.5}$$

For every k the stochastic dominance relation "$\succeq_{(k)}$" introduces a partial order among random variables in $\mathcal{L}^{k-1}(\Omega, \mathcal{F}, P)$ (see, e.g., [19] and the references therein). Partial orders appear in abstract optimization problems when the values of the objective operator are elements of a topological vector space (see, e.g., [15]). It is usually assumed that the partial order is generated by a convex cone. The stochastic dominance orders in $\mathcal{L}^k(\Omega, \mathcal{F}, P)$ are not generated by cones in this space, as we shall see in Proposition 2.4.

By definition, the kth-order dominance implies the $(k+1)$st-order dominance if the random variables in question are in $\mathcal{L}^k$.

Most important is the second-order dominance relation, because of its connections with risk-averse preferences, as described below (see also [9, 14]).

Changing the order of integration in (2.1) we get (see, e.g., [21])

$$F_2(X;\eta) = \mathbb{E}[(\eta - X)_+]. \tag{2.6}$$

Therefore, an equivalent representation of the second-order stochastic dominance relation is

$$\mathbb{E}[(\eta - X)_+] \le \mathbb{E}[(\eta - Y)_+] \quad \text{for all } \eta \in \mathbb{R}. \tag{2.7}$$

Let us consider the set $\mathcal{U}$ of concave nondecreasing functions $u : \mathbb{R} \to \mathbb{R}$ satisfying the following linear growth condition:

$$\lim_{t \to -\infty} u(t)/t < \infty. \tag{2.8}$$

For every random variable $X \in \mathcal{L}^1(\Omega, \mathcal{F}, P)$ and for every $u \in \mathcal{U}$ the quantity

$$\mathbb{E}[u(X)] = \int u(X(\omega))\, dP(\omega)$$

is well-defined and finite.

PROPOSITION 2.2. *For each* $X, Y \in \mathcal{L}^1(\Omega, \mathcal{F}, P)$ *the relation* $X \succeq_{(2)} Y$ *is equivalent to*

$$\mathbb{E}[u(X)] \ge \mathbb{E}[u(Y)] \quad \text{for all } u \in \mathcal{U}. \tag{2.9}$$

Proof. Suppose that $X \succeq_{(2)} Y$. It follows from (2.7) that for every function of form

$$u_N(x) = c_N - \sum_{i=1}^{N} \alpha_i (\eta_i - x)_+, \tag{2.10}$$

where $\alpha_i \geq 0$, $i = 1, \ldots, N$, we have

$$\mathbb{E}[u_N(X)] \geq \mathbb{E}[u_N(Y)]. \tag{2.11}$$

Let $u \in \mathcal{U}$. We shall construct a sequence of functions $\{u_N\}$ of the form (2.10).

For an integer M we introduce $2M^2 + 1$ discretization points

$$\eta_i = i/M, \quad -M^2 \leq i \leq M^2,$$

and we define

$$c_M = u(M),$$
$$\alpha_i = \begin{cases} \frac{u(\eta_i)-u(\eta_{i-1})}{\eta_i-\eta_{i-1}} & \text{for } i = M^2, \\ \frac{u(\eta_i)-u(\eta_{i-1})}{\eta_i-\eta_{i-1}} - \frac{u(\eta_{i+1})-u(\eta_i)}{\eta_{i+1}-\eta_i} & \text{for } i = M^2-1, M^2-2, \ldots, -M^2+1, \\ u(\eta_i) - u(\eta_i - 1) - \frac{u(\eta_{i+1})-u(\eta_i)}{\eta_{i+1}-\eta_i} & \text{for } i = -M^2. \end{cases}$$

By construction, the function

$$u_M(x) = c_M - \sum_{i=-M^2}^{M^2} \alpha_i(\eta_i - x)_+$$

is a piecewise linear approximation of $u(\cdot)$ with nodes $(\eta_i, u(\eta_i))$, $i = -M^2, \ldots, M^2$. It is elementary to see that

$$\lim_{M\to\infty} \mathbb{E}[u_M(X)] = \mathbb{E}[u(X)] \quad \text{for all } X \in \mathcal{L}^1(\Omega, \mathcal{F}, P).$$

Therefore (2.11) implies (2.9). On the other hand, if (2.9) is true, then it is true for particular functions $u(x) = -(\eta - x)_+$. In view of (2.7), this implies that $X \succeq_{(2)} Y$. □

Let us now analyze the structure of the set

$$A_2(Y) = \{X \in \mathcal{L}^1(\Omega, \mathcal{F}, P) : X \succeq_{(2)} Y\}.$$

Recall that the *recession cone* of a convex set A in a vector space $\mathcal{S}$ is defined as the set $A^\infty = \{H \in \mathcal{S} : A + \tau H \subset A \text{ for all } \tau \geq 0\}$.

PROPOSITION 2.3. *For every $Y \in \mathcal{L}^1(\Omega, \mathcal{F}, P)$ the set $A_2(Y)$ is convex and closed. Furthermore, its recession cone has the form*

$$A_2^\infty(Y) = \{H \in \mathcal{L}^1(\Omega, \mathcal{F}, P) : H \geq 0 \text{ a.s.}\}.$$

Proof. Let us consider the equivalent representation (2.7) of the stochastic dominance constraint. For every $\eta \in \mathbb{R}$ the functional $X \to \mathbb{E}[(\eta - X)_+]$ is convex and continuous in $\mathcal{L}^1(\Omega, \mathcal{F}, P)$ as a composition of the "max" function and the expectation operator. Consequently, the set $A_2(Y)$ is convex and closed.

If $H \geq 0$ a.s., then the distribution functions satisfy the inequality $F(Y+\tau H; \cdot) \leq F(Y; \cdot)$ for all $\tau \geq 0$. Thus, $F_2(Y + \tau H; \cdot) \leq F_2(Y; \cdot)$ for all $\tau \geq 0$. Consequently, H belongs to the recession cone of $A_2(Y)$.

Suppose that $H \in A_2^\infty(Y)$ and $P[H < 0] > 0$. By the definition of the recession cone,

$$F_2(Y + \tau H; \eta) \leq F_2(Y; \eta) \quad \text{for all } \eta \in \mathbb{R} \text{ and } \tau \geq 0. \tag{2.12}$$

We shall show that this is impossible. Since $P[H < 0] > 0$, there exists $\varepsilon > 0$ such that $\delta := P[H \le -\varepsilon] > 0$. For every $\eta \in \mathbb{R}$ and every $\tau > 0$ we have

$$\begin{aligned} F_2(Y + \tau H; \eta) &= \mathbb{E}[(\eta - Y - \tau H)_+] \\ &\ge P[H \le -\varepsilon]\, \mathbb{E}[(\eta - Y - \tau H)_+ \,|\, H \le -\varepsilon] \\ &\ge \delta \mathbb{E}[(\eta - Y + \tau\varepsilon)_+ \,|\, H \le -\varepsilon]. \end{aligned}$$

For any η we can find $\tau_0 > 0$ such that $\mathbb{E}[(\eta - Y + \tau_0\varepsilon)_+ \,|\, H \le -\varepsilon] > 0$. The last displayed expression is a convex function of τ, and it is increasing for $\tau > \tau_0$. Therefore

$$\lim_{\tau\to\infty} F_2(Y + \tau H; \eta) \ge \delta \lim_{\tau\to\infty} \mathbb{E}[(\eta - Y + \tau\varepsilon)_+ \,|\, H \le -\varepsilon] = \infty,$$

which contradicts (2.12). Thus $H \notin A_2^\infty(Y)$. □

The set $A_2(Y)$ is not a cone pointed at Y unless the reference outcome is deterministic.

PROPOSITION 2.4. *$A_2(Y) = Y + A_2^\infty(Y)$ if and only if Y is constant a.s.*

Proof. If Y is constant a.s., then $X \in A_2(Y)$ implies $F_2(X; \mathbb{E}[Y]) \le F_2(Y; \mathbb{E}[Y]) = 0$. Thus $X \ge \mathbb{E}[Y]$ a.s., which means that $X - Y \ge 0$ a.s. By Proposition 2.3, the direction $X - Y$ is an element of $A_2^\infty(Y)$.

Let us now consider the case when $P[Y \ne \mathbb{E}Y] > 0$. We shall construct a direction H such that $Y + \tau H \in A_2(Y)$ for all $\tau \in [0, 1]$, but $H \notin A_2^\infty(Y)$. Let $Y' = \mathbb{E}[Y]$ a.s. Clearly, $Y' \in A_2(Y)$. Define $H = Y' - Y$. By the convexity of $A_2(Y)$, $Y + \tau H \in A_2(Y)$ for all $\tau \in [0, 1]$. We have $P[H < 0] = P[Y > \mathbb{E}[Y]] > 0$. If follows from Proposition 2.3 that $H \notin A_2^\infty(Y)$. Therefore $A_2(Y)$ is not a cone. □

3. The optimization problem. To overcome serious technical difficulties associated with the dominance constraint we shall consider a relaxed version of problem (1.2)–(1.4), in which the dominance relation (2.7) is enforced on an interval $[a, b]$:

$$\max\ f(X) \tag{3.1}$$

$$\text{subject to } \mathbb{E}[(\eta - X)_+] \le \mathbb{E}[(\eta - Y)_+] \quad \text{for all } \eta \in [a, b], \tag{3.2}$$

$$X \in C. \tag{3.3}$$

Clearly, if all $X \in C$ have uniformly bounded distributions, (3.2) is equivalent to (1.3) for appropriately chosen a and b. However, if the distributions are not uniformly bounded, (3.2) is a relaxation of (1.3).

Constraint (3.2) involves a nonsmooth operator from the space $\mathcal{L}^1(\Omega, \mathcal{F}, P)$ to the space of continuous functions on $[a, b]$. Another way to formulate the problem is to introduce a decision vector $S : [a, b] \times \Omega \to \mathbb{R}$ to represent the shortfall. We obtain the problem:

$$\max\ f(X) \tag{3.4}$$

$$\text{subject to } X(\omega) + S(\eta, \omega) \ge \eta, \qquad \text{for a.a. } (\eta, \omega) \in [a, b] \times \Omega, \tag{3.5}$$

$$\mathbb{E}[S(\eta, \omega)] \le \mathbb{E}[(\eta - Y)_+] \qquad \text{for all } \eta \in [a, b], \tag{3.6}$$

$$S(\eta, \omega) \ge 0, \qquad \text{for a.a. } (\eta, \omega) \in [a, b] \times \Omega, \tag{3.7}$$

$$X \in C. \tag{3.8}$$

The abbreviation "a.a." is understood as "almost all with respect to the product of the Lebesgue measure on $[a, b]$ and the probability measure P on Ω." To complete the

definition of (3.4)–(3.8) we need to specify the space Σ of functions, from which S is to be selected. We assume that Σ is the space of all S such that $S(\cdot,\omega)$ is continuous for P-almost all ω and $S(\eta,\cdot)$ is integrable for all $\eta \in [a,b]$.

PROPOSITION 3.1. (i) *For every optimal solution $\hat{X}$ of* (3.1)–(3.3)*, the pair $(\hat{X},\hat{S})$, with $\hat{S}(\eta,\omega) = \max(0,\eta - \hat{X}(\omega))$, is an optimal solution of* (3.4)–(3.7)*.*

(ii) *For every optimal solution $(\hat{X},\hat{S})$ of* (3.4)–(3.8) *the point $\hat{X}$ is an optimal solution of* (3.1)–(3.3)*.*

The equivalence of the two formulations is evident and the proof is omitted.

If the reference point Y has a discrete distribution with finitely many realizations, both formulations simplify substantially.

PROPOSITION 3.2. *Assume that Y has a discrete distribution with realizations y_i, $i = 1,\dots,m$, where $a \le y_i \le b$ for all i. Then inequalities* (3.2) *are equivalent to*

$$\mathbb{E}[(y_i - X)_+] \le \mathbb{E}[(y_i - Y)_+], \quad i = 1,\dots,m. \tag{3.9}$$

Proof. With no loss of generality we may assume that $y_1 < y_2 < \cdots < y_m$. It is sufficient to prove that (3.9) imply that

$$F_2(X;\eta) \le F_2(Y;\eta) \quad \text{for all } \eta \in \mathbb{R}.$$

The function $F_2(Y;\cdot)$ is piecewise linear and has break points at y_i, $i = 1,\dots,m$. Let us consider three cases, depending on the value of η.

Case 1. If $\eta \le y_1$, we have

$$0 \le F_2(X;\eta) \le F_2(X;y_1) \le F_2(Y;y_1) = 0.$$

Therefore the required relation holds as an equality.

Case 2. Let $\eta \in [y_i, y_{i+1}]$ for some i. Since, for any X, the function $F_2(X;\cdot)$ is convex, inequalities (3.9) for i and $i+1$ imply that for all $\eta \in [y_i, y_{i+1}]$ one has

$$\begin{aligned} F_2(X;\eta) &\le \lambda F_2(X;y_i) + (1-\lambda) F_2(X;y_{i+1}) \\ &\le \lambda F_2(Y;y_i) + (1-\lambda) F_2(Y;y_{i+1}) = F_2(Y;\eta), \end{aligned}$$

where $\lambda = (y_{i+1} - \eta)/(y_{i+1} - y_i)$.

Case 3. For $\eta > y_m$ we have

$$\begin{aligned} F_2(Y;\eta) &= F_2(Y;y_m) + \eta - y_m \\ &\ge F_2(X;y_m) + \int_{y_m}^{\eta} F(X;\alpha)\, d\alpha = F_2(X;\eta), \end{aligned}$$

as required. □

If the entire space Ω has finitely many elementary events $\omega_1,\dots,\omega_m$, formulation (3.4)–(3.8) simplifies even further. Let $p_k = P[\{\omega_k\}]$, $v_i = \mathbb{E}[(y_i - Y)_+]$, and $x_k = X(\omega_k)$, $s_{ik} = S(y_i,\omega_k)$. Then the following finite system of linear inequalities is equivalent to (3.5)–(3.7):

$$x_k + s_{ik} \ge y_i, \quad i = 1,\dots,m, \quad k = 1,\dots,m, \tag{3.10}$$

$$\sum_{k=1}^{m} p_k s_{ik} \le v_i, \quad i = 1,\dots,m, \tag{3.11}$$

$$s_{ik} \ge 0, \quad i = 1,\dots,m, \quad k = 1,\dots,m. \tag{3.12}$$

This formulation can be used for numerical solution of dominance-constrained problems, as shown in section 8.

4. Optimality. We start from a specific form of constraint qualification for dominance constraints.

DEFINITION 4.1. *Problem* (3.1)–(3.3) *satisfies the* uniform dominance condition *if there exists a point* $\tilde{X} \in C$ *such that*

$$\inf_{\eta\in[a,b]} \Big\{ F_2(Y;\eta) - F_2(\tilde{X};\eta) \Big\} > 0.$$

We define the set $\mathcal{U}_1$ of functions $u(\cdot)$ satisfying the following conditions:

$$\begin{aligned} &u(\cdot) \text{ is concave and nondecreasing;}\\ &u(t) = 0 \text{ for all } t \geq b;\\ &u(t) = u(a) + c(t-a), \text{ with some } c > 0, \text{ for all } t \leq a. \end{aligned}$$

Clearly, $\mathcal{U}_1 \subset \mathcal{U}$. It is also evident that $\mathcal{U}_1$ is a convex cone.

Let us define the Lagrangian of (3.1)–(3.3), $L : C \times \mathcal{U}_1 \to \mathbb{R}$, as follows:

$$L(X,u) = f(X) + \mathbb{E}[u(X)] - \mathbb{E}[u(Y)]. \tag{4.1}$$

It is well-defined, because for every u in $\mathcal{U}_1$ and every $X \in \mathcal{L}^1(\Omega,\mathcal{F},P)$ the expected value $\mathbb{E}[u(X)]$ exists and is finite.

THEOREM 4.2. *Assume that the uniform dominance condition is satisfied. If* $\hat{X}$ *is an optimal solution of* (3.1)–(3.3)*, then there exists a function* $\hat{u} \in \mathcal{U}_1$ *such that*

$$L(\hat{X},\hat{u}) = \max_{X\in C} L(X,\hat{u}) \tag{4.2}$$

and

$$\mathbb{E}[\hat{u}(\hat{X})] = \mathbb{E}[\hat{u}(Y)]. \tag{4.3}$$

Conversely, if for some function $\hat{u} \in \mathcal{U}_1$ *an optimal solution* $\hat{X}$ *of* (4.2) *satisfies* (3.2) *and* (4.3)*, then* $\hat{X}$ *is an optimal solution of* (3.1)–(3.3)*.*

Proof. Let us rewrite (3.1)–(3.3) in the general form:

$$\begin{gathered} \max f(X)\\ \text{subject to } G(X) \in K,\\ X \in C, \end{gathered}$$

where $G : \mathcal{L}^1(\Omega,\mathcal{F},P) \to \mathcal{C}([a,b])$ is defined as

$$G(X)(\eta) := F_2(Y;\eta) - F_2(X;\eta), \quad \eta \in [a,b].$$

The set K is the cone of nonnegative functions in $\mathcal{C}([a,b])$. The operator G is concave with respect to the cone K; that is, for any X_1, X_2 in $\mathcal{L}^1(\Omega,\mathcal{F},P)$ and all $\lambda \in [0,1]$,

$$G(\lambda X_1 + (1-\lambda)X_2) - [\lambda G(X_1) + (1-\lambda)G(x_2)] \in K.$$

By the Riesz representation theorem, the space dual to $\mathcal{C}([a,b])$ is the space $\mathbf{rca}([a,b])$ of regular countably additive measures on $[a,b]$ having finite variation (see, e.g., [8]). We introduce the Lagrangian $\Lambda : C \times \mathbf{rca}([a,b]) \to \mathbb{R}$,

$$\Lambda(X,\mu) = f(X) + \int_a^b G(X)(\eta)\, d\mu(\eta). \tag{4.4}$$

Let us observe that the uniform dominance condition implies that the following generalized Slater condition is satisfied:

$$G(\tilde{X}) \in \operatorname{int} K.$$

Moreover, $\tilde{X} \in C$. By [4, Prop. 2.106], this is equivalent to the regularity condition:

$$0 \in \operatorname{int}[G(C) - K].$$

Therefore we can use the necessary conditions of optimality in abstract spaces (see, e.g., [4, Thm. 3.4]). We conclude that there exists a nonnegative measure $\hat{\mu} \in \mathbf{rca}([a,b])$ such that

$$\Lambda(\hat{X}, \hat{\mu}) = \max_{X \in C} \Lambda(X, \hat{\mu}) \tag{4.5}$$

and

$$\int_a^b [F_2(Y;\eta) - F_2(\hat{X};\eta)]\, d\hat{\mu}(\eta) = 0. \tag{4.6}$$

We shall transform these conditions to (4.2)–(4.3).

Every measure μ on $[a,b]$ can be extended to the whole real line by assigning measure 0 to Borel sets not intersecting $[a,b]$. Let us choose $M > 0$ such that $-M < a$. By changing the order of integration we obtain

$$\begin{aligned}
\int_a^b F_2(X;\eta)\, d\mu(\eta) &= \int_{-M}^b F_2(X;\eta)\, d\mu(\eta) \\
&= \int_{-M}^b \int_{-M}^{\eta} F(X;t)\, dt\, d\mu(\eta) + F_2(X;-M)\mu([a,b]) \\
&= \int_{-M}^b \int_t^b d\mu(\eta) F(X;t)\, dt + F_2(X;-M)\mu([a,b]) \\
&= \int_{-M}^b \mu([t,b])\, F(X;t)\, dt + F_2(X;-M)\mu([a,b]). \qquad (4.7)
\end{aligned}$$

A function $u : \mathbb{R} \to \mathbb{R}$ can be associated with every nonnegative measure μ as follows:

$$u(t) = \begin{cases} -\displaystyle\int_t^b \mu([\tau,b])\, d\tau, & t < b, \\ 0, & t \geq b. \end{cases}$$

Since $\mu \geq 0$, the function $\mu([\cdot,b])$ is nonnegative and nonincreasing, which implies that $u(\cdot)$ is nondecreasing and concave. We can rewrite (4.7) as

$$\int_a^b F_2(X;\eta)\, d\mu(\eta) = \int_{-M}^b F(X;t)\, du(t) + F_2(X;-M)\, \mu([a,b]).$$

Since $u(\cdot)$ is absolutely continuous and $F(X;\cdot)$ is monotone, we can integrate by parts to obtain

$$\begin{aligned}
\int_{-M}^b F(X;t)\, du(t) &= -F(X;-M)\, u(-M) - \int_{-M}^b u(t)\, dF(X;t) \\
&= -P[X \leq -M]\, u(-M) - \mathbb{E}[u(X)] + \mathbb{E}\Big[u(X) \mathbb{1}_{\{X \leq -M\}}\Big].
\end{aligned}$$

Putting together the last two equations we get

$$\int_a^b F_2(X;\eta)\,d\mu(\eta) = -\,\mathbb{E}[u(X)] - P[X \le -M]\,u(-M) + \mathbb{E}\Big[u(X)\mathbb{1}_{\{X\le -M\}}\Big] + F_2(X;-M)\,\mu([a,b]). \tag{4.8}$$

Let $M \to \infty$. Since $\mathbb{E}|X| < \infty$ and $|u(t)|$ grows linearly as $t \to -\infty$, we get

$$\lim_{M\to\infty} \mathbb{E}\Big[u(X)\mathbb{1}_{\{X\le -M\}}\Big] = 0.$$

By the monotonicity of u,

$$|u(-M)|\,P[X \le -M] \le \mathbb{E}\Big[|u(X)|\mathbb{1}_{\{X\le -M\}}\Big] \to 0 \text{ as } M \to \infty.$$

Consequently, (4.8) becomes

$$\int_a^b F_2(X;\eta)\,d\mu(\eta) = -\mathbb{E}[u(X)]. \tag{4.9}$$

In this way we have established a correspondence between nonnegative measures in $\mathbf{rca}([a,b])$ and functions in $\mathcal{U}_1$. Thus, the measure $\hat\mu$ corresponds to a function $\hat u \in \mathcal{U}_1$, condition (4.5) implies (4.2), and condition (4.6) implies (4.3).

Let us now prove the converse. If $u \in \mathcal{U}_1$, then the left derivative of u,

$$u'_-(t) = \lim_{\tau\uparrow t}[u(t) - u(\tau)]/(t-\tau),$$

is well-defined, nonincreasing, and continuous from the left. By Theorem 12.4 of [2], after an obvious adaptation, there exists a unique regular nonnegative measure μ satisfying

$$\mu([t,b]) = u'_-(t).$$

Thus the correspondence between nonnegative measures in $\mathbf{rca}([a,b])$ and functions in $\mathcal{U}_1$ is a bijection and formula (4.9) is always valid. Therefore, the maximizer $\hat X$ of (4.2) is also the maximizer of $\Lambda(X,\hat\mu)$, where $\hat\mu$ is derived from $\hat u$ in the way described above. It follows from sufficient conditions of optimality (see, e.g., [4, Prop. 3.3]) that if $\hat X$ satisfies (3.2) and (4.3), then it is optimal for (3.1)–(3.3). □

Remark 1. It is known (see, e.g., [4, Thm. 3.6]) that the set of Lagrange multipliers $\hat\mu$ corresponding to the Lagrangian Λ is convex, bounded, and weakly* closed in the dual space $\mathbf{rca}([a,b])$. Moreover, the same set of Lagrange multipliers corresponds to every optimal solution of (3.1)–(3.3). Therefore, the set $\hat U$ of functions in $\mathcal{U}_1$ satisfying (4.2)–(4.3) is convex, bounded, and weakly* closed in the following sense: if a sequence of functions $u^k \in \hat U$ and $u \in \mathcal{U}_1$ are such that

$$\lim_{k\to\infty} \mathbb{E}[u^k(X)] = \mathbb{E}[u(X)] \quad \text{for all } X \in \mathcal{L}^1(\Omega,\mathcal{F},P),$$

then $u \in \hat U$. The set $\hat U$ is the same for all optimal solutions of (3.1)–(3.3). These statements are derived by the application of the key equation (4.9).

5. Duality. For every function $u \in \mathcal{U}_1$ the problem

$$\max_{X \in C} \Big\{ f(X) + \mathbb{E}[u(X)] - \mathbb{E}[u(Y)] \Big\} \tag{5.1}$$

is a Lagrangian relaxation of problem (3.1)–(3.3). Its optimal value

$$D(u) = \sup_{X \in C} L(X, u)$$

is always greater than or equal to the optimal value of (3.1)–(3.3). Indeed, any feasible solution X of (3.1)–(3.3) is feasible for (5.1), and the dominance relation (3.2) implies that $\mathbb{E}[u(X)] \geq \mathbb{E}[u(Y)]$.

We define the dual problem as

$$\min_{u \in \mathcal{U}_1} D(u). \tag{5.2}$$

The set $\mathcal{U}_1$ is a closed convex cone in $\mathcal{C}([a,b])$ and $D(\cdot)$ is a convex functional, so (5.2) is a convex optimization problem. Employing general duality relations in convex programming (see [10, Thm. 2.1 and Chap. 3] and [4, Thm. 2.165]) we obtain the following result.

THEOREM 5.1. *Assume that the uniform dominance condition is satisfied and problem* (3.1)–(3.3) *has an optimal solution. Then problem* (5.2) *has an optimal solution and the optimal values of both problems coincide. Furthermore, the set of optimal solutions of* (5.2) *is the set of functions* $\hat{u} \in \mathcal{U}_1$ *satisfying* (4.2)–(4.3) *for an optimal solution* $\hat{X}$ *of* (3.1)–(3.3).

If Y has a discrete distribution with finitely many realizations $y_1 < y_2 < \cdots < y_m$, then the proof of Proposition 3.2 can also be used to show that the uniform dominance condition is equivalent to

$$F_2(\tilde{X}; y_i) < F_2(Y; y_i) \quad \text{for all} \quad y_i \in [a,b]. \tag{5.3}$$

Since $F_2(Y; y_1) = 0$, the uniform dominance condition cannot be satisfied unless $a > y_1$.

The constraint qualification condition simplifies when *all* distributions are finite. Then the functions $F_2(X; y_i)$, $i = 1, \ldots, m$, which appear at the left-hand side of (3.9), are convex polyhedral functions of the realizations x_i, $i = 1, \ldots, m$. Therefore, the dominance constraint is equivalent to a system of finitely many linear constraints. Consequently, in the discrete case, Theorems 4.2 and 5.1 are true under the following constraint qualification condition: there exists $\tilde{X} \in \operatorname{relint} C$ such that

$$F_2(\tilde{X}; y_i) \leq F_2(Y; y_i) \quad \text{for all} \quad y_i \in [a,b].$$

In this case we do not need to impose restrictions on a. In particular, we may have an interval $[a,b]$ covering all possible realizations y_i of the reference outcome. A sufficient condition for the existence of a nonempty relative interior of C is provided in [4, Thm. 2.197].

Moreover, the measure $\hat{\mu}$ is concentrated on the points y_i, $i = 1, \ldots, m$. The utility function $\hat{u}(\cdot)$ is concave, nondecreasing, piecewise linear, and has break points $y_1, \ldots, y_m$. Denoting by μ_i the Lagrange multipliers associated with (3.9), we obtain

the following representation of $\hat{u}(\cdot)$:

$$\hat{u}(t) = 0, \quad t \geq y_m,$$
$$\hat{u}(t) = \hat{u}(y_i) + (t - y_i)\sum_{j=i}^{m} \mu_j, \quad t \in [y_{i-1}, y_i], \quad i = m, m-1, \ldots, 2,$$
$$\hat{u}(t) = \hat{u}(y_1) + (t - y_1)\sum_{j=1}^{m} \mu_j, \quad t < y_1.$$

Equivalently,

$$\hat{u}(t) = -\sum_{i=1}^{N} \mu_i (y_i - t)_+,$$

which is a special case of (2.10).

6. Multiple dominance constraints. Let us now consider a problem with multiple dominance constraints introduced by several reference outcomes:

$$\max f(X) \tag{6.1}$$
$$\text{subject to } \mathbb{E}[(\eta - X)_+] \leq \mathbb{E}[(\eta - Y_j)_+] \quad \text{for all } \eta \in [a, b], \quad j = 1, \ldots, J, \tag{6.2}$$
$$X \in C. \tag{6.3}$$

Here $C \subset \mathcal{L}^1(\Omega, \mathcal{F}, P)$ and $Y_j \in \mathcal{L}^1(\Omega, \mathcal{F}, P)$, $j = 1, \ldots, J$. Denote by $\mathcal{U}_1^J$ the Cartesian product of J copies of $\mathcal{U}_1$. We define the Lagrangian $L : C \times \mathcal{U}_1^J \to \mathbb{R}$,

$$L(X, u) = f(X) + \mathbb{E}\sum_{j=1}^{J} [u_j(X) - u_j(Y_j)]. \tag{6.4}$$

We define the dual functional $D : \mathcal{U}_1^J \to \overline{\mathbb{R}}$,

$$D(u) = \sup_{X \in C} L(X, u),$$

and the dual problem,

$$\min_{u \in \mathcal{U}_1^J} D(u). \tag{6.5}$$

The uniform dominance condition takes on the following form: there exists $\tilde{X} \in \mathcal{L}^1(\Omega, \mathcal{F}, P)$ such that

$$\inf_{\eta \in [a,b]} \left\{ F_2(Y_j; \eta) - F_2(\tilde{X}; \eta) \right\} > 0, \quad j = 1, \ldots, J. \tag{6.6}$$

THEOREM 6.1. *Assume that condition* (6.6) *is satisfied. If* $\hat{X}$ *is an optimal solution of* (6.1)–(6.3)*, then there exists a function* $\hat{u} \in \mathcal{U}_1^J$ *such that*

$$L(\hat{X}, \hat{u}) = \max_{X \in C} L(X, \hat{u}) \tag{6.7}$$

and

$$\mathbb{E}[\hat{u}_j(\hat{X})] = \mathbb{E}[\hat{u}_j(Y_j)], \quad j = 1, \dots, J. \tag{6.8}$$

Conversely, if for some function $\hat{u} \in \mathcal{U}_1^J$ *an optimal solution* $\hat{X}$ *of* (6.7) *satisfies the dominance constraints* (6.2) *and* (6.8)*, then* $\hat{X}$ *is an optimal solution of* (6.1)–(6.3).

The proof follows the same line of argument as the proof of Theorem 4.2. Similarly, we have the duality theorem.

THEOREM 6.2. *Assume that condition* (6.6) *is satisfied and problem* (6.1)–(6.3) *has a solution. Then problem* (6.5) *has a solution and optimal values of both problems coincide. Moreover, the set of optimal solutions of* (6.5) *is the set of functions* $u \in \mathcal{U}_1^J$ *satisfying* (6.7)–(6.8) *for an optimal solution* $\hat{X}$ *of* (6.1)–(6.3).

7. Extension to higher order dominance. Our analysis extends to optimization problems involving higher order dominance constraints: problems of form (1.2)–(1.4) with (1.3) replaced by

$$X \succeq_{(k)} Y,$$

where $k > 2$. We assume that $C \subset \mathcal{L}^{k-1}(\Omega, \mathcal{F}, P)$ and that $Y \in \mathcal{L}^{k-1}(\Omega, \mathcal{F}, P)$, so that the kth-order dominance relation is well-defined.

Similarly to section 4 we focus on the problem in which the dominance constraint is enforced on an interval $[a, b]$:

$$\max f(X) \tag{7.1}$$

$$\text{subject to } F_k(X;\eta) \le F_k(Y;\eta) \quad \text{for all } \eta \in [a,b], \tag{7.2}$$

$$X \in C. \tag{7.3}$$

The set $\mathcal{U}_{k-1}$ of utility functions, which will play the role of Lagrange multipliers, contains all functions $u : \mathbb{R} \to \mathbb{R}$, for which there exists a nonnegative, nonincreasing, left-continuous, and bounded function $\varphi : [a, b] \to \mathbb{R}$ such that

$$\begin{array}{ll} u^{(k-1)}(t) = (-1)^k \varphi(t) & \text{for a.a. } t \in [a,b], \\ u^{(k-1)}(t) = (-1)^k \varphi(a) & \text{for } t < a, \\ u(t) = 0 & \text{for } t \ge b, \\ u^{(i)}(b) = 0, & i = 1, \dots, k-2. \end{array}$$

The symbol $u^{(i)}$ denotes the ith derivative of u and "a.a." means "almost all with respect to the Lebesgue measure." It is evident that

$$\lim_{t \to -\infty} \frac{|u(t)|}{|t|^{k-1}} < \infty.$$

The uniform dominance condition has the following form: there exists a point $\tilde{X} \in C$ such that

$$\inf_{\eta \in [a,b]} \Big\{ F_k(Y;\eta) - F_k(\tilde{X};\eta) \Big\} > 0. \tag{7.4}$$

Let us define the Lagrangian of (7.1)–(7.3), $L : C \times \mathcal{U}_{k-1} \to \mathbb{R}$, as follows:

$$L(X,u) = f(X) + \mathbb{E}[u(X)] - \mathbb{E}[u(Y)]. \tag{7.5}$$

Since $|u(t)|$ grows at the rate $|t|^{k-1}$, when $t \to -\infty$ and $X \in \mathcal{L}^{k-1}(\Omega, \mathcal{F}, P)$, the Lagrangian is well-defined.

THEOREM 7.1. *Assume that the uniform dominance condition* (7.4) *is satisfied. If* $\hat{X}$ *is an optimal solution of* (7.1)–(7.3)*, then there exists a function* $\hat{u} \in \mathcal{U}_{k-1}$ *such that*

$$L(\hat{X}, \hat{u}) = \max_{X \in C} L(X, \hat{u}) \tag{7.6}$$

and

$$\mathbb{E}[\hat{u}(\hat{X})] = \mathbb{E}[\hat{u}(Y)]. \tag{7.7}$$

Conversely, if for some function $\hat{u} \in \mathcal{U}_{k-1}$ *an optimal solution* $\hat{X}$ *of* (7.6) *satisfies* (7.2) *and* (7.7)*, then* $\hat{X}$ *is an optimal solution of* (7.1)–(7.3).

Proof. We can formulate (7.1)–(7.3) as an optimization problem in Banach spaces, as in the proof of Theorem 4.2. We introduce the Lagrangian $\Lambda : C \times \mathbf{rca}([a,b]) \to \mathbb{R}$,

$$\Lambda(X, \mu) = f(X) + \int_a^b [F_k(Y;\eta) - F_k(X;\eta)]\, d\mu(\eta). \tag{7.8}$$

Using necessary conditions of optimality, as in the proof of Theorem 4.2, we conclude that there exists a nonnegative measure $\hat{\mu} \in \mathbf{rca}([a,b])$ such that

$$\Lambda(\hat{X}, \hat{\mu}) = \max_{X \in C} \Lambda(X, \hat{\mu}) \tag{7.9}$$

and

$$\int_a^b [F_k(Y;\eta) - F_k(\hat{X};\eta)]\, d\hat{\mu}(\eta) = 0. \tag{7.10}$$

We extend the measure μ to the whole real line by assigning measure 0 to Borel sets not intersecting $[a,b]$. Using (2.2) and changing the order of integration in (7.8) we obtain

$$\int_a^b F_k(X;\eta)\, d\mu(\eta) = \int_{-\infty}^b \mu([t,b])\, F_{k-1}(X;t)\, dt. \tag{7.11}$$

Define the function $u \in \mathcal{U}_{k-1}$ as follows:

$$\begin{array}{rl} u(t) = 0, & t \geq b, \\ u^{(i)}(b) = 0, & i = 1, \dots, k-2, \\ u^{(k-1)}(t) = (-1)^k \mu([t,b]) & \text{for a.a. } t \leq b. \end{array}$$

Since $\mu \geq 0$, the function $u(\cdot)$ is nondecreasing and concave. We can rewrite (7.11) as

$$\int_a^b F_k(X;\eta)\, d\mu(\eta) = (-1)^k \int_{-\infty}^b F_k(X;t)\, du^{(k-1)}(t). \tag{7.12}$$

The $(k-1)$st derivative of u is monotone and $F_k(X;\cdot)$ is obtained by integrating $k-1$ times the monotone function $F(X;\cdot)$ (cf. (2.2)). Therefore, we can integrate (7.12) by parts $k-1$ times and get

$$(-1)^k \int_{-\infty}^b F_k(X;t)\, du^{(k-1)}(t) = -\int_{-\infty}^b u(t)\, dF(X;t) = -\mathbb{E}[u(X)].$$

All constants disappear, because the functions $F_i(X;\cdot)$ vanish at $-\infty$ and $u^{(k-i)}(b) = 0$, $i = 1, \ldots, k$. Substituting the last relation into (7.12) we obtain

$$\int_a^b F_k(X;\eta)\, d\mu(\eta) = -\mathbb{E}[u(X)]. \tag{7.13}$$

Thus, the measure $\hat{\mu}$ corresponds to a function $\hat{u} \in \mathcal{U}_{k-1}$, condition (7.9) implies (7.6), and condition (7.10) implies (7.7).

The reverse argument is similar. For every function $u \in \mathcal{U}_{k-1}$ we define the measure μ as

$$\mu([t,b]) = (-1)^k u_-^{(k-1)}(t),$$

where $u_-^{(k-1)}$ is the left derivative of $u^{(k-2)}$. Then we use (7.13) similarly to the proof of Theorem 4.2. □

Duality relations and optimality conditions for multiple constraints are analogous to the case of second-order dominance. The only difference is that the utility functions which play the role of multipliers have to be taken from the family $\mathcal{U}_{k-1}$.

8. Numerical illustration. To illustrate the features of the new models introduced in this paper, we consider the portfolio problem of Example 1. Table 8.1 contains historical data of returns of eight assets ($N = 8$) in 22 years. The assets are widely used indexes: U.S. three-month treasury bills, U.S. long-term government bonds, S&P 500, Willshire 5000, NASDAQ, Lehmann Brothers corporate bond index, EAFE foreign stock index, and gold. We use the returns in successive years as equally probable realizations.

We have chosen the reference random return Y as the return of the equally

TABLE 8.1
Asset returns (in %).

Year	Asset 1	Asset 2	Asset 3	Asset 4	Asset 5	Asset 6	Asset 7	Asset 8
1	7.5	-5.8	-14.8	-18.5	-30.2	2.3	-14.9	67.7
2	8.4	2	-26.5	-28.4	-33.8	0.2	-23.2	72.2
3	6.1	5.6	37.1	38.5	31.8	12.3	35.4	-24
4	5.2	17.5	23.6	26.6	28	15.6	2.5	-4
5	5.5	0.2	-7.4	-2.6	9.3	3	18.1	20
6	7.7	-1.8	6.4	9.3	14.6	1.2	32.6	29.5
7	10.9	-2.2	18.4	25.6	30.7	2.3	4.8	21.2
8	12.7	-5.3	32.3	33.7	36.7	3.1	22.6	29.6
9	15.6	0.3	-5.1	-3.7	-1	7.3	-2.3	-31.2
10	11.7	46.5	21.5	18.7	21.3	31.1	-1.9	8.4
11	9.2	-1.5	22.4	23.5	21.7	8	23.7	-12.8
12	10.3	15.9	6.1	3	-9.7	15	7.4	-17.5
13	8	36.6	31.6	32.6	33.3	21.3	56.2	0.6
14	6.3	30.9	18.6	16.1	8.6	15.6	69.4	21.6
15	6.1	-7.5	5.2	2.3	-4.1	2.3	24.6	24.4
16	7.1	8.6	16.5	17.9	16.5	7.6	28.3	-13.9
17	8.7	21.2	31.6	29.2	20.4	14.2	10.5	-2.3
18	8	5.4	-3.2	-6.2	-17	8.3	-23.4	-7.8
19	5.7	19.3	30.4	34.2	59.4	16.1	12.1	-4.2
20	3.6	7.9	7.6	9	17.4	7.6	-12.2	-7.4
21	3.1	21.7	10	11.3	16.2	11	32.6	14.6
22	4.5	-11.1	1.2	-0.1	-3.2	-3.5	7.8	-1

weighted portfolio $\bar{z} = [1/N\ 1/N\ \ldots\ 1/N]$. The return realizations are

$$y_k = \frac{1}{N}\sum_{n=1}^{N} r_{nk}, \quad k = 1, \ldots, m,$$

where $m = 22$, and r_{nk} denotes the return of asset n in year k. The probabilities of these realizations are $p_k = 1/m$, $k = 1, \ldots, m$. The expected return of the reference portfolio is equal to 10.6%.

Our objective is to maximize the expected return of the portfolio, under the condition that its return dominates the reference return Y in the second order.

Problem (3.1)–(3.3), with $f(X) = \mathbb{E}[X]$, owing to the transformation (3.10)–(3.12), can be formulated as the following linear programming problem:

$$\begin{aligned}
\max\ & \sum_{k=1}^{m}\sum_{n=1}^{N} p_k r_{nk} z_n & & \\
\text{subject to}\ & \sum_{n=1}^{N} r_{nk} z_n + s_{ik} \geq y_i, & & i = 1, \ldots, m,\ k = 1, \ldots, m, \\
& \sum_{k=1}^{m} p_k s_{ik} \leq v_i, & & i = 1, \ldots, m, \\
& s_{ik} \geq 0, & & i = 1, \ldots, m,\ k = 1, \ldots, m, \\
& \sum_{n=1}^{N} z_n = 1, & & \\
& z_n \geq 0, & & n = 1, \ldots, N.
\end{aligned}$$

The optimal portfolio has the form

$$\hat{z} = [0, 0, 0.068036, 0.188003, 0, 0.391376, 0.230924, 0.121661].$$

The expected return of this portfolio is 11.0%. It is slightly above the reference return and is much below the maximum expected return of a single asset, which is 14.1% (for Asset 7). This difference is due to the dominance constraint, which reflects risk aversion.

The shortfall functions $F_2(X;\cdot)$ and $F_2(Y;\cdot)$ appearing in the dominance constraints are illustrated in Figure 8.1. As we can see, the dominance constraint is active at several target values. The optimal measure $\hat{\mu}$ has the following atoms:

$$\begin{aligned}
\hat{\mu}(\{-4.49\%\}) &= 1.44366, \\
\hat{\mu}(\{-3.64\%\}) &= 0.59416, \\
\hat{\mu}(\{-0.84\%\}) &= 0.48158, \\
\hat{\mu}(\{23.39\%\}) &= 0.42786.
\end{aligned}$$

The corresponding utility function $\hat{u}$ appearing in the necessary conditions of optimality and duality relations is shown in Figure 8.2. It should be stressed that $\hat{u}$ is the Lagrange multiplier; that is, the optimal solution of the problem maximizes also the sum of the expected return and the expected value of $\hat{u}$. As we can see, negative returns are heavily penalized by this multiplier.

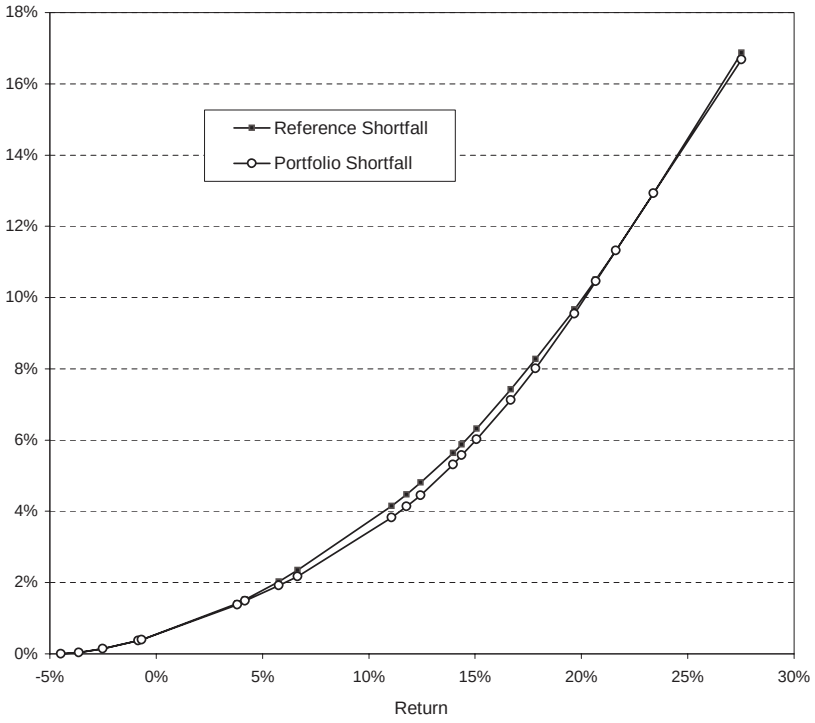


FIG. 8.1. *The shortfall functions for the portfolios of indices.*

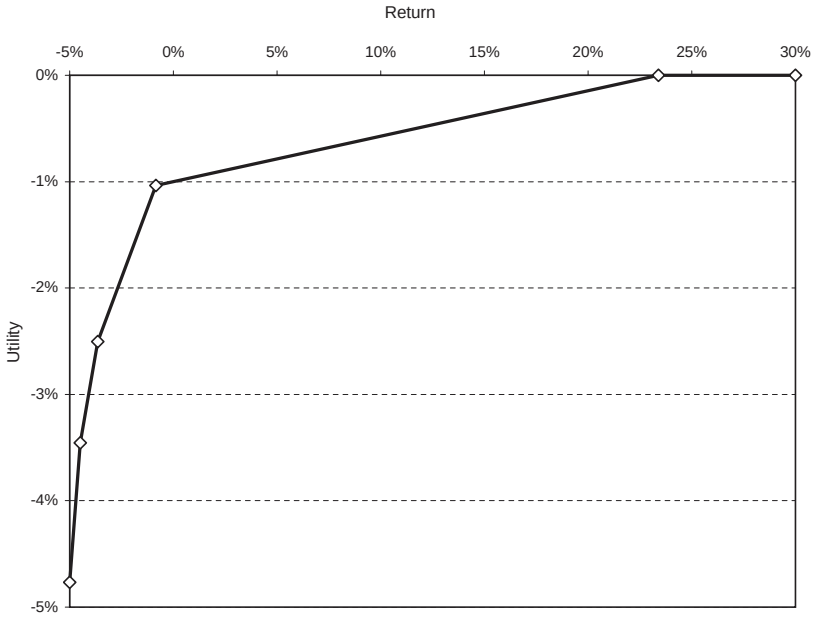


FIG. 8.2. *The optimal utility function for the portfolio of indices.*

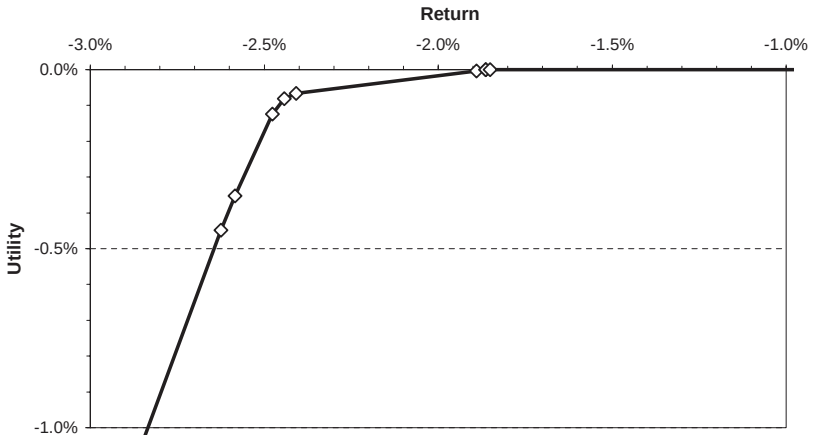


FIG. 8.3. *The optimal utility function for dominating the S&P* 500 *index.*

Our second example uses daily returns of the S&P 500 index as the reference outcome and creates a portfolio that dominates this reference outcome and has the highest expected return. As an illustration we used 248 daily returns from the year 2001 of the index and of 719 stocks from our database. The optimal portfolio is composed of only 7 stocks with weights 10.98%, 7.08%, 21.79%, 13.19%, 36.51%, 4.41%, 6.04%, respectively. It has the expected return of 0.64%, as compared to the expected return of −0.0359% of the S&P 500 index. The optimal utility function is shown in Figure 8.3.

The portfolio optimization problem is explored in much detail in our follow-up paper [7]. We also present there a specialized numerical method for solving such problems.

We want to stress that although our examples are drawn from the area of finance, our models and theoretical results are general.

REFERENCES

[1] I.M.L. BEALE, *On minimizing a convex function subject to linear inequalities,* J. Roy. Statist. Soc. Ser B, 17 (1955), pp. 173–184.
[2] P. BILLINGSLEY, *Probability and Measure*, John Wiley & Sons, New York, 1995.
[3] J.R BIRGE AND F.V. LOUVEAUX, *Introduction to Stochastic Programming*, Springer-Verlag, New York, 1997.
[4] J.F. BONNANS AND A. SHAPIRO, *Perturbation Analysis of Optimization Problems*, Springer-Verlag, New York, 2000.
[5] A. CHARNES, W.W. COOPER, AND G.H. SYMONDS, *Cost horizons and certainty equivalents: An approach to stochastic programming of heating oil,* Management Sci., 4 (1958), pp. 235–263.
[6] G.B. DANTZIG, *Linear programming under uncertainty*, Management Sci., 1 (1955), pp. 197–206.
[7] D. DENTCHEVA AND A. RUSZCZYŃSKI, *Portfolio optimization with stochastic dominance constraints*, J. Banking and Finance, submitted.
[8] N. DUNFORD AND J.T. SCHWARTZ, *Linear Operators*, Interscience, New York, 1958.
[9] P.C. FISHBURN, *Decision and Value Theory*, John Wiley & Sons, New York, 1964.
[10] E.G. GOL'SHTEIN, *Duality Theory in Mathematical Programming and Its Applications*, Nauka, Moscow, 1971 (in Russian).
[11] J. HADAR AND W. RUSSELL, *Rules for ordering uncertain prospects*, Amer. Econom. Rev., 59 (1969), pp. 25–34.
[12] G. HANOCH AND H. LEVY, *The efficiency analysis of choices involving risk*, Rev. Econom.

Stud., 36 (1969), pp. 335–346.
[13] W.K. Klein Haneveld and M.H. van der Vlerk, *Integrated Chance Constraints: Reduced Forms and an Algorithm*, Research Report 02A33, University of Groningen, The Netherlands; available online from http://www.ub.rug.nl/eldoc/som/a/02A33/.
[14] H. Levy, *Stochastic dominance and expected utility: Survey and analysis*, Management Sci., 38 (1992), pp. 555–593.
[15] D.T. Luc, *Theory of Vector Optimization*, Lecture Notes in Econom. and Math. Systems, Springer-Verlag, Berlin, 1989.
[16] H.M. Markowitz, *Portfolio selection*, J. Finance, 7 (1952), pp. 77–91.
[17] H.M. Markowitz, *Mean-Variance Analysis in Portfolio Choice and Capital Markets*, Blackwell, Oxford, 1987.
[18] L.B. Miller and H. Wagner, *Chance-constrained programming with joint constraints*, Oper. Res., 13 (1965), pp. 930–945.
[19] K. Mosler and M. Scarsini, eds., *Stochastic Orders and Decision under Risk*, Institute of Mathematical Statistics, Hayward, CA, 1991.
[20] A. Müller and D. Stoyan, *Comparison Methods for Stochastic Models and Risks*, John Wiley & Sons, Chichester, UK, 2002.
[21] W. Ogryczak and A. Ruszczyński, *From stochastic dominance to mean-risk models: Semideviations as risk measures*, European J. Oper. Res., 116 (1999), pp. 33–50.
[22] W. Ogryczak and A. Ruszczyński, *On consistency of stochastic dominance and mean-semideviation models*, Math. Program., 89 (2001), pp. 217–232.
[23] W. Ogryczak and A. Ruszczyński, *Dual stochastic dominance and related mean-risk models*, SIAM J. Optim., 13 (2002), pp. 60–78.
[24] A. Prékopa, *On probabilistic constrained programming*, in Proceedings of the Princeton Symposium on Mathematical Programming, Princeton University Press, Princeton, NJ, 1970, pp. 113–138.
[25] A. Prékopa, *Stochastic Programming*, Kluwer Academic, Dordrecht, Boston, 1995.
[26] J.P. Quirk and R. Saposnik, *Admissibility and measurable utility functions*, Rev. Econom. Stud., 29 (1962), pp. 140–146.
[27] S.T. Rachev and W. Römisch, *Quantitative stability in stochastic programming: The method of probability metrics*, Math. Oper. Res., 27 (2002), pp. 792–818.
[28] R.T. Rockafellar, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.
[29] R.T. Rockafellar and S. Uryasev, *Conditional value-at-risk for general loss distributions,* J. Banking and Finance, 26 (2002), pp. 1443–1471.
[30] R.T. Rockafellar and R.J.-B. Wets, *Stochastic convex programming: Basic duality*, Pacific J. Math., 62 (1976), pp. 173–195.
[31] R.T. Rockafellar and R.J.-B. Wets, *Stochastic convex programming: Singular multipliers and extended duality singular multipliers and duality,* Pacific J. Math., 62 (1976), pp. 507–522.
[32] R.T. Rockafellar and R.J.-B. Wets, *Stochastic convex programming: Relatively complete recourse and induced feasibility,* SIAM J. Control Optim., 14 (1976), pp. 574–589.
[33] M. Rothschild and J.E. Stiglitz, *Increasing risk:* I. *A definition*, J. Econom. Theory, 2 (1969), pp. 225–243.
[34] A. Ruszczyński and R. Vanderbei, *Frontiers of stochastically nondominated portfolios*, Econometrica, 71 (2003), pp. 1287–1297.
[35] S. Uryasev and R.T. Rockafellar, *Conditional value-at-risk: Optimization approach,* in Stochastic Optimization: Algorithms and Applications (Gainesville, FL, 2000), Appl. Optim. 54, Kluwer Academic, Dordrecht, The Netherlands, 2001, pp. 411–435.
[36] J. von Neumann and O. Morgenstern, *Theory of Games and Economic Behavior*, Princeton University Press, Princeton, NJ, 1947.
[37] R.J.-B. Wets, *Programming under uncertainty: The equivalent convex program,* SIAM J. Appl. Math., 14 (1966), pp. 89–105.
[38] R.J.-B. Wets, *Stochastic programs with fixed recourse: The equivalent deterministic program,* SIAM Rev., 16 (1974), pp. 309–339.

SIAM J. OPTIM.
Vol. 14, No. 2, pp. 567–583

ON THE LOCAL CONVERGENCE OF PATTERN SEARCH*

ELIZABETH D. DOLAN†, ROBERT MICHAEL LEWIS‡, AND VIRGINIA TORCZON§

Abstract. We examine the local convergence properties of pattern search methods, complementing the previously established global convergence properties for this class of algorithms. We show that the step-length control parameter which appears in the definition of pattern search algorithms provides a reliable asymptotic measure of first-order stationarity. This gives an analytical justification for a traditional stopping criterion for pattern search methods. Using this measure of first-order stationarity, we both revisit the global convergence properties of pattern search and analyze the behavior of pattern search in the neighborhood of an isolated local minimizer.

Key words. pattern search, local convergence analysis, global convergence analysis, stopping criteria, desultory rate of convergence

AMS subject classifications. 65K05, 90C56

DOI. S1052623400374495

1. Introduction.

Pattern search methods are a class of direct search methods for solving nonlinear optimization problems. In a series of papers [16, 11, 12, 13, 14] we established the global convergence properties of pattern search for both constrained and unconstrained problems. In this paper, we consider the local convergence properties of pattern search and revisit the global convergence properties in light of these new results.

For simplicity, our discussion will focus on the case of unconstrained minimization:

$$\min_{x \in \mathbb{R}^n} \quad f(x),$$

where $f : \mathbb{R}^n \to \mathbb{R}$. Results similar to those we present here can also be derived for the general case of bound and linear constraints [12, 13]. However, the underlying ideas are simpler to explain for the unconstrained case.

*Received by the editors June 26, 2000; accepted for publication (in revised form) March 9, 2003; published electronically November 6, 2003.
http://www.siam.org/journals/siopt/14-2/37449.html

†Industrial Engineering and Management Sciences, Northwestern University and Mathematics and Computer Science Division, Argonne National Laboratory, Argonne, IL 60439–4844 (dolan@mcs.anl.gov). This author's research was supported by the National Science Foundation under grant CCR–9734044 while the author was in residence at the College of William & Mary; by the Mathematical, Information, and Computational Sciences Division subprogram of the Office of Advanced Scientific Computing Research, Office of Science, U.S. Department of Energy, under contract W-31-109-Eng-38; and by the National Science Foundation (Challenges in Computational Science) grant CDA-9726385 and (Information Technology Research) grant CCR-0082807.

‡Department of Mathematics, College of William & Mary, P.O. Box 8795, Williamsburg, VA 23187–8795 (buckaroo@math.wm.edu). This author's research was supported by the National Aeronautics and Space Administration under NASA contract NAS1–97046 while the author was in residence at the Institute for Computer Applications in Science and Engineering (ICASE), NASA Langley Research Center, Hampton, VA and by the Computer Science Research Institute at Sandia National Laboratories.

§Department of Computer Science, College of William & Mary, P.O. Box 8795, Williamsburg, VA 23187–8795 (va@cs.wm.edu). This author's research was supported by the National Science Foundation under grant CCR–9734044, by the National Aeronautics and Space Administration under NASA contract NAS1–97046 while the author was in residence at the Institute for Computer Applications in Science and Engineering (ICASE), NASA Langley Research Center, Hampton, VA, and by the Computer Science Research Institute at Sandia National Laboratories.

We first show how the pattern size parameter, which plays a central role in the definition of pattern search methods and tacitly serves as a step-length control mechanism, also provides a reliable asymptotic measure of first-order stationarity. This gives an analytical justification for the traditional use of the pattern size parameter as a stopping criterion. We also establish a local convergence result concerning the behavior of the sequence of iterates produced by a pattern search algorithm in the neighborhood of an isolated local minimizer x_*. These analytical results are illustrated with some simple numerical experiments on quadratic objectives.

What is interesting about the analysis presented here is that we can establish local convergence properties despite the fact that direct search methods do not employ an explicit representation of the gradient of the objective and, as a consequence, cannot enforce a notion of sufficient decrease. We proved global convergence results for pattern search by showing that all iterates lie on a rational lattice. It is this restriction on the form of the steps that allows us to relax the notion of sufficient decrease and yet still prove global convergence. Pattern search may accept any point on the current integer lattice so long as it produces simple decrease on the value of the objective function at the current iterate. However, key to the global analysis is the notion of having searched in a sufficient number of directions from the current iterate to guarantee that we have not overlooked a potential direction of descent. It is only after searching over a sufficient set of directions that we are allowed to reduce the current step-length control parameter—which has the effect of refining the lattice over which we are searching.

This notion of sufficient local information at iterations at which we reduce the step-length control parameter allows us to show that the pattern size, as measured by the step-length control parameter, provides a reliable asymptotic measure of first-order stationarity. This analytical result is gratifying since it vindicates the long-standing use of the step-length control parameter as a stopping criterion for direct search methods (see, for instance, section 4 of [8]). The result on the correlation of the step-length control parameter and stationarity then enables us to study the local convergence properties of pattern search.

Notation. We use $\mathcal{L}(x_0)$ to denote the set $\{\, x \mid f(x) \leq f(x_0) \,\}$. We use ∂ to denote the boundary of a given set. It is assumed, unless otherwise noted, that all norms are Euclidean vector norms or the associated operator norm. Given x and $r > 0$, we denote by $\mathcal{B}(x, r)$ the open ball of radius r centered at x so that $\mathcal{B}(x, r) = \{\, y \mid \| y - x \| < r \,\}$. We also acknowledge an abuse of notation that is nonetheless convenient: if y is a vector and A is a matrix, we use the notation $y \in A$ to mean that the vector y is contained in the set of columns of the matrix A.

2. Pattern search. We first review the elements of pattern search that play a role in our local analysis. There are rigorous formal definitions of pattern search [16, 11], several features of which we will recall shortly. However, pattern search can perhaps be most quickly understood with the following simple example of a pattern search algorithm. At iteration k, we have an iterate $x_k \in \mathbb{R}^n$ and a step-length control parameter $\Delta_k > 0$. Let e_j, $j = 1, \ldots, n$, be the standard unit basis vectors. For the purposes of this example, we represent the set of directions that we will use for the search as the set $\mathcal{D} \equiv \{d_i\}_{i=1}^{2n} \equiv \{e_1, \ldots, e_n, -e_1, \ldots, -e_n\}$, though, as we discuss shortly, many other choices are possible. We now have several algorithmic options open to us. We consider the simple opportunistic strategy, which is to look successively at the points $x_+ = x_k + \Delta_k d_i$, $i \in \{1, \ldots, 2n\}$, until either we find an x_+ for which $f(x_+) < f(x_k)$ or we exhaust all $2n$ possibilities. Figure 2.1 illustrates the

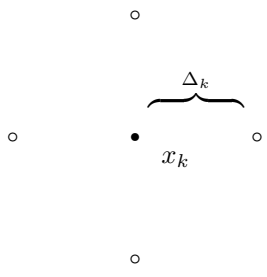


FIG. 2.1. *A simple instance of a pattern in* $\mathbb{R}^2$.

pattern of points among which we search for x_+ when $n = 2$.

If we find no x_+ such that $f(x_+) < f(x_k)$, then we call the iteration *unsuccessful*; otherwise, we consider the iteration *successful* since we have found a new iterate that produces decrease on f at x_k. When the iteration is unsuccessful, we set $x_{k+1} = x_k$ and are required to reduce Δ_k (typically, by a half) before continuing; otherwise, for a successful iteration, we set $x_{k+1} = x_+$ and leave the step-length control parameter alone; i.e., $\Delta_{k+1} = \Delta_k$ (though the analysis also allows us to increase Δ_k if the iteration is a success). We repeat this process until some suitable stopping criterion is satisfied.

Note that overall our requirements on the outcome of the search at each iteration are light: if after searching over all the points defined by $\Delta_k d_i$, $i = 1, \ldots, 2n$, we fail to find a point $x_+ = x_k + \Delta_k d_i$ that reduces the value of f at x_k, then we must try again with a smaller value of Δ_k. Otherwise, we accept as our new iterate the first point in the pattern that produces decrease. In the latter case, we may choose to increase Δ_k. In either case, we are free to make changes to the set of search directions $\mathcal{D}$ to be used in the next iteration, though we leave $\mathcal{D}$ unchanged in the example given previously. In general, changes to either the step-length control parameter or the set of search directions are subject to certain algebraic conditions, outlined fully in [11].

A distinguishing characteristic of pattern search methods is that they sample the function over a predefined pattern of points, all of which lie on a rational lattice. By enforcing structure on the form of the points in the pattern, as well as some simple rules on both the outcome of the search and the subsequent updates, standard global convergence results can be obtained [16, 11].

There remains the question of what constitutes an acceptable set of search directions. A pattern must form a positive spanning set for $\mathbb{R}^n$ [5]. A set of vectors $\{a_1, \ldots, a_p\}$ *positively spans* $\mathbb{R}^n$ if any vector $x \in \mathbb{R}^n$ can be written as a nonnegative linear combination of the vectors in the set; i.e.,

$$x = \alpha_1 a_1 + \cdots + \alpha_p a_p, \quad \alpha_i \geq 0 \quad \forall i.$$

The set $\{a_1, \ldots, a_p\}$ is called *positively dependent* if one of the a_i's is a nonnegative combination of the others; otherwise, the set is *positively independent*. A *positive basis* is a positively independent set whose positive span is $\mathbb{R}^n$.

It is straightforward to verify that the set of vectors $\{e_1, \ldots, e_n, -e_1, \ldots, -e_n\}$ we used to define the pattern for our simple example is a positive spanning set.

2.1. Prior results. Before proceeding to our local convergence results, we recall the following proposition from [11], which we state here without proof.

PROPOSITION 2.1. *Given any set* $\{a_1, \ldots, a_r\}$ *that positively spans* $\mathbb{R}^n$, $a_i \neq 0$ *for* $i = 1, \ldots, r$, *there exists* $c_{2.1} > 0$ *such that for all* $x \in \mathbb{R}^n$, *we can find an* a_i *for which*

$$x^T a_i \geq c_{2.1} \| x \| \| a_i \|.$$

Note that this is a purely geometric property of positive spanning sets.

2.2. Some formal definitions. We also need to recall some notation regarding both the pattern and the form of the search. For the details, we refer the reader to [16, 11].

We have noted already that the pattern must form a positive spanning set for $\mathbb{R}^n$. In fact, we represent the pattern using two components, a *basis matrix* and a *generating matrix*.

The basis matrix can be any nonsingular matrix $B \in \mathbb{R}^{n \times n}$.

The generating matrix is an integral matrix $C_k \in \mathbb{Z}^{n \times p_k}$, where $p_k > n + 1$. We require C_k to contain a minimum of $n + 2$ columns because the minimum number of vectors in a positive spanning set is $n + 1$ [5]; for convenience, we require a column of zeros to denote the zero step. We further partition the generating matrix to reveal the positive basis that guarantees that the pattern positively spans $\mathbb{R}^n$. We call the columns associated with the positive basis the *core pattern*, which we denote Γ_k; any remaining columns in the positive spanning set are denoted L_k:

$$C_k = [\ \Gamma_k \quad L_k \quad 0\ ]. \tag{2.1}$$

We further require that $\Gamma_k \in \mathbf{\Gamma}$, where $\mathbf{\Gamma}$ comprises a finite set of integral matrices, the columns of which form a positive basis for $\mathbb{R}^n$.

A *pattern* is then represented by the columns of the matrix $P_k = BC_k$. For convenience, we use the partition of the generating matrix C_k given in (2.1) to partition P_k as follows:

$$P_k = BC_k = [\ B\Gamma_k \quad BL_k \quad 0\ ].$$

To tie this notation back to the example that introduces section 2, we note that $B = I$, $\Gamma_k = [\, I \ -I\,]$, and $L_k \equiv \emptyset$. Since the choices of Γ_k and L_k are fixed in our example, $P_k \equiv [\, I \ -I \ \ 0\,]$ for all k.

Now, given the step-length control parameter $\Delta_k \in \mathbb{R}$, $\Delta_k > 0$, we define a *trial step* s_k^i to be any vector of the form $s_k^i = \Delta_k B c_k^i$, where c_k^i is a column of C_k.

In Figure 2.2 we state the general form of a pattern search method for unconstrained minimization.

We have remarkable latitude in our choice of the step s_k. For the global convergence analysis to hold, we only need to satisfy the hypotheses on the outcome of the unconstrained exploratory moves given in Figure 2.3.

A few comments on these hypotheses are in order. The first hypothesis is straightforward: the step returned must be a column in the current pattern matrix P_k, scaled by the current value of the step-length control parameter Δ_k. This condition ensures that the steps we consider remain on the rational lattice; arbitrary steps are not allowed.

For our purposes, the second hypothesis is the more interesting one. Notice that in Figure 2.2, a successful iteration of pattern search requires only that the step s_k produces simple decrease; i.e., $f(x_k + s_k) < f(x_k)$. Thus, *any* nonzero step defined by a column of $\Delta_k P_k$ that satisfies the condition $f(x_k + s_k) < f(x_k)$ may be returned

Let $x_0 \in \mathbb{R}^n$ and $\Delta_0 > 0$ be given.
For $k = 0, 1, \ldots,$ until convergence do:
1. Compute $f(x_k)$.
2. Determine a step s_k using an unconstrained exploratory moves algorithm.
3. If $f(x_k + s_k) < f(x_k)$, then $x_{k+1} = x_k + s_k$. Otherwise, $x_{k+1} = x_k$.
4. Update C_k and Δ_k.

FIG. 2.2. *Generalized pattern search for unconstrained minimization.*

1. $s_k \in \Delta_k P_k$.
2. If $\min \{ f(x_k + y) \mid y \in \Delta_k B\Gamma_k \} < f(x_k)$, then $f(x_k + s_k) < f(x_k)$.

FIG. 2.3. *Hypotheses on the outcome of the unconstrained exploratory moves.*

1. If all $f(x_k + s_k) \geq f(x_k)$, then $\Delta_{k+1} = \theta\Delta_k$, where $\theta \in (0, 1)$.
2. If any $f(x_k + s_k) < f(x_k)$, then $\Delta_{k+1} = \lambda_k\Delta_k$, where $\lambda_k \geq 1$.

FIG. 2.4. *Basic rules for updating Δ_k.*

by the exploratory moves since it immediately satisfies both of the hypotheses given in Figure 2.3—even if we do not explicitly find $\min \{ f(x_k + y) \mid y \in \Delta_k B\Gamma_k \}$.

The second hypothesis in Figure 2.3 ensures that we have sufficient information about the local behavior of f to declare an iteration *unsuccessful*, accept the zero step $s_k = 0$ (so that $x_{k+1} \equiv x_k$), and reduce Δ_k to continue the search with smaller steps at the next iteration. The second hypothesis implicitly decrees that we may return the zero step, and thus reduce Δ_k, only when we have looked at all the steps defined by the core pattern; i.e., all steps of the form $y \in \Delta_k B\Gamma_k$.

The core pattern $B\Gamma_k$ must be a positive basis. This means that even though we do not have an explicit representation of $\nabla f(x_k)$ (assuming that f is differentiable), the geometric property of positive spanning sets captured in Proposition 2.1 gives us a positive lower bound, which is independent of k, on the angle between $-\nabla f(x_k)$ (assuming it is nonzero) and some a_i in the positive spanning set. At any given iteration, we do not know for which a_i this lower bound holds. However, this guaranteed lower bound, when combined with the second hypothesis in Figure 2.3, ensures that at the end of an unsuccessful iteration, we have significant information about the local behavior of f at x_k. Furthermore, the quality of our local information improves as we reduce Δ_k.

Finally, we make a brief comment on the basic rules for updating Δ_k, which are given in Figure 2.4. We also must impose additional conditions on the choice of θ and λ_k to ensure that Theorem 3.2 from [16] holds. Rather than detail these conditions here, since they are outlined fully in [16] (other options are discussed in [10]), we note the two essential consequences. First, if our choices for θ and λ_k ensure that Theorem 3.2 from [16] holds, then all the iterates lie on a translated integer lattice. Second, the rules for updating Δ_k ensure that Δ_k is *reduced* after any unsuccessful iteration since $\theta \in (0, 1)$. The latter means that after any unsuccessful iteration, pattern search refines the lattice of points over which the search resumes.

We can capitalize on the structure of pattern search refinement to construct local convergence results. The subsequence of *unsuccessful* iterates, which is what interests us here, is well-defined: they are the iterates at which we must reduce Δ_k to ensure

that the search can make further progress. We reduce Δ_k only after we have sufficient local information about the behavior of f to justify this action: we have considered all the steps defined by the columns of $\Delta_k\Gamma_k$, and none of them has produced descent on f at x_k. We presently use this fact to assess stationarity.

3. Measuring first-order stationarity. The following theorem shows that the step-length control parameter Δ_k, when small enough, provides a reasonable measure of first-order stationarity at an unsuccessful iterate. For simplicity, we assume that $\nabla f(x)$ is Lipschitz continuous. For the reader interested in greater generality, we note that a similar result can be proven under the assumption of uniform continuity.

THEOREM 3.1. *Suppose that for some $\rho > 0$, $\nabla f(x)$ is Lipschitz continuous, with Lipschitz constant $\mathcal{K}$, on the open neighborhood $\Omega = \cup_{x\in\mathcal{L}(x_0)} \mathcal{B}(x,\rho)$ of $\mathcal{L}(x_0)$. Then there exist $\delta_{3.1} > 0$ and $c_{3.1} > 0$ for which the following holds. If x_k is an unsuccessful iterate and $\Delta_k < \delta_{3.1}$, then*

$$\| \nabla f(x_k) \| \le c_{3.1}\Delta_k.$$

Proof. Let $r = \frac{1}{2}\min\{1,\rho\}$. If $x \in \mathcal{L}(x_0)$, then the ball $\mathcal{B}(x,r)$ is contained in Ω. We are interested in steps of the form $s = \Delta_k B c_k^i$, where c_k^i is a column of the core matrix Γ_k. Since $\Gamma_k \in \mathbf{\Gamma}$ and $\mathbf{\Gamma}$ is finite, $\| s \| \le \Delta_k \| B \| \Gamma^*$, where Γ^* is the maximum norm of any column of the matrices in the set $\mathbf{\Gamma}$. Set $\delta_{3.1} = r/(\| B \| \Gamma^*)$.

By the definition of pattern search, for any $\Gamma_k \in \mathbf{\Gamma}$ the set $\{ s \mid s \in \Delta_k B\Gamma_k \}$ forms a positive basis for $\mathbb{R}^n$. Thus Proposition 2.1 assures us of the existence of a step s for which

$$-\nabla f(x_k)^T s \ge c_{2.1} \| \nabla f(x_k) \| \| s \|. \tag{3.1}$$

Since iteration k is unsuccessful, it follows that

$$f(x_k + s) - f(x_k) \ge 0 \quad \forall s \in \Delta_k B\Gamma_k.$$

Since we assume $\Delta_k < \delta_{3.1}$, $(x_k + s) \in \mathcal{B}(x_k, r) \subset \Omega$, and we can apply the mean value theorem. In addition, using (3.1) and the Cauchy–Schwarz inequality, for some ξ in the line segment connecting x_k and $x_k + s$ we have

$$\begin{aligned} 0 &\le f(x_k + s) - f(x_k) \\ &= \nabla f(x_k)^T s + (\nabla f(\xi) - \nabla f(x_k))^T s \\ &\le -c_{2.1} \| \nabla f(x_k) \| \| s \| + \| \nabla f(\xi) - \nabla f(x_k) \| \| s \|, \end{aligned}$$

where s is the step for which (3.1) holds. Thus

$$c_{2.1} \| \nabla f(x_k) \| \le \| \nabla f(\xi) - \nabla f(x_k) \|.$$

Again, since $\mathcal{B}(x_k, r) \subset \Omega$, the Lipschitz continuity of $\nabla f(x)$ gives us

$$c_{2.1} \| \nabla f(x_k) \| \le \mathcal{K} \| \xi - x_k \| \le \mathcal{K} \| s \| \le \mathcal{K}\, \Delta_k \| B \| \Gamma^*.$$

Therefore,

$$\| \nabla f(x_k) \| \le c_{3.1}\Delta_k,$$

with $c_{3.1} = \mathcal{K} \| B \| \Gamma^*/c_{2.1}$. ∎

Theorem 3.1 gives a theoretical justification for a traditional stopping criterion for pattern search methods. In the long literature on direct search methods, one frequently encounters the suggestion that a direct search method be terminated when some measure of the step size first falls below a value deemed suitably small [8, 4, 2]. In the case of pattern search, Theorem 3.1 vindicates this intuition. At unsuccessful iterations, the step size in pattern search (as measured by Δ_k) provides a bound on first-order stationarity. At the same time, it is after the unsuccessful iterations that Δ_k is decreased. Thus, decrease in Δ_k provides a simple measure of progress which can be used reliably to test for convergence. We discuss further the use of Δ_k to measure progress when we present some numerical examples in section 5.

A similar relation between Δ_k and constrained stationarity in the case of pattern search for bound constrained problems is explicitly used in the pattern search augmented Lagrangian algorithm in [14]. The result plays a critical role in allowing successive inexact minimization of an augmented Lagrangian without an explicit estimate of the gradient. A relation similar to Theorem 3.1 for linearly constrained pattern search appears in [13].

The global convergence analysis of pattern search in [16] says that if $\mathcal{L}(x_0)$ is compact, then $\liminf_{k\to\infty} \Delta_k = 0$ and $\liminf_{k\to\infty} \| \nabla f(x_k) \| = 0$. The former result and Theorem 3.1 allow us to sharpen the latter result. Let the set $\mathcal{U}$ represent a subsequence of unsuccessful iterates for which $\lim_{k\to\infty, k\in\mathcal{U}} \Delta_k = 0$ (such a subsequence exists since $\liminf_{k\to\infty} \Delta_k = 0$). Then Theorem 3.1 says that we have $\lim_{k\to\infty, k\in\mathcal{U}} \| \nabla f(x_k) \| = 0$.

The general result $\liminf_{k\to\infty} \| \nabla f(x_k) \| = 0$ for pattern search leaves open the possibility that $\| \nabla f(x_k) \|$ does not converge. In [1], Audet shows that this actually can occur by constructing a pattern search algorithm and an objective function for which $\{x_k\}$ has infinitely many limit points, one of which is not a stationary point of the objective. However, in his example, the subsequence of iterates converging to the nonstationary point of the objective function are all *successful* iterates. Theorem 3.1 reassures us that in practice we need not worry about convergence to nonstationary points. If we stop the algorithm at the first *unsuccessful* iterate for which $\Delta_k < \Delta_*$ for some suitably small stopping tolerance Δ_*, then Theorem 3.1 says that we may reasonably expect $\| \nabla f(x_k) \|$ to be small.

4. Local convergence. We now consider the local convergence of pattern search methods. We begin with a collection of hypotheses and definitions we will need.

The first condition is a mild hypothesis on the generating matrices C_k that allows us to bound the size of the steps $\{s_k\}$.

HYPOTHESIS 0. *The columns of the generating matrices* $C_k = \left[c_k^1 \cdots c_k^{p_k}\right]$ *remain bounded in norm; i.e., there exists* $C_0 > 0$ *such that for all* k, $C_0 > \left\| c_k^i \right\|$, *for all* $i = 1, \ldots, p_k$. *Thus, there exists a constant* $c_0 > 0$ *such that any step* s_k *satisfies*

$$\| s_k \| \leq c_0 \Delta_k.$$

We also impose the following condition on the step-length control parameter Δ_k.

HYPOTHESIS 1. *There exists* N *for which* Δ_k *is monotonically nonincreasing for all* $k \geq N$.

Note that this is a condition we can explicitly enforce by not allowing increases in Δ_k after some iteration N; Δ_k can stay the same or decrease.

The local convergence results are concerned with the behavior of pattern search in a neighborhood of an isolated local minimizer x_*. We make the following assumptions about the behavior of f in a neighborhood of x_*.

HYPOTHESIS 2. *We assume the existence of an open ball $\mathcal{B}(x_*, r)$, $r > 0$, for which $f(x)$ is twice continuously differentiable on $\mathcal{B}(x_*, r)$, $\nabla^2 f(x)$ is positive definite on $\mathcal{B}(x_*, r)$, $\nabla f(x_*) = 0$, and lower and upper bounds $\sigma_{\min}$ and $\sigma_{\max}$ on the singular values of $\nabla^2 f(x)$ on $\mathcal{B}(x_*, r)$ exist. We further assume $\sigma_{\min} > 0$.*

We then define

$$\kappa = \sigma_{\max}/\sigma_{\min}. \tag{4.1}$$

We also define

$$\eta = r/(\| B \| \Gamma^* + 1). \tag{4.2}$$

This choice ensures that if $\| x_k - x_* \| < \eta$ and $\Delta_k < \eta$, then for any step $s \in \Delta_k B\Gamma_k$ we have $\| (x_k + s) - x_* \| < r$.

Our first result relates Δ_k to $\| x_k - x_* \|$ at unsuccessful iterates.

PROPOSITION 4.1. *Under Hypothesis 2, there exists $c_{4.1} > 0$ for which the following holds. If x_k is an unsuccessful iterate, $\Delta_k < \eta$, and $\| x_k - x_* \| < \eta$ (where η is as in (4.2)), then*

$$\| x_k - x_* \| \leq c_{4.1} \Delta_k.$$

Proof. Proposition 2.1 assures us of the existence of a step $s \in \Delta_k B\Gamma_k$ for which

$$-\nabla f(x_k)^T s \geq c_{2.1} \| \nabla f(x_k) \| \, \| s \|. \tag{4.3}$$

If iteration k is unsuccessful, it follows that

$$f(x_k + s) - f(x_k) \geq 0 \quad \forall s \in \Delta_k B\Gamma_k.$$

Because $\Delta_k < \eta$, we know that $(x_k + s) \in \mathcal{B}(x_*, r)$, where f is differentiable, and we can apply the mean value theorem. In addition, using (4.3) and the Cauchy–Schwarz inequality, for some ξ in the line segment connecting x_k and $x_k + s$ we have

$$0 \leq f(x_k + s) - f(x_k) \leq -c_{2.1} \| \nabla f(x_k) \| \, \| s \| + \| \nabla f(\xi) - \nabla f(x_k) \| \, \| s \|,$$

where s is the step for which (4.3) holds. Thus

$$c_{2.1} \| \nabla f(x_k) \| \leq \| \nabla f(\xi) - \nabla f(x_k) \|. \tag{4.4}$$

By the integral form of the mean value theorem,

$$\begin{aligned} \| \nabla f(\xi) - \nabla f(x_k) \| &= \left\| \int_0^1 \left[\nabla^2 f(x_k + t(\xi - x_k))(\xi - x_k) \right] dt \right\| \\ &\leq \sigma_{\max} \| \xi - x_k \| \leq \sigma_{\max} \Delta_k \| B \| \Gamma^*. \end{aligned}$$

Meanwhile, since $\nabla f(x_*) = 0$, we have

$$\begin{aligned} \| \nabla f(x_k) \| &= \| \nabla f(x_k) - \nabla f(x_*) \| \\ &= \left\| \int_0^1 \left[\nabla^2 f(x_* + t(x_k - x_*))(x_k - x_*) \right] dt \right\| \geq \sigma_{\min} \| x_k - x_* \|. \end{aligned} \tag{4.5}$$

Combining (4.4) and (4.5) yields

$$c_{2.1} \, \sigma_{\min} \| x_k - x_* \| \leq c_{2.1} \| \nabla f(x_k) \| \leq \sigma_{\max} \| B \| \Gamma^* \, \Delta_k.$$

Setting $c_{4.1} = (\sigma_{\max} \| B \| \Gamma^*)/(c_{2.1}\, \sigma_{\min})$ completes the proof. ∎

Next we have the following elementary result concerning the level sets of f near an isolated local minimizer x_*.

PROPOSITION 4.2. *Under Hypothesis* 2, *if* $x, y \in \mathcal{B}(x_*, \eta)$ *and* $f(x) \le f(y)$, *then*

$$\| x - x_* \| \le \kappa^{\frac{1}{2}} \| y - x_* \|, \tag{4.6}$$

where κ *is as defined in* (4.1).

Proof. Suppose $x, y \in \mathcal{B}(x_*, \eta)$ and $f(x) \le f(y)$. From Taylor's theorem with remainder and the fact that $\nabla f(x_*) = 0$, we have

$$f(y) = f(x_*) + \frac{1}{2}(y - x_*)^T \nabla^2 f(\xi)(y - x_*),$$
$$f(x) = f(x_*) + \frac{1}{2}(x - x_*)^T \nabla^2 f(\omega)(x - x_*)$$

for ξ and ω on the line segments connecting x_* with y and x, respectively. Since $f(x) \le f(y)$, we obtain

$$0 \le f(y) - f(x) = \frac{1}{2}(y - x_*)^T \nabla^2 f(\xi)(y - x_*) - \frac{1}{2}(x - x_*)^T \nabla^2 f(\omega)(x - x_*),$$

whence

$$0 \le \sigma_{\max} \| y - x_* \|^2 - \sigma_{\min} \| x - x_* \|^2,$$

and thus (4.6). ∎

We use the previous proposition to show that if we start sufficiently close to x_* with a sufficiently small step-length control parameter Δ_k and we have stopped allowing increases in Δ_k (Hypothesis 1), then pattern search will not move away from a neighborhood of x_*.

PROPOSITION 4.3. *Under Hypotheses* 0, 1, *and* 2, *there exist* $\delta_{4.3} > 0$ *and* $\varepsilon_{4.3} > 0$ *for which the following holds. For* $k \ge N$, *where* N *is as defined in Hypothesis* 1, *if* x_k *is an iterate for which* $\Delta_k < \delta_{4.3}$ *and* $\| x_k - x_* \| < \varepsilon_{4.3}$, *then for all* $\ell \ge k$,

$$\| x_\ell - x_* \| < \eta,$$

where η *is as in* (4.2).

Proof. Choose $\delta_{4.3}$ and $\varepsilon_{4.3}$ to satisfy

$$\delta_{4.3} < \frac{\eta}{2c_0},$$
$$\varepsilon_{4.3} < \frac{1}{2}\kappa^{-\frac{1}{2}}\eta,$$

where the constant c_0 comes from Hypothesis 0 and the definition of κ appears as (4.1). Observe that the definition of κ means that for any choice of $\eta > 0$,

$$\frac{1}{2}\kappa^{-\frac{1}{2}}\eta \le \frac{\eta}{2}.$$

The proof is by induction. First consider $x_{k+1} = x_k + s_k$. Hypothesis 0 gives us $\| x_{k+1} - x_k \| = \| s_k \| \le c_0 \Delta_k$. We have, a priori,

$$\| x_{k+1} - x_* \| \le \| x_{k+1} - x_k \| + \| x_k - x_* \| < c_0 \Delta_k + \varepsilon_{4.3} < \eta.$$

Now consider any $\ell \geq k+1$, and suppose

$$\| x_\ell - x_* \| < \eta.$$

Then

$$\| x_{\ell+1} - x_* \| \leq \| x_{\ell+1} - x_\ell \| + \| x_\ell - x_* \|. \tag{4.7}$$

Hypothesis 1 assures us that $\Delta_\ell \leq \Delta_k$ for $\ell \geq k$, so

$$\| x_{\ell+1} - x_\ell \| \leq c_0 \Delta_\ell \leq c_0 \Delta_k.$$

Meanwhile, by the induction hypothesis, $x_\ell \in \mathcal{B}(x_*, \eta)$. Since $f(x_\ell) \leq f(x_k)$ as well, Proposition 4.2 and the assumption $\| x_k - x_* \| < \varepsilon_{4.3}$ say that

$$\| x_\ell - x_* \| \leq \kappa^{\frac{1}{2}} \| x_k - x_* \| < \kappa^{\frac{1}{2}} \varepsilon_{4.3}.$$

Thus (4.7) yields

$$\| x_{\ell+1} - x_* \| < c_0 \Delta_k + \kappa^{\frac{1}{2}} \varepsilon_{4.3} < \eta. \qquad \square$$

An immediate consequence of Proposition 4.3 is the following, which is simply a localized version of Theorem 3.3 from [16].

PROPOSITION 4.4. *Suppose Hypotheses* 0–2 *hold. Let* $\delta_{4.3} > 0$ *and* $\varepsilon_{4.3} > 0$ *be as in Proposition* 4.3. *If for some* $k \geq N$, *where* N *is as defined in Hypothesis* 1, *we have* $\Delta_k < \delta_{4.3}$ *and* $\| x_k - x_* \| < \varepsilon_{4.3}$, *then* $\lim_{j \to \infty} \Delta_j = 0$.

Proof. The proof proceeds by contradiction. Suppose $\lim_{j \to \infty} \Delta_j \neq 0$. Then Δ_j has some minimum value $\Delta_{\min} > 0$, which implies that after some iteration k we have an infinite number of successful iterations. From Proposition 4.3 and Hypothesis 0, we see that all possible iterates after k remain in a bounded set. As discussed in section 2.2, the structure of pattern search algorithms is such that all possible iterates must lie on a translated integer lattice that depends on $\Delta_{\min}$. The intersection of a bounded set with a translated integer lattice is finite. So if we do not reduce Δ_j beyond $\Delta_{\min}$, there is only a finite number of points that we can consider that remain in the bounded set. Thus, if there is an infinite number of successful iterations, there must exist at least one point $\hat{x}$ in the lattice for which $x_j = \hat{x}$ for more than one value of j. This leads to a contradiction because we can have a successful iteration and avoid decreasing Δ_j only if $f(x_j) < f(x_{j-1})$. Therefore, we must have $\lim_{j \to \infty} \Delta_j = 0$. $\square$

This argument is analogous to the basic reasoning found in the proof of Theorem 3.3 in [16], in which it is shown that $\liminf_{k \to +\infty} \Delta_k = 0$ under the assumption that $\mathcal{L}(x_0)$ is compact.

Putting the pieces together, we obtain the following local convergence result. It says that if at some iteration the entire set of trial points is sufficiently close to a local minimizer x_* satisfying Hypothesis 2, then the sequence of subsequent iterates will converge to x_*. We use the suggestive notation $x_k + \Delta_k P_k$ to represent the set of all possible trial points at iteration k, $\left\{ x_k + \Delta_k B c_k^i \mid i = 1, \ldots, p_k \right\}$, where B is the basis matrix and c_k^i is a column of the generating matrix C_k.

THEOREM 4.5. *Given a pattern search algorithm satisfying Hypotheses* 0–1, *let* N *be as in Hypothesis* 1. *Suppose Hypothesis* 2 *holds and that, in particular,* x_* *is a point satisfying Hypothesis* 2.

Then there exist $\rho > 0$ and $c_{4.5} > 0$ for which the following hold. Suppose that at some iteration K, $K \geq N$, we have $x_K + \Delta_K P_K \subset \mathcal{B}(x_*, \rho)$. Let $\bar{K}$ be the first unsuccessful iteration after K. Then for all $k > \bar{K}$,

$$\| x_k - x_* \| \leq c_{4.5} \Delta_{m(k)}, \tag{4.8}$$

where $m(k)$ is the last unsuccessful iteration preceding or including k. As a consequence, we have $\lim_{k\to\infty} x_k = x_*$.

Proof. We begin by noting that the integrality of the generating matrix C_k guarantees that for all k,

$$\min_{i \in \{1,\ldots,(p_k-1)\}} \left\| c_k^i \right\| \geq 1. \tag{4.9}$$

The bound in (4.9) excludes the last column of C_k, which allows the zero step. We also know that for all $k \geq 0$ any trial step $s_k^i \in \Delta_k P_k$ satisfies

$$\left\| s_k^i \right\| \;=\; \Delta_k \left\| B c_k^i \right\| \;\geq\; \Delta_k \sigma_n(B) \left\| c_k^i \right\|, \tag{4.10}$$

where $\sigma_n(B)$ denotes the smallest singular value of the basis matrix B.

Our assumption that $x_K + \Delta_K P_K \subset \mathcal{B}(x_*, \rho)$ means that for any $s_K^i \in \Delta_K P_K$ we have

$$\left\| s_K^i \right\| < 2\rho. \tag{4.11}$$

Combining (4.9), (4.10), and (4.11), we obtain

$$\Delta_K < \frac{2\rho}{\sigma_n(B) \left\| c_K^i \right\|} \leq \frac{2\rho}{\sigma_n(B)}$$

for all $i \in \{1, \ldots, (p_K - 1)\}$. The assumption that $x_K + \Delta_K P_K \subset \mathcal{B}(x_*, \rho)$ also yields

$$\| x_K - x_* \| < \rho.$$

Thus we can choose $\rho > 0$ to be so small that if $x_K + \Delta_K P_K \subset \mathcal{B}(x_*, \rho)$, then

$$\Delta_K < \min\{\eta, \delta_{4.3}\} \quad \text{and} \quad \| x_K - x_* \| < \min\{\eta, \varepsilon_{4.3}\},$$

where η is as in (4.2) and $\delta_{4.3}$, $\varepsilon_{4.3}$ are as in Proposition 4.3. Proposition 4.3 then gives us

$$\| x_k - x_* \| < \eta \quad \forall k \geq K. \tag{4.12}$$

By assumption, $\bar{K}$ is the first unsuccessful iteration after K. We now consider two cases.

First, for all unsuccessful iterates x_k with $k \geq \bar{K}$, Proposition 4.1 gives us

$$\| x_k - x_* \| \leq c_{4.1} \Delta_k. \tag{4.13}$$

Since x_k is an unsuccessful iterate and Δ_k has not yet been reduced, $k = m(k)$; and we can restate (4.13) as

$$\| x_{m(k)} - x_* \| \leq c_{4.1} \Delta_{m(k)} \quad \forall k \geq \bar{K}. \tag{4.14}$$

Second, for all successful iterations $k > \bar{K}$, we have $f(x_k) < f(x_{m(k)})$. Since $k > \bar{K} \geq K$, (4.12) assures us that $x_k, x_{m(k)} \in \mathcal{B}(x_*, \eta)$. It then follows from Proposition 4.2 that

$$\| x_k - x_* \| \leq \kappa^{\frac{1}{2}} \| x_{m(k)} - x_* \|, \tag{4.15}$$

where κ is as in (4.1).

Together (4.14) and (4.15) imply that for all $k > \bar{K}$, (4.8) holds with $c_{4.5} = \kappa^{\frac{1}{2}} c_{4.1}$ since the definition of κ in (4.1) ensures that $\kappa^{\frac{1}{2}} \geq 1$.

Finally, since Proposition 4.4 says $\Delta_k \to 0$, it follows that $\lim_{k\to\infty} x_k = x_*$. ∎

This theorem complements Theorem 3.7 of [16], where it is shown, under different hypotheses and a more stringent criterion for accepting a step, that $\| \nabla f(x_k) \| \to 0$. The trade-off is that while here we relax the criterion for accepting a step, Hypothesis 2 places stronger assumptions on f than those used in [16], where all that is assumed about f is that it is continuously differentiable on a neighborhood of $\mathcal{L}(x_0)$.

Theorem 4.5 is similar to local convergence results for other minimization algorithms. The standard convergence results for Newton's method and quasi-Newton methods (with exact gradients) say that if we start sufficiently close to a point x_* satisfying Hypothesis 2, then the sequence of subsequent iterates will converge to x_* [15]. Our result is even more like the local convergence results for minimization with finite-difference estimates of the gradient, with which pattern search can be aptly compared. Theorem 5.1 in [3], an example of such a result, requires the points from whose objective values the finite-difference estimate of the gradients is constructed to lie sufficiently close to x_*. Our requirement that the entire pattern be close to x_* is similar.

Theorem 4.5 says that for the subsequence of *unsuccessful* iterates, the rate of convergence is R-linear. Theorem 4.5 says nothing about what may happen at the successful iterations nor how many such iterations there may be between unsuccessful iterations. The obstruction to sharpening the rate of convergence result is that we do not know a priori how much improvement we obtain in $f(x)$ at the successful iterations. We have a sort of multistep R-linear rate of convergence but one for which we do not know and, as our numerical tests reported in section 5.3 suggest, cannot predict the number of intervening steps. For want of an existing term for this notion of convergence, we call it *desultory R-linear convergence*.

More positively, Theorem 4.5 suggests how one can "accelerate" the local convergence of pattern search algorithms. One only need rename the formerly unsuccessful iterates successful iterates and drop the formerly successful iterates from discussion. Then, mirabile dictu, this simple modification makes the successful iterates an R-linearly convergent sequence.

All joking aside, this suggestion is based on the observation that we can rewrite pattern search algorithms to have an inner iteration/outer iteration structure. The outer iterations consist of those iterates at which we reduce Δ_k because no more local reduction in f can be found using the current pattern $\Delta_k P_k$. The inner iterations comprise those iterates which identify simple decrease for some $s_k \in \Delta_k P_k$. Theorem 4.5 allows us to say something about the asymptotic behavior of such "outer" iterations in pattern search.

In that sense, our results are similar to the local convergence for, say, steepest descent with a line search strategy. In steepest descent, the line search is an inner iteration that may require multiple evaluations of the objective in order to generate the ostensible next iterate. In this way both pattern search and steepest descent

generate R-linearly convergent sequences. However, we do not see a way to say anything, asymptotically, about the behavior of the pattern search "inner" iterations. By contrast, one can bound, asymptotically, the number of steps required for the inner iterations of steepest descent devoted to the exercise of a line search strategy, since in the worst case one builds a quadratic model of the objective along the search direction. Once again, for the pattern search analysis the gap lies in the lack of both an explicit estimate of the gradient and a local model of f with which to work. Faster local convergence seems to require better local models.

We close by noting that the only other local convergence result for pattern search similar to Theorem 4.5 of which we are aware is due to Yu [18]. The result is restricted to positive definite quadratic functions (though the extension to nonlinear objectives is straightforward). The fact that f is a quadratic figures explicitly in the derivation of a result similar to (4.8).

5. Numerical results. We now present some numerical experiments that illustrate the practical implications of our convergence analysis. The first round of testing, summarized in section 5.2, supports the analysis; the second round, summarized in section 5.3, shows its limitations. The numerical results regarding the effectiveness of Δ_k as a measure of stationarity, reported in section 5.2, summarize some of the numerical results reported in [6]. The second round of results, given in section 5.3, was generated using the implementation of pattern search from [6].

5.1. The testing environment. Full details of the numerical experiments can be found in [6]. The tests we report here were done with randomly generated quadratic functions. This is a reasonable choice, since we are interested in the local convergence behavior of pattern search, and any function that is twice continuously differentiable looks like a convex quadratic in the neighborhood of an isolated local minimizer. The quadratics tested were of the form $f(x) = x^T Ax + c$, where $A = H^T H$ and $H \in \mathbb{R}^{(n+2)\times n}$ is a matrix with entries that are normal random variates with means of zero and standard deviations of one. The absence of a linear term may be thought of as shifting the quadratic so that the solution lies at the origin, which simplifies our calculations of $\| x_k - x_* \|$. The constant term c is not interesting for the purposes of the optimization but provides a useful tag for identifying individual functions. For the testing in [6], $2 \le n \le 5$; we show a result for $n = 5$.

In addition to randomly generating the entries of the matrix H, we also randomly generated Δ_0 and the entries of the vector x_0. The entries for the starting point x_0 were also normal random variates with means of zero and standard deviations of one. The choice for Δ_0 was an exponential variate with a mean of one. Since the starting points are randomly generated, the absence of a linear term in the quadratic should not unduly influence the outcome of the searches.

The software described in [6] was written in C++ to make use of C++ classes, a convenient way to establish the key features of pattern search and then easily derive specific variants. Several of these variants were implemented and tested, as described in [6]. We show results using `HJSearch`, an implementation of the classical pattern search algorithm of Hooke and Jeeves [8]; `CompassSearch`, the pattern search algorithm described in section 2; and `NLessSearch`, a pattern search algorithm that takes advantage of the fact that a minimal positive basis requires only $n+1$ vectors [11], as opposed to the $2n$ coordinate vectors used in most traditional pattern search methods, including compass search and Hooke and Jeeves.

TABLE 5.1
`HJSearch` *in five variables.*

Δ_k	$\| \nabla f(x_k) \|$	$\| f(x_k) - f(x_*) \|$	$\| x_k - x_* \|$
0.696226813823902	3.718628968450993	3.96639084353257	2.396301558944381
0.348113406911951	1.370661155865317	0.44618879006458	0.698389592313846
0.174056703455976	0.993386770046628	0.19091214014793	0.450386903073632
0.087028351727988	0.236893510661273	0.01477525409286	0.153943970082610
0.043514175863994	0.314026005456998	0.01421309666224	0.119315177505950
0.021757087931997	0.131650296045321	0.00223337373949	0.034002609804365
0.010878543965999	0.042526372212693	0.000 28996796577	0.015791910616849
0.005439271982999	0.032921235371376	0.000 18078437086	0.014678346820778
0.002719635991500	0.012854930063180	0.000 14567060113	0.016582990396810
0.001359817995750	0.005667414556147	0.0000 1023084696	0.003757596625046
0.000679908997875	0.004101406209192	0.00000 429756349	0.002612852810391
0.000339954498938	0.001396318029208	0.000000 50775161	0.000854609846084
0.000169977249469	0.000833651146770	0.000000 49903818	0.000985750547712
0.000084988624734	0.000563050121378	0.0000000 2356890	0.000074244150774
0.000042494312367	0.000112117511534	0.00000000 325088	0.000043510325021
0.000021247156184	0.000097664692564	0.00000000 266601	0.000032689236837
0.000010623578092	0.000035578092711	0.000000000 26108	0.000013878637584
0.000005311789046	0.000010624256256	0.000000000 15362	0.000017458315183

5.2. Measuring stationarity. The first question we ask is, how effective is Δ_k as a measure of stationarity? Not too surprisingly, the results of our tests showed that Δ_k is a reliable measure of progress toward a solution. Furthermore, our numbers make quite clear the R-linear convergence of the subsequence of unsuccessful iterates.

After any unsuccessful iteration, a pattern search method is required to reduce Δ_k. We used the standard reduction factor of $\frac{1}{2}$ so that after an unsuccessful iteration, $\Delta_{k+1} = \frac{1}{2}\Delta_k$. Before proceeding to the next iteration, we recorded the value of Δ_k, $\| \nabla f(x_k) \|$, $| f(x_k) - f(x_*) |$, and $\| x_k - x_* \|$ (though since we knew $x_* \equiv 0$, we simply had to compute $\| x_k \|$). Representative results from one particular test are given in Table 5.1.

The point of the results we report in Table 5.1 is not to demand close scrutiny of each entry but rather to demonstrate the trends in each of the four quantities measured. We clearly see the R-linear behavior the analysis tells us to expect: by the time we halve Δ_k, we have roughly halved the error in the solution.

We report here the results from only one experiment, but they are representative of results from ten thousand runs over multiple quadratics, in multiple dimensions, from multiple starting points, with multiple choices of Δ_0, using four different pattern search methods [6]. We found that across all these tests, Δ_k gave us a consistent measure of the accuracy of the solution. Further, these results conform both with a long-standing recommendation for a stopping criterion (see [8]) as well as with our observations when applying pattern search algorithms to general (i.e., nonquadratic) functions.

One practical benefit of using Δ_k as a measure of stationarity is that it is already present in pattern search algorithms; no additional computation is required. Another good reason for using Δ_k as a measure of stationarity is that it is largely insusceptible to numerical error. Since pattern search methods often are recommended when the evaluations of the objective function are subject to numerical "noise," the fact that Δ_k will not be affected by numerical noise in the computed values of the objective function suggests that Δ_k provides a particularly suitable stopping criterion. One last observation to be made about the practical utility of Δ_k as a measure of stationarity is

that pattern searches require only ranking, or order, information to drive the search—no numeric values for the objective are necessary [11]. In such a setting, Δ_k is a feasible measure of progress, whereas measures based on the numeric values of the objective function are not.

We close with the observation that the conditioning of the Hessian does play a role in the progress of the search, as is true for steepest descent. For the example in Table 5.1 this is not an issue since the smallest singular value for the Hessian is 0.4661 and the condition number of the Hessian is 37.5767. However, in limited tests, we parameterized the Hessian of a two-dimensional quadratic to control the condition number of A. As the Hessian became increasingly less well-conditioned, the number of iterations between each unsuccessful iteration grew; however, we still saw the same trends evident in Table 5.1. The effect the conditioning of the Hessian has on our experimental results should not be surprising since the constant $c_{4.1}$ in Proposition 4.1 explicitly depends on $\sigma_{\min}$; as $\sigma_{\min} \to 0$, $c_{4.1} \to \infty$. For a similar observation regarding the connection between conditioning and the performance of steepest descent with finite-difference gradients, see [3].

5.3. How many successful iterates? Theorem 4.5 says that the subsequence of unsuccessful iterates converges R-linearly once we are in a neighborhood of a solution. A natural question to then ask is, how many iterations occur in practice between each iterate included in this subsequence? If a reasonable a priori bound for the number of intervening iterations could be derived, then we could establish the rate of convergence for the entire sequence of iterates. Since we could see no analytical approach to answering this question, as discussed at the end of section 4, we decided to conduct some numerical studies. As it happens, our experiments shed little light on the question. We give only a few specific results in Figures 5.1–5.2.

In all instances, we terminate the search when $\Delta_{k+1} < 2 \times 10^{-8}$. Along the horizontal axis, we list the number of unsuccessful iterations; i.e., the number of times we halve Δ_k before it is less than the stopping tolerance. Each bar then represents the number of successful iterations that preceded an unsuccessful iteration plus the (single) unsuccessful iteration so that summing all the entries gives us the total number of iterations for the search.

Notice that for the three algorithms we tested the scale on the vertical axes varies considerably. For `NLessSearch`, the number of successful iterations preceding an unsuccessful iteration can be considerably higher than, say, for `HJSearch`, but over all of our tests, the results are mixed. We cannot predict how many successful iterations may precede an unsuccessful iteration, nor does there seem to be any particular trend. However, a few useful observations emerged.

One trend that can be seen in Figures 5.1–5.2 is the apparent superiority, in terms of the total number of iterations required to satisfy our stopping criterion, of the algorithm of Hooke and Jeeves when applied to quadratic functions. This is consistent with the results in [6]. As yet we can offer no analytical explanation for this behavior, but it seems that the "pattern step" in the Hooke and Jeeves algorithm, which captures some limited history of prior successes and potentially enables a much longer trial step than allowed by the core pattern, helps the overall progress of the search.

Another point is illustrated by the example shown in Figure 5.2. The poor scaling of the graphs in Figure 5.2, a consequence of the relatively huge number of iterations taken before the first reduction in Δ, precludes close examination—but that underscores the point we wish to make.

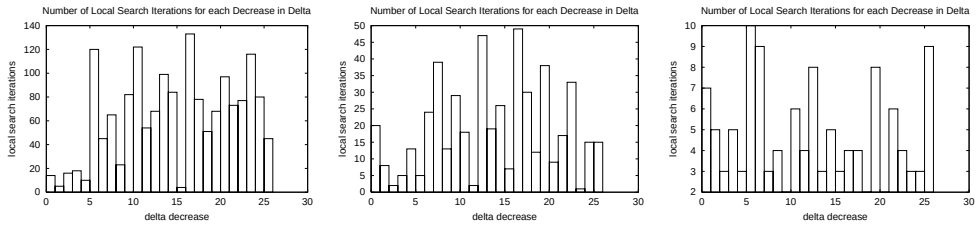


FIG. 5.1. `NLessSearch` *(left),* `CompassSearch` *(middle), and* `HJSearch` *(right) in eight variables.*

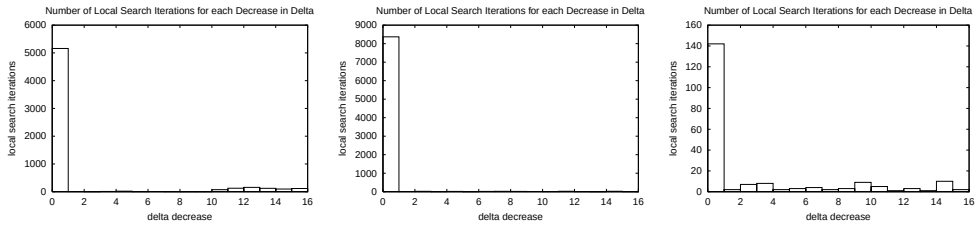


FIG. 5.2. `NLessSearch` *(left),* `CompassSearch` *(middle), and* `HJSearch` *(right) in four variables.*

The relatively huge number of successful iterations taken before Δ_k is ever reduced is due to the small initial value of Δ_0. For our experiments, the value of Δ_0 was drawn randomly. In this example it is so small (0.001128116614106) that initially there is a long sequence of successful iterations, but progress is remarkably slow because we start with such a small choice of Δ_0 that all the trial steps are quite short. After the first reduction in Δ, the number of iterations between each subsequent reduction in Δ demonstrates the same unpredictable behavior we see in the graphs in Figure 5.1.

This suggests two conjectures. The first is that in general it is best to start the search with a relatively large value of Δ_0. This is consistent with pattern search/direct search lore (e.g., see the discussion found in [17] on choosing the size of the initial simplex). The second conjecture is that there is merit to allowing Δ_k to increase so as to recover from an inappropriate choice of Δ_0. While the analyses in [16, 11] support such a specification for pattern search algorithms, most analyses require Δ_k to be monotonically nonincreasing. Furthermore, we are aware of only two publicly available implementations of pattern search methods [7, 9] that allow Δ_k to increase. Even the analysis we present here assumes that eventually Δ_k is monotonically nonincreasing. The practical compromise, implicit in Hypothesis 1, is that we allow increases in Δ_k only up to some finite number of iterations, after which we require Δ_k to be nonincreasing. This allows for some initial adjustments in the step-length control parameter if the first few iterations of the search suggest that the choice of Δ_0 may have been too conservative. However, if we disable any further increases in Δ_k once $k \geq N$, then we preserve the global and local convergence properties presented in sections 3 and 4.

6. Conclusion. The results given here round out the convergence analysis of pattern search. The analysis and numerical experiments reported here show that Δ_k can be used as a reliable stopping criterion. Moreover, these tests show that the correlations predicted by Theorems 3.1 and 4.5 between Δ_k, $\|\nabla f(x_k)\|$, and $\|x_k - x_*\|$ are manifest in practice. These results vindicate the intuition of the early developers of direct search methods.

Acknowledgments. We are indebted to Natalia Alexandrov for a conversation that led to the term "desultory convergence" in connection with Theorem 4.5. We thank Stephen Nash for a lively discussion about stopping criteria.

We also thank both referees and the associate editor for their careful reading of the earlier drafts of this paper. They caught an oversight in one of the proofs and made many helpful suggestions for improving the overall presentation. We greatly appreciate their efforts.

REFERENCES

[1] C. Audet, *Convergence Results for Pattern Search Algorithms Are Tight*, Tech. report 98–24, Department of Computational and Applied Mathematics, Rice University, Houston, TX, 1998.

[2] M. Avriel, *Nonlinear Programming: Analysis and Methods*, Prentice-Hall, Englewood Cliffs, NJ, 1976.

[3] P. T. Boggs and J. J. E. Dennis, *A stability analysis for perturbed nonlinear iterative methods*, Math. Comp., 30 (1976), pp. 199–215.

[4] M. J. Box, D. Davies, and W. H. Swann, *Non-Linear Optimization Techniques*, ICI Monograph 5, Oliver and Boyd, Edinburgh, Scotland, 1969.

[5] C. Davis, *Theory of positive linear dependence*, Amer. J. Math., 76 (1954), pp. 733–746.

[6] E. D. Dolan, *Pattern Search Behavior in Nonlinear Optimization*, Honors Thesis, Department of Computer Science, College of William & Mary, Williamsburg, VA, 1999, accepted with highest honors, http://www.cs.wm.edu/∼va/CS495/.

[7] A. P. Gurson, *Simplex Search Behavior in Nonlinear Optimization*, Honors Thesis, Department of Computer Science, College of William & Mary, Williamsburg, VA, 2000, accepted with highest honors, http://www.cs.wm.edu/∼va/CS495/.

[8] R. Hooke and T. A. Jeeves, *Direct search solution of numerical and statistical problems*, J. Assoc. Comput. Mach., 8 (1961), pp. 212–229.

[9] P. D. Hough, T. G. Kolda, and V. J. Torczon, *Asynchronous parallel pattern search for nonlinear optimization*, SIAM J. Sci. Comput., 23 (2001), pp. 134–156.

[10] T. G. Kolda and V. J. Torczon, *On the convergence of asynchronous parallel pattern search*, SIAM J. Optim., to appear.

[11] R. M. Lewis and V. J. Torczon, *Rank Ordering and Positive Bases in Pattern Search Algorithms*, Tech. report 96–71, Institute for Computer Applications in Science and Engineering, NASA Langley Research Center, Hampton, VA, 1996.

[12] R. M. Lewis and V. Torczon, *Pattern search algorithms for bound constrained minimization*, SIAM J. Optim., 9 (1999), pp. 1082–1099.

[13] R. M. Lewis and V. Torczon, *Pattern search methods for linearly constrained minimization*, SIAM J. Optim., 10 (2000), pp. 917–941.

[14] R. M. Lewis and V. Torczon, *A globally convergent augmented Lagrangian pattern search algorithm for optimization with general constraints and simple bounds*, SIAM J. Optim., 12 (2002), pp. 1075–1089.

[15] S. G. Nash and A. Sofer, *Linear and Nonlinear Programming*, McGraw-Hill, New York, 1996.

[16] V. Torczon, *On the convergence of pattern search algorithms*, SIAM J. Optim., 7 (1997), pp. 1–25.

[17] F. H. Walters, L. R. Parker, Jr., S. L. Morgan, and S. N. Deming, *Sequential Simplex Optimization*, Chemometrics Series, CRC Press, Boca Raton, FL, 1991.

[18] W. Yu, *Positive basis and a class of direct search techniques*, Sci. Sinica, Special Issue of Mathematics, 1 (1979), pp. 53–67.

SIAM J. OPTIM.
Vol. 14, No. 2, pp. 584–607

CONSTRAINT QUALIFICATION, THE STRONG CHIP, AND BEST APPROXIMATION WITH CONVEX CONSTRAINTS IN BANACH SPACES*

CHONG LI† AND K. F. NG‡

Abstract. Several fundamental concepts such as the basic constraint qualification (BCQ), the strong conical hull intersection property (CHIP), and the perturbations for convex systems of inequalities in Banach spaces (over $\mathbb{R}$ or $\mathbb{C}$) are extended and studied; here the systems are not necessarily finite. Their relationships with each other in connection with the best approximations are investigated. As applications, we establish results on the unconstrained reformulation of best approximations with infinitely many constraints in Hilbert spaces; also we give several characterizations of best restricted range approximations in $C(Q)$ under quite general constraints.

Key words. convex inequality system, the strong CHIP, the basic constraint qualification, best approximation, perturbation, best restricted range uniform approximation

AMS subject classifications. Primary, 41A65; Secondary, 41A29

DOI. S1052623402415846

1. Introduction.

For the study of best approximation problems with a finite system of inequality constraints in $\mathbb{R}^N$ (or in Hilbert spaces), the strong CHIP (the strong conical hull intersection property) and other constraint qualification concepts have played important roles in dual reformulation of the best approximation problems. See, e.g., [6, 7, 13, 14, 15, 22, 23, 26, 27]. In this paper these concepts are extended and studied in connection with more general systems. The system (of convex inequalities) that we will focus on is

$$g_i(x) \le 0, \quad i \in I, \tag{CIS}$$

where I is an index set (finite or otherwise), $x \in X$, each g_i is a real continuous convex function on X, and X is a Banach space (say, over the real field $\mathbb{R}$, but later we will also consider the case when X is over the complex field $\mathbb{C}$).

In what follows we always assume that the solution set S of the system (CIS) is nonempty, i.e.,

$$S := \{x \in X : \ g_i(x) \le 0 \quad \text{for all } i \in I\} \ne \emptyset. \tag{1.1}$$

Let $G(\cdot)$ denote the sup-function [18] of $\{g_i\}$:

$$G(x) := \sup_{i \in I} g_i(x) \quad \text{for all } x \in X.$$

Then S is also the solution set of the convex inequality

$$G(x) \le 0. \tag{SCIS}$$

*Received by the editors October 8, 2002; accepted for publication (in revised form) May 14, 2003; published electronically November 6, 2003.

http://www.siam.org/journals/siopt/14-2/41584.html

†Department of Mathematics, Zhejiang University, Hangzhou 310027, People's Republic of China (cli@zju.edu.cn). This author was supported in part by the National Natural Science Foundation of China (grant 10271025).

‡Department of Mathematics, Chinese University of Hong Kong, Hong Kong, People's Republic of China (kfng@math.cuhk.edu.hk). This author was supported by a direct grant (CUHK) and an earmarked grant from the Research Grant Council of Hong Kong.

In this paper we assume throughout that

$$G(x) < +\infty \quad \text{for all } x \in X \tag{1.2}$$

and that G is continuous on X. These blanket assumptions are automatically satisfied if $\{g_i : i \in I\}$ is locally uniformly bounded. Moreover, the continuity of G automatically follows from (1.2) if X is of finite dimension.

Let C be a closed convex subset of X and let K consist of all $x \in C$ satisfying the system (CIS). For a subset Z of X, we use P_Z to denote the projection operator defined by

$$P_Z(x) = \{y \in Z : \|x - y\| = d_Z(x)\},$$

where $d_Z(x)$ denotes the distance from x to Z.

Recently, studies have been done on establishing the dual formulation of the best approximation problem in the setting of real Hilbert spaces; see [6, 7, 13, 14, 15, 26, 27] for finite systems of linear inequalities and [22, 23] for finite systems of nonlinear inequalities. However, there are many problems in Banach spaces (over $\mathbb{R}$ or $\mathbb{C}$) that have infinitely many convex constraints. One typical example is the problem of best restricted range approximations in $C(Q)$, the space of all continuous complex-valued functions defined on a compact metric space Q; see [21, 33, 34, 35, 36, 37]; this problem can be reformulated as an approximation problem with constraints defined by an infinite system of convex inequalities. This motivates us to consider the following question: Can the results on the dual formulation of the best constrained approximation in Hilbert spaces for finite systems be extended to infinite systems in general Banach spaces? We shall study the relationships between the basic constraint qualification (BCQ) and the CHIP in Banach spaces (over $\mathbb{R}$ or $\mathbb{C}$) in section 3. As applications, we establish some results on the unconstrained reformulation of best approximations with infinitely many constraints in Hilbert spaces. This is done in section 4, where we begin with a general result (applicable to both real and complex Hilbert spaces) relating the BCQ and the dual formulation of the best approximation problem. Our result, on the complex Hilbert space X, is in a very general setting: $\{\Omega_i : i \in I\}$ is a family of closed convex subsets of $\mathbb{C}$, $\{h_i : i \in I\} \subseteq X \setminus \{0\}$, C is a closed convex subset of X, and $\hat{C}_i := \{x \in X : \langle h_i, x\rangle \in \Omega_i\}$. Theorem 4.2 shows that the family $\{C, \hat{C}_i : i \in I\}$ has the strong CHIP if and only if a dual formulation in terms of the projections P_C and $P_{(\cap_{i\in I}\hat{C}_i)\cap C}$ holds. It is worth noting in particular that $\{\Omega_i\}$ is not necessarily explicitly given by (CIS) at the outset. Another application of our results is given in section 5, where several characterizations of best restricted range approximations in $C(Q)$ are given for a class of quite general constraints.

To end this section, we describe some basic notation, most of which is standard (cf. [8, 18]). In particular, for a set Z in X (or in $\mathbb{R}^n$), the interior (resp., closure, convex hull, convex cone hull, linear hull, negative polar, boundary) of Z is defined by $\operatorname{int} Z$ (resp., $\overline{Z}$, $\operatorname{conv} Z$, $\operatorname{cone} Z$, $\operatorname{span} Z$, $Z^{\ominus}$, $\operatorname{bd} Z$); the normal cone of Z at z_0 is denoted by $N_Z(z_0)$ and defined by $N_Z(z_0) = (Z - z_0)^{\ominus}$. Let $\operatorname{ext} Z$ denote the set of all extreme points of Z and let $\mathbb{R}_-$ denote the subset of $\mathbb{R}$ consisting of all nonpositive real numbers. For a proper extended real-valued convex function on X, the subdifferential of f at $x \in X$ is denoted by $\partial f(x)$ and defined by

$$\partial f(x) = \{z^* \in X^* : \ f(x) + \langle z^*, y - x\rangle \le f(y) \quad \text{for all } y \in X\},$$

where $\langle z^*, x\rangle$ denotes the value of a functional z^* in X^* at $x \in Z$, i.e., $\langle z^*, x\rangle = z^*(x)$.

Remark 1.1. (a) Let f be a continuous convex function f on X and $x \in X$ with $f(x) = 0$. It is easy to see that $\operatorname{cone}\big(\partial f(x)\big) \subseteq N_{f^{-1}(\mathbb{R}_-)}(x)$ and that the equality holds if f is an affine function or if x is not a minimizer of f; see [8, Corollary 1, p. 56].

(b) The directional derivative of the function f at x in the direction d is denoted by $f'_+(x,d)$:

$$f'_+(x,d) := \lim_{t\to 0+} \frac{f(x+td)-f(x)}{t}. \tag{1.3}$$

We recall [8, Proposition 2.2.7] (see also [28]) that

$$\partial f(x) = \{z^* \in X^* :\ \langle z^*, d\rangle \le f'_+(x,d) \quad \text{for all } d \in X\} \tag{1.4}$$

and

$$f'_+(x,d) = \max\{\langle z^*, d\rangle :\ z^* \in \partial f(x)\}. \tag{1.5}$$

2. Preliminaries. Let $\{A_i :\ i \in J\}$ be a family of subsets of X. The set $\sum_{i\in J} A_i$ is defined by

$$\sum_{i\in J} A_i = \begin{cases} \big\{\sum_{i\in J_0} a_i :\quad a_i \in A_i, \quad J_0 \subseteq J \text{ being finite}\big\} & \text{if } J \neq \emptyset, \\ \{0\} & \text{if } J = \emptyset. \end{cases} \tag{2.1}$$

Consider (CIS) as before with the solution set denoted by S. For $x \in X$, let $I(x)$ denote the set of all active indices $i : I(x) = \{i \in I :\ g_i(x) = G(x) = 0\}$. Following [17, 24], we define

$$N'(x) := \sum_{i\in I(x)} \operatorname{cone}\big(\partial g_i(x)\big), \quad x \in X. \tag{2.2}$$

Note that, by (2.1), $N'(x) = \operatorname{cone}\big(\bigcup_{i\in I(x)} \partial g_i(x)\big)$ if $I(x) \neq \emptyset$ and $N'(x) = \{0\}$ if $I(x) = \emptyset$.

In the remainder of this paper, we let $K := C \cap S$, where S denotes the solution set of (CIS). The following concepts are well known in the case when I is finite or X is of finite dimension; see, e.g., [24, 22, 23].

DEFINITION 2.1. *Let $x \in K$. The system* (CIS) *is said to satisfy the BCQ relative to C at x if*

$$N_K(x) = N_C(x) + N'(x). \tag{2.3}$$

Remark 2.1. (CIS) satisfies the BCQ at each $x \in C \cap \operatorname{int} S$ because (2.3) holds trivially in this case.

The following concept of the strong CHIP is due to [13, 14] in the case when I is finite and plays an important role in optimization theory; see, e.g., [1, 2, 9, 12, 32].

DEFINITION 2.2. *Let $\{C_i :\ i \in I\}$ be a collection of closed convex subsets of X and $x \in \bigcap_{i\in I} C_i$. The collection is said to have the strong CHIP at x if*

$$N_{\bigcap_{i\in I} C_i}(x) = \sum_{i\in I} N_{C_i}(x). \tag{2.4}$$

Remark 2.2. (a) If $g_i(x) < 0$, then $x \in \operatorname{int}(g_i^{-1}(\mathbb{R}_-))$ and $N_{g_i^{-1}(\mathbb{R}_-)}(x) = \{0\}$. Hence

$$\sum_{i \in I(x)} N_{g_i^{-1}(\mathbb{R}_-)}(x) = \sum_{i \in I} N_{g_i^{-1}(\mathbb{R}_-)}(x).$$

(b) Let $x \in C \cap \operatorname{bd} S$. Then

$$\begin{aligned} &\text{The system (CIS) satisfies BCQ relative to } C \text{ at } x \\ &\Longrightarrow \{C, g_i^{-1}(\mathbb{R}_-) : i \in I\} \text{ has the strong CHIP at } x. \end{aligned}$$

(c) Let $x \in C \cap \operatorname{bd} S$ and suppose that, for each $i \in I(x)$, either g_i is affine or there exists $x_i \in C$ such that $g_i(x_i) < 0$ (so $\operatorname{cone}\big(\partial g_i(x)\big) = N_{g_i^{-1}(\mathbb{R}_-)}(x)$ by Remark 1.1). Then

$$\begin{aligned} &\text{The system (CIS) satisfies BCQ relative to } C \text{ at } x \\ &\Longleftrightarrow \{C, g_i^{-1}(\mathbb{R}_-) : i \in I\} \text{ has the strong CHIP at } x. \end{aligned}$$

(This assertion is of course trivial when $x \in C \cap \operatorname{int} S$.)

(d) When each g_i is affine, $\{g_i^{-1}(\mathbb{R}_-) : \ i \in I\}$ has the strong CHIP at x automatically if I is finite. However, this is not necessarily true if I is infinite; see [24, Example 1].

DEFINITION 2.3. *We say that the system* (CIS) *satisfies the Slater condition on C if there exists a point $\bar{x} \in C$ such that $G(\bar{x}) < 0$. In this case, $\bar{x}$ is called a Slater point of* (CIS) *on C.*

The following theorem, which is known (cf. [18, 24]) in the special case when X is of finite dimension, will play a key role in section 5.

THEOREM 2.1. *Assume that I is a compact metric space and that the function $i \mapsto g_i(x)$ is upper semicontinuous on I at each $x \in X$. Let C be a nonempty closed convex subset of X such that* $\operatorname{span} C$ *is of finite dimension. Suppose that there exists a Slater point $\bar{x}$ of* (CIS) *on C. Then the system* (CIS) *satisfies the BCQ relative to C at every point $x \in K$.*

Proof. As the result is trivial if $x \in C \cap \operatorname{int} S$, we may assume that $x \in C \cap \operatorname{bd} S$. We divide the proof into two steps. First we show that

$$N_C(x) + \partial G(x) \subseteq N_C(x) + N'(x) \quad \text{for all } x \in C \cap \operatorname{bd} S. \tag{2.5}$$

Let $\tilde{G}$ and $\tilde{g}_i$, respectively, denote the restrictions of G and g_i on $\operatorname{span} C$, where $i \in I$. Then

$$\tilde{G}(z) = \sup_{i \in I} \tilde{g}_i(z) \quad \text{for all } z \in \operatorname{span} C. \tag{2.6}$$

By assumptions and [18, Theorem 4.4.2, p. 267] (see also [24, Theorem 3.1]), for any $x \in C \cap \operatorname{bd} S$, we have that

$$\partial \tilde{G}(x) = \operatorname{conv}\left(\bigcup_{i \in I(x)} \partial \tilde{g}_i(x) \right). \tag{2.7}$$

For any $y^* \in \partial G(x)$, y^* can be viewed as an element of $\partial \tilde{G}(x)$. Thus, by (2.7), there exist $\tilde{y}_j^* \in \partial \tilde{g}_{i_j}(x)$, $\lambda_j \geq 0$, $i_j \in I(x)$, $j = 1, 2, \ldots, m$, such that $\sum_{j=1}^m \lambda_j = 1$ and

$$\langle y^*, z \rangle = \left\langle \sum_{j=1}^m \lambda_j \tilde{y}_j^*, z \right\rangle \quad \text{for all } z \in \operatorname{span} C. \tag{2.8}$$

Noting

$$\langle \tilde{y}_j^*, z\rangle \le \tilde{g}'_{i_j+}(x,z) = g'_{i_j+}(x,z) \quad \text{for all } z \in \operatorname{span} C, \tag{2.9}$$

and making use of the Hahn–Banach extension theorem, there exists $y_j^* \in X^*$ satisfying

$$\langle y_j^*, z\rangle = \langle \tilde{y}_j^*, z\rangle \quad \text{for all } z \in \operatorname{span} C \tag{2.10}$$

and such that $\langle y_j^*, z\rangle \le g'_{i_j+}(x,z)$ holds for all z in X. This implies that $y_j^* \in \partial g_{i_j}(x)$. Let $y_0^* = y^* - \sum_{j=1}^m \lambda_j y_j^*$. Then, by (2.8) and (2.10), one has

$$\langle y_0^*, z\rangle = \left\langle y^* - \sum_{j=1}^m \lambda_j y_j^*, z \right\rangle = 0 \quad \text{for all } z \in C - x,$$

in particular, $y_0^* \in N_C(x)$. This implies that $y^* = y_0^* + \sum_{j=1}^m \lambda_j y_j^* \in N_C(x) + \operatorname{cone}\big(\bigcup_{i\in I(x)} \partial g_i(x)\big)$; hence (2.5) is established.

Next, by the assumed Slater condition, it follows from [14, Proposition 2.3] (the proof given there is valid for an arbitrary Banach space although it was stated for Hilbert spaces) that $\{C, S\}$ has the strong CHIP at every point $x \in C \cap \operatorname{bd} S$:

$$N_K(x) = N_C(x) + N_S(x) \quad \text{for all } x \in C \cap \operatorname{bd} S. \tag{2.11}$$

Since $G(\bar{x}) < 0$, Remark 1.1(a) implies

$$N_S(x) = \operatorname{cone}\big(\partial G(x)\big) \quad \text{for all } x \in C \cap \operatorname{bd} S.$$

Then, by (2.11) and (2.5),

$$N_K(x) = N_C(x) + \operatorname{cone}\big(\partial G(x)\big) = N_C(x) + N'(x).$$

Thus Theorem 2.1 is proved. □

In the remainder of this paper, we will assume that X is a Banach space over the complex field $\mathbb{C}$ or the real field $\mathbb{R}$. When X is a Banach space over the complex field $\mathbb{C}$, let X_R denote the corresponding real Banach space by restricting the scalar multiplication to the reals. In this case, for any subset Z of X and $z_0 \in X$, one has two different versions for normal cones:

$$\tilde{N}_Z(z_0) = \{z^* \in X_R^* : \langle z^*, x - z_0\rangle \le 0 \quad \text{for all } x \in X\}, \tag{2.12}$$

$$N_Z(z_0) = \{z^* \in X^* : \operatorname{Re}\langle z^*, x - z_0\rangle \le 0 \quad \text{for all } x \in X\}. \tag{2.13}$$

Likewise, if f is a proper convex function on X and $x \in X$, then one can define

$$\tilde{\partial} f(x) = \{z^* \in X_R^* : \ f(x) + \langle z^*, y - x\rangle \le f(y) \quad \text{for all } y \in X\}, \tag{2.14}$$

$$\partial f(x) = \{z^* \in X^* : \ f(x) + \operatorname{Re}\langle z^*, y - x\rangle \le f(y) \quad \text{for all } y \in X\}. \tag{2.15}$$

Finally in addition to (2.2), one can define $\tilde{N}'(x)$ in the above manner. In view of the Bohnenblust–Sobczyk theorem ($x^* \mapsto \operatorname{Re} x^*$ is a real-isometry from X^* onto X_R^*; cf. [39, p. 192]), such distinctions are immaterial; for example, regarding Definition 2.1,

the system (CIS) in X satisfies the BCQ relative to C at x in the sense of (2.3) if and only if it does in X_R. Thus, the results in this section, such as Theorem 2.1, can be applied to spaces over $\mathbb{C}$.

We now introduce some new concepts. Recall that $K := C \cap S$, where S denotes the solution set of (CIS). The index set I is not assumed to have any topological structure.

DEFINITION 2.4. *Let $x \in K$. An element $d \in X$ is called*

(a) *a linearized feasible direction of* (CIS) *at x if*

$$\operatorname{Re}\langle z^*, d\rangle \le 0 \quad \text{for all } z^* \in \bigcup_{i\in I(x)} \operatorname{ext}\partial g_i(x); \tag{2.16}$$

(b) *a sequentially feasible direction of K at x if there exist a sequence $d_k \to d$ and a sequence of positive real numbers $\delta_k \to 0$ such that $\{x + \delta_k d_k\} \subseteq K$.*

Remark 2.3. When I is finite and each g_i is differentiable at x, the definition of a linearized feasible direction of (CIS) at x in a real space X coincides with the corresponding definition introduced in [25, 38]; see also [22].

Let $\mathrm{LFD}(x)$ (resp., $\mathrm{SFD}(x)$) denote the set of all d satisfying (a) (resp., (b)) in Definition 2.4. Note that $\mathrm{LFD}(x)$ is a closed convex cone (so it contains the origin) while $\mathrm{SFD}(x)$ is a closed cone (but not necessarily convex). Note also that $\mathrm{LFD}(x) = X$ if $I(x) = \emptyset$.

DEFINITION 2.5. *Let $x \in K$. Let $K_S(x)$ and $K_L(x)$ be defined, respectively, by*

$$K_S(x) = \left(x + \overline{\operatorname{conv}(\mathrm{SFD}(x))}\right) \bigcap C \tag{2.17}$$

and

$$K_L(x) = (x + \mathrm{LFD}(x)) \bigcap C. \tag{2.18}$$

Note that the two sets are closed convex sets. We have the following well-known inclusion relationship.

PROPOSITION 2.1. *Let $x \in C \cap S$. Then* $\mathrm{SFD}(x) \subseteq \mathrm{LFD}(x)$ *and*

$$K \subseteq K_S(x) \subseteq K_L(x). \tag{2.19}$$

Let $x_0 \in K$ and suppose that $I(x_0) \neq \emptyset$. In the study of the system (CIS), it would be useful to consider the following associated (linearized) system on X:

$$\operatorname{Re}\langle z^*, x - x_0\rangle \le 0, \quad z^* \in \bigcup_{i\in I(x_0)} \operatorname{ext}\partial g_i(x_0). \tag{2.20}$$

Let

$$\hat{S}_{z^*}(x_0) := \{x \in X : \operatorname{Re}\langle z^*, x - x_0\rangle \le 0\} \quad \text{for all } z^* \in \bigcup_{i\in I(x_0)} \operatorname{ext}\partial g_i(x_0)$$

and

$$\hat{S}(x_0) := \bigcap \left\{ \hat{S}_{z^*}(x_0) : z^* \in \bigcup_{i\in I(x_0)} \operatorname{ext}\partial g_i(x_0) \right\}. \tag{2.21}$$

Moreover, we define $\hat{S}(x_0) = X$ if $I(x_0) = \emptyset$. Then

$$x_0 + \mathrm{LDF}(x_0) = \hat{S}(x_0) \quad \text{and} \quad K_L(x_0) = \hat{S}(x_0) \cap C, \tag{2.22}$$

whether or not $I(x_0) \neq \emptyset$. For our convenience we state the following elementary lemma. We omit its proof as it is straightforward.

LEMMA 2.1. *Let $z^* \in X^*$, $x_0 \in X$, and let $\varphi : X \to \mathbb{R}$ be defined by*

$$\varphi(x) = \mathrm{Re}\,\langle z^*, x - x_0\rangle \quad \text{for all } x \in X.$$

Then $\partial\varphi(x_0) = z^*$. Consequently, $N'(x_0)$ defined by (2.2) *with respect to the system* (CIS) *coincides with the corresponding one with respect to the system* (2.20).

Recall that the duality map J from X to 2^{X^*} is defined by

$$J(x) := \{x^* \in X^* : \ \langle x^*, x\rangle = \|x\|^2, \ \|x^*\| = \|x\|\}. \tag{2.23}$$

In fact, $J(x) = \partial\phi(x)$, where $\phi(x) := \frac{1}{2}\|x\|^2$. Thus a Banach space X is smooth if and only if for each $x \in X$ the duality map is single-valued.

The following proposition will be useful later. This result was established independently by Deutsch [10] and Rubenstein [29] (see also [3]). We thank the two anonymous referees for their helpful comments. One of the referees kindly suggested the above references as well as the formulation of Corollary 4.3.

PROPOSITION 2.2. *Let Z be a closed convex set in X. Then for any $x \in X$, $z_0 \in P_Z(x)$ if and only if $z_0 \in Z$ and there exists $x^* \in J(x - z_0)$ such that $\mathrm{Re}\,\langle x^*, z - z_0\rangle \leq 0$ for any $z \in Z$; that is, $J(x - z_0) \cap N_Z(z_0) \neq \emptyset$. In particular, when X is smooth, $z_0 \in P_Z(x)$ if and only if $z_0 \in Z$ and $J(x - z_0) \in N_Z(z_0)$.*

3. Best constrained approximations in Banach spaces. Before proving the main theorem of this section we recall two lemmas. These two lemmas were stated in the Hilbert space setting in [22, 23]. The proof given in [22, Theorem 3.1] for the first lemma is valid for Banach spaces, while the proof of the second lemma given in [23, Lemma 3.1] for Hilbert space will need to be modified to suit our purpose here.

LEMMA 3.1. *Let K be a nonempty closed convex subset of X, and let $x_0 \in K$. Then, for any $x \in X$, we have*

$$x_0 \in P_K(x) \Longleftrightarrow x_0 \in P_{K_S(x_0)}(x). \tag{3.1}$$

LEMMA 3.2. *Suppose that X is reflexive and smooth. Let C be a closed convex set, let $x_0 \in C$, and let T_1, T_2 be closed convex cones in X. Then the following statements are equivalent:*

(i) $C \cap (x_0 + T_1) \subseteq C \cap (x_0 + T_2)$.

(ii) $x_0 \in P_{C\cap(x_0+T_1)}(x)$ *whenever* $x \in X$ *and* $x_0 \in P_{C\cap(x_0+T_2)}(x)$.

Proof. We modify the proof that is given in [23] for the special case when X is a Hilbert space. Since X is assumed smooth, the map $x \mapsto J(x)$ is a (single-valued) weak*-continuous map from X to X^*.

Suppose that (i) does not hold; take $\bar{x} \in C \cap (x_0 + T_1)$ such that $\bar{x} \notin x_0 + T_2$. Let $x_0 + e \in P_{x_0+T_2}(\bar{x})$, where $e \in T_2$. Denote $h = \bar{x} - (x_0 + e)$. Then, by Proposition 2.2,

$$\langle J(h), (x_0 + z) - (x_0 + e)\rangle \leq 0 \quad \text{for all } z \in T_2.$$

Therefore,

$$\langle J(h), e\rangle = 0, \tag{3.2}$$

and $P_{x_0+T_2}(x_t) = x_0$ for each $t > 0$, where $x_t := x_0 + th$. By (ii), it follows that

$$P_{C\cap(x_0+T_1)}(x_t) = x_0. \tag{3.3}$$

Let $\bar{x}_t = (x_t - \bar{x})/t$ for $t > 0$. Then $\bar{x}_t = (1-1/t)h - e/t$ and $\lim_{t\to+\infty} \bar{x}_t = h$; hence,

$$\lim_{t\to+\infty} \langle J(\bar{x}_t) - J(h), h + e\rangle = 0. \tag{3.4}$$

Consequently, by (3.2) and (3.4),

$$\begin{aligned}\|\bar{x}_t\|^2 &= \langle J(\bar{x}_t), \bar{x}_t\rangle \\ &= \langle J(\bar{x}_t), h\rangle - \langle J(\bar{x}_t) - J(h), (h+e)/t\rangle - \langle J(h), (h+e)/t\rangle \\ &\le \|\bar{x}_t\| \cdot \|h\| + |\langle J(\bar{x}_t) - J(h), (h+e)\rangle|/t - \|h\|^2/t \\ &< \|\bar{x}_t\| \cdot \|h\|,\end{aligned}$$

and so $\|x_t - \bar{x}\| < t\|h\|$ for $t > 1$ large enough. Since $\bar{x} \in C\cap(x_0+T_1)$, this contradicts (3.3). The proof is complete. □

Remark 3.1. The result of Lemma 3.2 characterizes the smoothness of X (among reflexive Banach spaces). Indeed, suppose that there exists a unit vector $x_0 \in X$ such that $J(x_0)$ contains two distinct elements $x_1^*,\ x_2^*$. Write $x_0^* = \frac{x_1^*+x_2^*}{2},\ x_3^* = \frac{2}{3}x_1^* + \frac{1}{3}x_2^*$, and define

$$T_1 = \{x \in X :\ \langle x_3^*, x\rangle \ge 0\}, \quad T_2 = \{x \in X :\ \langle x_0^*, x\rangle \ge 0\}.$$

Then $x_0+T_1 \not\subseteq x_0+T_2$ although, for each $x \in X$, $x_0 \in P_{x_0+T_2}(x) \Longrightarrow x_0 \in P_{x_0+T_1}(x)$. In fact, if $x_0 \in P_{x_0+T_2}(x)$, then, by Proposition 2.2, there exists $x^* \in J(x - x_0)$ such that $\langle x^*, z\rangle \le 0$ for all $z \in T_2$. This implies that $x^* = -\|x - x_0\|x_0^*$; hence $\langle x_0^*, x_0 - x\rangle = \|x - x_0\|$. Consequently, $\langle x_i^*, x_0 - x\rangle = \|x - x_0\|$ for $i = 1, 2, 3$. Thus, for each $z \in T_1$,

$$\|x - (x_0 + z)\| \ge \langle x_3^*, x_0 + z - x\rangle \ge \langle x_3^*, x_0 - x\rangle = \|x - x_0\|;$$

hence $x_0 \in P_{x_0+T_1}(x)$, as claimed.

Let Z^* be a subset of X^* and $Z \subseteq X$. Let $z^*|_Z$ denote the restriction of z^* on Z; i.e., $z^*|_Z$ is viewed as a functional defined on Z instead of X. Set

$$Z^*|_Z = \{z^*|_Z :\ z^* \in Z^*\}. \tag{3.5}$$

Recall that $K := C \cap S$, where S denotes the solution set of (CIS). Let $x_0 \in K$, and let $\hat{S}(x_0)$ and $\hat{S}_{z^*}$ be defined as in (2.21). By Remark 2.2(c) (applied to the system (2.20) in place of (CIS)), we have the following equivalence:

(3.6) The system (2.20) satisfies the BCQ relative to C at x_0 if and only if the family $\{C, \hat{S}_{z^*}(x_0) :\ z^* \in \bigcup_{i\in I(x_0)} \operatorname{ext}\partial g_i(x_0)\}$ has the strong CHIP at x_0.

Thus one has (ii)$\Longleftrightarrow$(ii*) in the following theorem.

THEOREM 3.1. *Let $x_0 \in K$. Consider the following statements:*

(i) *the system* (CIS) *satisfies the BCQ relative to C at x_0;*

(ii) *$K_S(x_0) = K_L(x_0)$, and the family $\{C, \hat{S}_{z^*}(x_0) :\ z^* \in \bigcup_{i\in I(x_0)} \operatorname{ext}\partial g_i(x_0)\}$ has the strong CHIP at x_0;*

(ii*) *$K_S(x_0) = K_L(x_0)$, and the system* (2.20) *satisfies the BCQ relative to C at x_0;*

(iii) *for each* $x \in X$, $x_0 \in P_K(x)$ *if and only if*

$$J(x - x_0) \cap \big(N_C(x_0) + N'(x_0)\big) \neq \emptyset; \tag{3.7}$$

(iv) *for each* $x \in X$, $x_0 \in P_K(x)$ *if and only if*

$$J(x - x_0)|_{C-x_0} \cap \big(N_C(x_0)|_{C-x_0} + N'(x_0)|_{C-x_0}\big) \neq \emptyset. \tag{3.8}$$

Then the following implications hold:

(1) (i)$\Longrightarrow$(iii)$\Longrightarrow$(iv); (ii)$\Longleftrightarrow$(ii*)$\Longrightarrow$(iii)$\Longrightarrow$(iv);

(2) (i)$\Longleftrightarrow$(ii)$\Longrightarrow$(iii)$\Longrightarrow$(iv) *if* X *is reflexive;*

(3) (i)$\Longleftrightarrow$(ii)$\Longleftrightarrow$(iii)$\Longrightarrow$(iv) *if* X *is both reflexive and smooth.*

Proof. The results are trivial when $x_0 \in C \cap \operatorname{int} S$ since each of (i)–(iv) in Theorem 3.1 holds automatically. Hence we assume that $x_0 \in C \cap \operatorname{bd} S$.

(1) Suppose that (i) holds. Then (3.7) can be rewritten as $J(x-x_0)\cap N_K(x_0) \neq \emptyset$; hence (iii) holds by Proposition 2.2. Therefore (i)$\Longrightarrow$(iii). Thus assuming that (ii*) holds, and applying this implication to the system (2.20) in place of (CIS), one has, for each $x \in X$,

$$x_0 \in P_{C\cap\hat{S}(x_0)}(x) \Longleftrightarrow J(x - x_0) \bigcap (N_C(x_0) + N'(x_0)) \neq \emptyset \tag{3.9}$$

(see Lemma 2.1). Consequently, by (2.22),

$$x_0 \in P_{K_L(x_0)}(x) \Longleftrightarrow J(x - x_0) \bigcap (N_C(x_0) + N'(x_0)) \neq \emptyset. \tag{3.10}$$

Further, by (3.1) and the assumption $K_S(x_0) = K_L(x_0)$ in (ii), we have that, for each $x \in X$,

$$x_0 \in P_K(x) \Longleftrightarrow x_0 \in P_{K_L(x_0)}(x). \tag{3.11}$$

Therefore, combining (3.10) and (3.11), we have established that (ii)$\Longleftrightarrow$(ii*)$\Longrightarrow$(iii).

Since (3.7) implies (3.8), to prove that (iii) implies (iv), it suffices to show that if (3.8) holds, then $x_0 \in P_K(x)$. By (3.8) and $N_C(x_0)|_{C-x_0} + N'(x_0)|_{C-x_0} \subseteq N_K(x_0)|_{C-x_0}$, we obtain that there exists $x^* \in J(x - x_0)$ such that

$$\operatorname{Re}\langle x^*, x' - x_0\rangle \leq 0 \quad \text{for all } x' \in K. \tag{3.12}$$

Hence, for any $x' \in K$, we have that

$$\|x^*\| \cdot \|x - x_0\| = \operatorname{Re}\langle x^*, x - x_0\rangle \leq \operatorname{Re}\langle x^*, x - x'\rangle \leq \|x^*\| \cdot \|x - x'\|.$$

This shows that $x_0 \in P_K(x)$, as required. Therefore (iii)$\Longrightarrow$(iv).

(2) Suppose that (3) is valid, and that X is reflexive. Then, by a known result in Banach space theory (cf. [16, p. 186]), there exists an equivalent norm on X such that X is smooth under the new norm. Then (3) implies that (i) and (ii) are equivalent. Other implications in (2) have already been proved in (1).

(3) By statement (1), we only need to show that (iii) implies (i) and (ii*). Suppose that (iii) holds. Let $z^* \in N_K(x_0)$. By the reflexivity of X, there exists $\bar{x} \in X$ such that $\langle z^*, \bar{x}\rangle = \|z^*\|\|\bar{x}\| = \|z^*\|^2$. Let $x = \bar{x} + x_0$. Then $z^* = J(x - x_0)$ by the smoothness, and $x_0 \in P_K(x)$ by Proposition 2.2. It follows from (iii) that $z^* \in N_C(x_0) + N'(x_0)$. This shows that $N_K(x_0) \subseteq N_C(x_0) + N'(x_0)$ and so (i) holds. Therefore (iii)$\Longrightarrow$(i).

To prove (iii)$\Longrightarrow$(ii*), noting from (2.19) that $K \subseteq K_S(x_0) \subseteq K_L(x_0)$, we have, for each $x \in X$,

$$x_0 \in P_{K_L(x_0)}(x) \Longrightarrow x_0 \in P_{K_S(x_0)}(x) \Longrightarrow x_0 \in P_K(x). \tag{3.13}$$

Conversely, let $x_0 \in P_K(x)$. Then, by (iii), $J(x - x_0) \in N_C(x_0) + N'(x_0)$. By (2.22) and (2.16), one has $N'(x_0) \subseteq N_{K_L(x_0)}(x_0)$. Since $K_L(x_0) \subseteq C$, it follows that $J(x - x_0) \in N_{K_L(x_0)}(x_0)$. Consequently, by Proposition 2.2, $x_0 \in P_{K_L(x_0)}(x)$. Hence, we have proved that, for each $x \in X$,

$$x_0 \in P_{K_S(x_0)}(x) \Longleftrightarrow x_0 \in P_{K_L(x_0)}(x) \Longleftrightarrow x_0 \in P_K(x). \tag{3.14}$$

It follows from Lemma 3.2 that $K_S(x_0) = K_L(x_0)$. Furthermore, by (3.14) and (iii), we obtain that $x_0 \in P_{K_L(x_0)}(x) \Longleftrightarrow J(x - x_0) \in N_C(x_0) + N'(x_0)$. Applying the just proved implication (iii)$\Longrightarrow$(i), we see that the system (2.20) satisfies the BCQ relative to C at x_0. This completes the proof of (iii)$\Longrightarrow$(ii*). □

Remark 3.2. The proof given for Theorem 3.1 is valid even if $I(x_0) = \emptyset$.

Remark 3.3. Example 3.1 (a) and (b) below show that neither the condition that X is smooth nor the condition that X is reflexive can be dropped for the implication (iii)$\Longrightarrow$(i) in Theorem 3.1.

Example 3.1 (cf. [24, Example 1]). (a) Let X be the Banach space $\mathbb{R}^2$ endowed with the l_1 norm defined as follows:

$$\|x\| = |t_1| + |t_2| \quad \text{for all } x = (t_1, t_2) \in \mathbb{R}^2. \tag{3.15}$$

Let $C = X,\ I = \{1, 2, \dots\}$, and define

$$g_i(x) = t_1 + \frac{1}{i} t_2 \quad \text{for all } x = (t_1, t_2) \in \mathbb{R}^2,\ \ i \in I.$$

Then, for any $x = (t_1, t_2) \in \mathbb{R}^2$,

$$G(x) := \sup_{i \in I} g_i(x) = \begin{cases} t_1 & \text{if } t_2 \le 0, \\ t_1 + t_2 & \text{if } t_2 \ge 0; \end{cases}$$

in particular, G is continuous. Furthermore,

$$K := C \cap S = S = \{x = (t_1, t_2) \in X :\ t_1 \le 0,\ t_1 + t_2 \le 0\}.$$

Take $x_0 = (0, 0)$. Then

$$N_K(x_0) = \{(t_1, t_2) \in \mathbb{R}^2 :\ 0 \le t_2 \le t_1\},$$

$$N'(x_0) = \{(t_1, t_2) \in \mathbb{R}^2 :\ 0 < t_2 \le t_1\} \cup \{(0, 0)\}.$$

Hence, the system (CIS) does not satisfy the BCQ relative to C at x_0. On the other hand, for any $x = (t_1, t_2) \in X$, from (3.15), $x_0 \in P_K(x)$ if and only if x lies in the first quadrant W of $\mathbb{R}^2$. Moreover, one has

$$J(x - x_0) = \begin{cases} [-1, 1] \times \operatorname{sgn} t_2 & \text{if } x = (0, t_2) \neq 0, \\ \operatorname{sgn} t_1 \times [-1, 1] & \text{if } x = (t_1, 0) \neq 0, \\ \big(\operatorname{sgn} t_1, \operatorname{sgn} t_2\big) & \text{if } x = (t_1, t_2),\ t_1 \neq 0, t_2 \neq 0, \end{cases}$$

where $\operatorname{sgn} t$ denotes the sign of t. Hence (3.7) holds if and only if $x \in W$. Thus, (iii) of Theorem 3.1 holds.

(b) Let X be any nonreflexive Banach space. By the well-known James theorem (cf. [19]; see also [31, Corollary 2.4, p. 99]), there exists a nonzero functional $x_0^* \in X^*$ such that it does not attain its norm on the unit ball of X. Set $C = \{x \in X : \langle x, x_0^* \rangle \leq 2\}$, $I = \{0, 1, \dots\}$. Define

$$g_0(x) = -\langle x, x_0^* \rangle, \quad x \in X,$$

and

$$g_i(x) = \langle x, x_0^* \rangle - \frac{1}{i}, \quad x \in X,\ i = 1, 2, \dots.$$

Then

$$K = C \cap S = \{x \in X :\ \langle x, x_0^* \rangle = 0\}.$$

Taking $x_0 = 0$, we have that $I(x_0) = \{0\}$ and that

$$N_K(x_0) = \{x^* \in X^* :\ \langle x, x^* \rangle = 0 \quad \text{for all } x \in K\} = \operatorname{span}\{x_0^*\},$$

$$N_C(x_0) = 0, \quad N'(x_0) = \operatorname{cone}\big(\partial g_0(x_0)\big) = \{-\lambda x_0^* :\ \lambda \geq 0\}. \tag{3.16}$$

In particular,

$$N_C(x_0) + N'(x_0) \neq N_K(x_0),$$

and hence the system

$$g_i(x) \leq 0, \quad i = 0, 1, 2, \dots,$$

does not satisfy the BCQ relative to C at x_0. Moreover, by our choice of x_0^* and (3.16), it is easy to see that (3.7) holds if and only if $x - x_0 = 0$. Recalling from [31, p. 100] that $P_K(z) \neq \emptyset$ implies $z \in K$, it follows that (iii) holds. □

Remark 3.4. When I is finite and g_i is both convex and differentiable for each $i \in I$, the equivalence of (i) and (ii) in Theorem 3.1 was established in [22] for Hilbert spaces. Theorem 3.1 is new even in the case when $C = X = \mathbb{R}^n$. Two new features here are worth noting: I is not necessarily finite and g_i is not necessarily smooth. Moreover, our treatments are in the general Banach space setting.

4. Best constrained approximation in Hilbert spaces. Throughout this section, let X denote a Hilbert space (over $\mathbb{R}$ or $\mathbb{C}$). Let C be a closed convex subset of X and let K be the set of $x \in C$ that satisfies (CIS). Since X is a Hilbert space, $X^* = X$. In particular, (2.15) can be redefined as

$$\partial f(x) = \{z \in X :\ f(x) + \operatorname{Re}\langle z, y - x \rangle \leq f(y) \quad \text{for all } y \in X\}.$$

Similarly, $N_Z(z_0) = \{y \in X :\ \operatorname{Re}\langle y, z - z_0 \rangle \leq 0 \quad \text{for all } z \in Z\}$.

Dual formulation of the constrained best approximation problem in Hilbert spaces has been extensively investigated for finite systems of linear inequality constraints, e.g., [6, 7, 13, 14, 15, 26, 27], and for that of nonlinear inequalities, e.g., [22, 23]. In this section, we will establish similar results for infinite systems of convex inequalities. The first main result is as follows. Notation is as in the preceding sections (see (2.17), (2.18), and (2.21) in particular).

THEOREM 4.1. *Let $x_0 \in K$. Then the following statements are equivalent:*

(i) *the system* (CIS) *satisfies the BCQ relative to C at x_0;*

(ii) *$K_S(x_0) = K_L(x_0)$ and the family $\{C, \hat{S}_{z^*}(x_0) : \ z^* \in \bigcup_{i\in I(x_0)} \operatorname{ext} \partial g_i(x_0)\}$ has the strong CHIP at x_0;*

(iii) *for any $x \in X$, $P_K(x) = x_0$ if and only if there exists a finite (possibly empty) set $I_0 \subseteq I(x_0)$ such that $P_C(x - \sum_{i\in I_0} \lambda_i h_i) = x_0$ for some $\lambda_i \geq 0$ and $h_i \in \partial g_i(x_0)$ with each $i \in I_0$.*

Proof. By Theorem 3.1, it suffices to show that (3.7) holds if and only if there exists a finite set $I_0 \subseteq I(x_0)$ such that $P_C(x - \sum_{i\in I_0} \lambda_i h_i) = x_0$ for some $\lambda_i \geq 0$ and $h_i \in \partial g_i(x_0)$ with each $i \in I_0$. In view of the definition of $N'(x_0)$ and since $J(x - x_0) = x - x_0$ in a Hilbert space, $J(x - x_0) \in N_C(x_0) + N'(x_0)$ if and only if there exist a finite set $I_0 \subseteq I(x_0)$, $\lambda_i \geq 0$, and $h_i \in \partial g_i(x_0)$ such that

$$x - \sum_{i\in I_0} \lambda_i h_i - x_0 \in N_C(x_0). \tag{4.1}$$

By Proposition 2.2, (4.1) holds if and only if $P_C(x - \sum_{i\in I_0} \lambda_i h_i) = x_0$. Thus the result is clear. ∎

COROLLARY 4.1. *Consider the system* (CIS) *as before but suppose that, for each $i \in I$, g_i is an affine function defined by*

$$g_i(x) = \operatorname{Re}\langle h_i, x\rangle - b_i \quad \text{for all } x \in X, \tag{4.2}$$

where $\{h_i : \ i \in I\} \subset X \setminus \{0\}$ and $\{b_i\} \subseteq \mathbb{R}$. Let $C_i \subseteq X$ be defined by

$$C_i = \{x \in X : \ \operatorname{Re}\langle h_i, x\rangle \leq b_i\}. \tag{4.3}$$

Let C be a closed convex set in X and let $x_0 \in C \bigcap \left(\bigcap_{i\in I} C_i\right)$. Then the following statements are equivalent:

(i) *the family $\{C, C_i : \ i \in I\}$ has the strong CHIP at x_0;*

(ii) *for any $x \in X$, $P_K(x) = x_0$ if and only if there exists a finite (possibly empty) set $I_0 \subseteq I(x_0)$ such that $P_C(x - \sum_{i\in I_0} \lambda_i h_i) = x_0$ for some $\lambda_i \geq 0$ with each $i \in I_0$.*

More generally, let C be a closed convex set in X, $\{h_i : \ i \in I\} \subset X \setminus \{0\}$, and let $\{\Omega_i : \ i \in I\}$ be a family of nonempty closed convex subsets of the scalar field. Define

$$\hat{C}_i = \{x \in X : \ \langle h_i, x\rangle \in \Omega_i\}, \quad i \in I, \tag{4.4}$$

and

$$\hat{K} = C \bigcap \left(\bigcap_{i\in I} \hat{C}_i\right). \tag{4.5}$$

Let $x_0 \in \hat{K}$, and define

$$\hat{I}(x_0) := \{i \in I : \ \langle h_i, x_0\rangle \in \operatorname{bd} \Omega_i\}.$$

For convenience, we shall write $\tilde{h}_i(\cdot)$ for the function $\langle h_i, \cdot\rangle$ on X, and h_i^0 for the scalar $\langle h_i, x_0\rangle$. Then we have the following perturbation theorem.

THEOREM 4.2. *Let X be a Hilbert space (over $\mathbb{C}$ or $\mathbb{R}$), and let $x_0 \in \hat{K}$. Then the following statements are equivalent:*

(i) *the collection of convex sets $\{C, \hat{C}_i : \ i \in I\}$ has the strong CHIP at x_0;*

(ii) *for any* $x \in X$, $P_{\hat{K}}(x) = x_0$ *if and only if there exists a finite (possibly empty) set* $I_0 \subseteq \hat{I}(x_0)$ *such that* $P_C(x - \sum_{i\in I_0} \overline{\alpha}_i h_i) = x_0$ *for some* $\alpha_i \in N_{\Omega_i}(h_i^0)$ *with each* $i \in I_0$.

Proof. We may assume that X is over $\mathbb{C}$ (the case when X is over $\mathbb{R}$ is similar). For each $i \in I$, let $F_i(\cdot)$ be any (real-valued) convex function on $\mathbb{C}$ such that

$$\Omega_i = \{x \in \mathbb{C} : \ F_i(x) \le 0\} \tag{4.6}$$

(see (4.9) below, for example). Then we have that

$$\partial(F_i \circ \tilde{h}_i)(x_0) = \{\overline{\alpha} h_i : \ \alpha \in \partial F_i(h_i^0)\}. \tag{4.7}$$

In fact, it is easy to verify that the set on the right-hand side of (4.7) is contained in the set on the left-hand side. Conversely, let $x^* \in \partial(F_i \circ \tilde{h}_i)(x_0) : \operatorname{Re}\langle x^*, x - x_0\rangle \le (F_i \circ \tilde{h}_i)(x) - (F_i \circ \tilde{h}_i)(x_0)$ for all $x \in X$. Treating the corresponding real space X_R as in section 2, it follows that the real part $\operatorname{Re} x^* \in \tilde{\partial}(F_i \circ \tilde{h}_i)(x_0)$, where $\operatorname{Re} x^* : x \mapsto \operatorname{Re}\langle x^*, x\rangle$. Thus, by [23, Proposition 2.3], there exists $\alpha \in \partial F_i(h_i^0)$ such that

$$\operatorname{Re}\langle x^*, x\rangle = \operatorname{Re}\ \overline{\alpha}\langle h_i, x\rangle \quad \text{for all } x \in X. \tag{4.8}$$

This implies that $x^* = \overline{\alpha} h_i$; hence (4.7) is proved.

Define

$$\hat{g}_i(x) = d_{\Omega_i}(\langle h_i, x\rangle) \quad \text{for all } x \in X,\ i \in I, \tag{4.9}$$

where $d_{\Omega_i}(\cdot)$ denotes the distance function from the set Ω_i. Note that $\hat{g}_i^{-1}(\mathbb{R}_-) = \hat{C}_i$. Also, by (4.7) and [18, Example 3.3, p. 259], we get

$$\partial \hat{g}_i(x_0) = \{\overline{\alpha} h_i : \ \alpha \in N_{\Omega_i}(h_i^0),\ |\alpha| \le 1\}. \tag{4.10}$$

Consequently, by Theorem 4.1, (ii) holds if and only if the following system on X,

$$\hat{g}_i(x) \le 0, \quad i \in I, \tag{4.11}$$

satisfies the BCQ relative to C at x_0, that is,

$$N_{C\cap\left(\cap_{i\in I}\hat{g}_i^{-1}(\mathbb{R}_-)\right)}(x_0) = N_C(x_0) + \sum_{i\in\hat{I}(x_0)} \operatorname{cone}\left(\partial \hat{g}_i(x_0)\right) = N_C(x_0) + \sum_{i\in I} \operatorname{cone}\left(\partial \hat{g}_i(x_0)\right),$$

where the last equality holds because, for each $i \in I \setminus \hat{I}(x_0)$, one has $N_{\Omega_i}(h_i^0) = 0$ (and hence, by (4.10), that $\partial \hat{g}_i(x_0) = 0$). Note also that $N_{\hat{C}_i}(x_0) = 0$ for each $i \in I \setminus \hat{I}(x_0)$. Thus, to complete the proof, it suffices, by (4.10), to prove that

$$N_{\hat{C}_i}(x_0) = \{\overline{\alpha} h_i : \ \alpha \in N_{\Omega_i}(h_i^0)\} \quad \text{for all } i \in \hat{I}(x_0). \tag{4.12}$$

Let $i \in \hat{I}(x_0)$ and divide the case in two: $\operatorname{int}\Omega_i \ne \emptyset$ and $\operatorname{int}\Omega_i = \emptyset$. In the first case, take a convex function F_i on $\mathbb{C}$ such that $\Omega_i = \{z \in \mathbb{C} : \ F_i(z) \le 0\}$ and $\operatorname{int}\Omega_i = \{z \in \mathbb{C} : \ F_i(z) < 0\}$ (e.g., $F_i(\cdot) = \hat{q}_i(\cdot - \hat{z}_i) - 1$, where $\hat{q}_i$ denotes the Minkowski functional (cf. [30, p. 24]) of the set $\Omega_i - \hat{z}_i$ for some $\hat{z}_i \in \operatorname{int}\Omega_i$). Then, by Remark 1.1(a),

$$N_{\Omega_i}(h_i^0) = \operatorname{cone}\left(\partial F_i(h_i^0)\right). \tag{4.13}$$

Similarly, note that $\hat{C}_i = \{x \in X : \ (F_i \circ \tilde{h}_i)(x) \leq 0\}$ and that x_0 is not a minimizer of the convex function $F_i \circ \tilde{h}_i$ on X; again, by Remark 1.1, we have that

$$N_{\hat{C}_i}(x_0) = \text{cone}\big(\partial(F_i \circ \tilde{h}_i)(x_0)\big). \tag{4.14}$$

Hence, by (4.7), (4.13), and (4.14), (4.12) holds. It remains to consider the second case: Ω_i is of empty interior. Then the convex set Ω_i in $\mathbb{C}$ must be of one dimension and hence can be expressed as the intersection of at most four real half-spaces in $\mathbb{C}$ (e.g., a bounded closed line-segment in $\mathbb{R}^2$ is the intersection of four half-spaces). Thus there are affine functionals, say $\hat{F}_j$, $j = 1, \ldots, m$ with $m \leq 4$, such that $\Omega_i = \bigcap_{j=1}^m \hat{F}_j^{-1}(\mathbb{R}_-)$. Write $\hat{f}_j$ for the function $\hat{F}_j \circ \tilde{h}_i$ $(j = 1, \ldots, m)$ and denote $J_0 := \{j : \ \hat{f}_j(x_0) = 0, \ j = 1, \ldots, m\} = \{j : \ \hat{F}_j(h_i^0) = 0, \ j = 1, \ldots, m\}$. Then by Remark 1.1(a) we have that, for each $j \in J_0$,

$$N_{\hat{F}_j^{-1}(\mathbb{R}_-)}(h_i^0) = \text{cone}\big(\partial \hat{F}_j(h_i^0)\big) \tag{4.15}$$

and

$$N_{\hat{f}_j^{-1}(\mathbb{R}_-)}(x_0) = \text{cone}\big(\partial \hat{f}_j(x_0)\big). \tag{4.16}$$

In addition, it is clear that $\hat{C}_i = \bigcap_{j=1}^m \hat{f}_j^{-1}(\mathbb{R}_-)$. It follows from Remark 2.2(d) and (4.16) that

$$N_{\hat{C}_i}(x_0) = \sum_{j=1}^m N_{\hat{f}_j^{-1}(\mathbb{R}_-)}(x_0) = \sum_{j \in J_0} \text{cone}\big(\partial \hat{f}_j(x_0)\big). \tag{4.17}$$

Similarly, we also have that

$$N_{\Omega_i}(h_i^0) = \sum_{j \in J_0} \text{cone}\big(\partial \hat{F}_j(h_i^0)\big). \tag{4.18}$$

Thus, by (4.7), (4.17), and (4.18), we get

$$N_{\hat{C}_i}(x_0) = \sum_{j \in J_0} \{\overline{\alpha} h_i : \ \alpha \in \text{cone}\big(\partial \hat{F}_j(h_i^0)\big)\} = \{\overline{\alpha} h_i : \ \alpha \in N_{\Omega_i}(h_i^0)\}, \tag{4.19}$$

and so (4.12) holds. The proof is complete. □

Let g_i be defined by

$$g_i(x) = \langle h_i, x \rangle - b_i \quad \text{for all } x \in X, \tag{4.20}$$

where $\{h_i : \ i \in I\} \subset X \setminus \{0\}$ and $\{b_i\} \subseteq \mathbb{C}$, and let $\tilde{S} = \bigcap_{i \in I} S_i$, where

$$S_i = \{x \in X : \ \langle h_i, x \rangle = b_i\}, \quad i \in I. \tag{4.21}$$

Applying Theorem 4.2 to the case when $\Omega_i = \{b_i\}$ for each i, we have the following corollary.

COROLLARY 4.2. *Let X be a Hilbert space over $\mathbb{R}$ (resp., $\mathbb{C}$) and let $x_0 \in C \cap \tilde{S}$. Then the following statements are equivalent:*

(i) *$\{C, S_i : \ i \in I\}$ has the strong CHIP at x_0;*

(ii) *for each* $x \in X$, $P_{C\cap\tilde{S}}(x) = x_0$ *if and only if there exists a finite (possibly empty) set* $I_0 \subseteq I(x_0)$ *such that* $P_C(x - \sum_{i\in I_0} \lambda_i h_i) = x_0$ *for some* $\lambda_i \in \mathbb{R}$ *(resp.,* $\mathbb{C}$*) with each* $i \in I_0$.

Remark 4.1. In the case when I is finite, each of (i) and (ii) of Corollary 4.2 is equivalent to the condition (cf. [11, 13, 14]) that

(i*) $\{C, \cap_{i\in I} S_i\}$ has the strong CHIP at x_0.

This is no longer true if I is infinite, as shown by the following example.

Example 4.1. Let X be the (real or complex) Hilbert space l^2 consisting of all infinite (real or complex) sequences (x_i) satisfying $\sum_{i=1}^{\infty} |x_i|^2 < \infty$. Let C be the closed unit ball of X. Let $I = \{2, 3, \dots\}$, and define

$$g_i(x) = x_i \quad \text{for all } x = (x_1, x_2, \dots) \in X, \quad i \in I.$$

Then $\tilde{S} = \{(x_1, 0, \dots) : x_1 \in \mathbb{R}\}$. Let $x_0 = 0$. Since $\operatorname{int} C \cap \tilde{S} \neq \emptyset$, $\{C, \tilde{S}\}$ has the strong CHIP at x_0. However, since $N_{S_i}(x_0) = \{x = (x_1, x_2, \dots) \in X : x_j = 0, j \neq i\}$ for each $i \in I$, $\sum_{i\in I} N_{S_i}(x_0)$ is not closed, and hence $\{C, S_i : i \in I\}$ does not have the strong CHIP at x_0. By Corollary 4.2, (ii) of Corollary 4.2 does not hold. □

Remark 4.2. Note that $\operatorname{int} C \cap (\cap_{i\in I} S_i) \neq \emptyset$ in Example 4.1. Thus Proposition 2.3(2) of [14] is not longer true if the index set I is infinite. Moreover, it is easy to verify that C itself is the only extremal subset of C containing $C\cap\tilde{S}$. Consequently the extremal subset C_b of C introduced in [15, Definition 4.1] is equal to C. Therefore the perturbation results in [15, Theorem 4.5] cannot be extended directly to the infinite case.

Remark 4.3. Results in this section have been presented as local ones; namely, we characterize conditions that hold at a single point x_0 of the set $C \cap (\cap_{i\in I} S_i)$. It is simple but sometimes desirable to describe the global analogue of the local results. For example, corresponding to Corollary 4.2, we have the following.

COROLLARY 4.3. *Let* X *be a Hilbert space. We write* $\tilde{S}$ *for* $\cap_{i\in I} S_i$. *Then the following statements are equivalent:*

(i) $\{C, S_i : i \in I\}$ *has the strong CHIP at each point of the intersection* $C \cap \tilde{S}$;

(ii) *for each* $x \in X$, *there exist a finite (possibly empty) set* I_x *of* I *and scalars* λ_i *such that*

$$P_{C\cap\tilde{S}}(x) = P_C\left(x - \sum_{i\in I_x} \lambda_i h_i\right).$$

Remark 4.4. By considering the whole space X in place of the unit ball in Example 4.1, we have a family $\{S_i : i \in I\}$ of polyhedra (in fact, maximal subspaces) which does not have the strong CHIP. In Example 4.2, we exhibit an infinite collection of polyhedra that has the strong CHIP.

Example 4.2. Let X be the real Hilbert space l^2 and let $I = \{1, 2, \dots\}$. Define, for each $i \in I$,

$$C_i = \{x = (x_n) \in X : x_i \leq 1\}.$$

Let $C = \cap_{i\in I} C_i$. Then $\{C_i : i \in I\}$ has the strong CHIP at each point x of C. Indeed, since $x = (x_n) \in l^2$, there exists an $N \in \mathbb{N}$ such that $|x_n| \leq 1/2$ for all $n \geq N$. Let U denote the ball with center x and radius $1/2$. Then $U \subset \cap_{i\geq N} C_i$. This shows that $x \in \operatorname{int}(\cap_{i\geq N} C_i)$ and hence that $N_{\cap_{i\geq N} C_i}(x) = 0$. Since

$$N_C(x) = N_{\cap_{i\leq N} C_i}(x) + N_{\cap_{i\geq N} C_i}(x)$$

and $\{C_1, C_2, \ldots, C_N\}$ has the strong CHIP, we have

$$N_C(x) = \sum_{i=1}^{N} N_{C_i}(x) = \sum_{i=1}^{\infty} N_{C_i}(x).$$

5. Best constrained approximation in $C(Q)$. Let $C(Q)$ denote the Banach space of all complex-valued continuous functions on a compact metric space Q endowed with the uniform norm:

$$\|f\| = \max_{t \in Q} |f(t)| \quad \text{for all } f \in C(Q).$$

Let $\mathcal{P}$ be a finite-dimensional subspace of $C(Q)$, and let $\{\Omega_t : \ t \in Q\}$ be a family of nonempty closed convex sets in the complex plane $\mathbb{C}$. For brevity, we write $\{\Omega_t\}$ for $\{\Omega_t : \ t \in Q\}$. Set

$$\mathcal{P}_\Omega = \{p \in \mathcal{P} : \ p(t) \in \Omega_t \text{ for all } t \in Q\}. \tag{5.1}$$

The problem considered here is that of finding an element $p^* \in \mathcal{P}_\Omega$ for $f \in C(Q)$ such that

$$\|f - p^*\| = \inf_{p \in \mathcal{P}_\Omega} \|f - p\| \tag{5.2}$$

(such p^* is called a best restricted range approximation to f from $\mathcal{P}$ with respect to $\{\Omega_t\}$). This problem was first presented and formulated by Smirnov and Smirnov in [33, 34]; their approach followed the standard path for the corresponding issue in the real-valued continuous function space theory (see, for example, [5, 20] and the relevant references therein). In [34], while it was pointed out that this problem for the general class of restrictions was quite difficult, they took up the special case when each Ω_t is a disk in $\mathbb{C}$. Later, in [35, 36, 37], a more general case was considered in that the family $\{\Omega_t\}$ was assumed to have the following properties:

(i) there exists an element $p_0 \in \mathcal{P}$ satisfying $p_0(t) \in \operatorname{int} \Omega_t$ for each $t \in Q$ (such an element p_0 of $\mathcal{P}$ will be called an interior point with respect to $\mathcal{P}$ and $\{\Omega_t\}$);

(ii) Ω_t is a strictly convex set with "smooth" boundary for each $t \in Q$;

(iii) the set-valued map $t \mapsto \Omega_t$ is continuous with respect to the Hausdorff metric.

It was pointed out in [21] that (i) and (iii) imply that there exists a function F on the product space $\mathbb{C} \times Q$ with the following properties:

(C1) $F(\cdot, t)$ is convex on $\mathbb{C}$ for each $t \in Q$;

(C2) $\operatorname{bd} \Omega_t = \{z \in \mathbb{C} : \ F(z,t) = 0\}$ for all $t \in Q$;

(C3) $\operatorname{int} \Omega_t = \{z \in \mathbb{C} : \ F(z,t) < 0\}$ for all $t \in Q$;

(C4) F is continuous on $\mathbb{C} \times Q$.

This observation led the first author of the present paper to study, in [21], a more general setting in that a function with properties (C1)–(C4) is given, $\Omega_t := \{z \in \mathbb{C} : F(z,t) \le 0\}$, and an interior point (in the above sense) exists. Thus (ii) and (iii) need not be satisfied.

For the remainder of this section, let $\mathcal{P}$ be a finite-dimensional subspace of $C(Q)$, let Q be a compact metric space, and let $\{\Omega_t : \ t \in Q\}$ be a family of nonempty closed convex subsets of $\mathbb{C}$ satisfying the following:

(D1) the set-valued function $t \mapsto \Omega_t$ is lower semicontinuous on Q;

(D2) there exists $p_0 \in \mathcal{P}$ such that

$$0 \in \operatorname{int}\left(\bigcap_{t\in Q}\left(\Omega_t - p_0(t)\right)\right) \tag{5.3}$$

(such an element p_0 of $\mathcal{P}$ will be called a strong interior point with respect to $\mathcal{P}$ and $\{\Omega_t\}$).

The following remarks show in particular that the present setting is more general than that of [21] (and [33, 34, 35, 36, 37]).

Remark 5.1. (a) In the case when (C1)–(C4) are satisfied, the map $t \mapsto \Omega_t$ is both upper (in the sense of Kuratowski; see [24, p. 37]) and lower semicontinuous on Q. In fact, the upper semicontinuity is trivial while the lower semicontinuity holds because for any $t_0 \in Q$ and $x_0 \in \operatorname{int}\Omega_{t_0}$ there exists an open neighborhood $V(t_0)$ of t_0 such that $x_0 \in \operatorname{int}\Omega_t$ for all $t \in V(t_0)$.

(b) One can prove that properties (C1)–(C4) imply that $p_0 \in \mathcal{P}$ is a strong interior point if and only if $\sup_{t\in Q} F(p_0(t), t) < 0$. Hence, in this case, $p_0 \in \mathcal{P}$ is an interior point if and only if it is a strong interior point.

(c) For a family $\{\Omega_t\}$ satisfying (D1) and (D2), there exist many functions $F(\cdot,\cdot)$ on $\mathbb{C} \times Q$ with properties (C1)–(C3). One such function which is given below has additional properties that will be useful for us. Let $t \in Q$ and $p_0 \in \mathcal{P}$ be such that (5.3) holds. Define $\hat{F} : \mathbb{C} \times Q \to \mathbb{R}$ by

$$\hat{F}(z,t) = \hat{q}_t(z - p_0(t)) - 1, \tag{5.4}$$

where $\hat{q}_t$ denotes the Minkowski functional (cf. [30, p. 24]) of the closed convex set $\hat{\Omega}_t$ in $\mathbb{C}$ defined by

$$\hat{\Omega}_t = \Omega_t - p_0(t); \tag{5.5}$$

thus $\hat{q}_t(z) = \inf\{\lambda > 0 : z \in \lambda\hat{\Omega}_t\}$. It is easy to verify that $\hat{F}$ does have the properties (C1)–(C3) stated for F. On the other hand, there are many examples for which (C1)–(C3) are satisfied but without any associated function F with the properties (C1)–(C4).

Before giving the main theorem of this section, we need some preliminary results.

LEMMA 5.1. *For each $t \in Q$, let q_t be defined by*

$$q_t(z) := \hat{q}_t(z - p_0(t)) \quad \text{for all } z \in \mathbb{C}; \tag{5.6}$$

that is, $q_t(\cdot) = \hat{F}(\cdot, t) + 1$. Then

(i) *there exists a constant $\gamma > 0$ such that*

$$|q_t(z) - q_t(z')| \le \gamma|z - z'| \quad \text{for all } t \in Q,\ z, z' \in \mathbb{C}; \tag{5.7}$$

(ii) *for each $z \in \mathbb{C}$, the function $t \mapsto q_t(z)$ is upper semicontinuous.*

Proof. (i) By (D2), there exists a ball $B(0,\delta)$ in $\mathbb{C}$ with center 0 and radius $\delta > 0$ such that

$$B(0,\delta) \subseteq \hat{\Omega}_t \quad \text{for all } t \in Q. \tag{5.8}$$

By the definition of Minkowski functionals (cf. [30, p. 24]), it follows that

$$\hat{q}_t(z) \le \frac{1}{\delta}\|z\| \quad \text{for all } t \in Q,\ z \in \mathbb{C}. \tag{5.9}$$

Hence, by the subadditivity of $\hat{q}_t$ and (5.6), (5.7) holds with $\gamma := \frac{1}{\delta}$.

(ii) Let $z \in \mathbb{C}$ and $t_0 \in Q$. We have to show that $\limsup_{t\to t_0} \hat{q}_t(z) \le \hat{q}_{t_0}(z)$. Take a sequence $(t_n) \to t_0$ such that $\lim_{t_n\to t_0} \hat{q}_{t_n}(z) = l$ for some $l \in \mathbb{R}$. It suffices to show that $l \le \hat{q}_{t_0}(z)$. To this end, let $\varepsilon > 0$. Then, by the definition of Minkowski functionals, $z \in (\hat{q}_{t_0}(z) + \varepsilon)\,\hat{\Omega}_{t_0}$. Let $\lambda = \hat{q}_{t_0}(z) + \varepsilon$. Then $\frac{z}{\lambda} \in \hat{\Omega}_{t_0}$. By the lower semicontinuity, considering subsequences if necessary, we may assume that there exists $(z_n) \to \frac{z}{\lambda}$ with $z_n \in \hat{\Omega}_{t_n}$ for each n; we may assume further that $|z_n - \frac{z}{\lambda}| \le \frac{\varepsilon}{\lambda\gamma}$ for each n. Then it follows from (i) that

$$\hat{q}_{t_n}\left(\frac{z}{\lambda}\right) = \hat{q}_{t_n}\left(\frac{z}{\lambda}\right) - \hat{q}_{t_n}(z_n) + \hat{q}_{t_n}(z_n) \le \gamma\left|\frac{z}{\lambda} - z_n\right| + 1 \le 1 + \frac{\varepsilon}{\lambda}, \tag{5.10}$$

and so $\hat{q}_{t_n}(z) \le \varepsilon + \hat{q}_{t_0}(z) + \varepsilon$. Since $\varepsilon > 0$ is arbitrary, letting $\epsilon \to 0$, we have $l \le \hat{q}_{t_0}(z)$, as required. □

Let $\mathcal{P}$, p_0, and $\{\Omega_t : t \in Q\}$ be given with the properties (D1) and (D2). A key step to establishing our main result in this section is to apply Theorem 2.1 to (CIS) with $I = Q$, $X = C(Q)$, and g_t, where $g_t : C(Q) \to \mathbb{R}$ defined by

$$g_t(u) = q_t(u(t)) - 1 \quad \text{for all } u \in C(Q), \quad t \in Q. \tag{5.11}$$

Note, by (5.4) and (5.6), that

$$g_t(u) = \hat{q}_t(u(t) - p_0(t)) - 1 = \hat{F}(u(t), t) \quad \text{for all } t \in Q,\ u \in C(Q). \tag{5.12}$$

Thus, each g_t is a continuous convex function on $C(Q)$. Let $\hat{S}$ denote the solution set of the following system of inequalities:

$$g_t(\cdot) \le 0, \quad t \in Q. \tag{5.13}$$

Then $\hat{S}$ is nonempty, since p_0 is a Slater point for (5.13) as $g_t(p_0) = -1$ for each $t \in Q$. Note also that, by definition,

$$\hat{S} = \{u \in C(Q) : u(t) \in \Omega_t \quad \text{for all } t \in Q\}. \tag{5.14}$$

Let $\hat{G}$ denote the sup-function of $\{g_t : t \in Q\}$:

$$\hat{G}(u) = \sup_{t\in Q} g_t(u). \tag{5.15}$$

In a lemma below, we will show that $\hat{G}$ is continuous and that, for each $u \in C(Q)$, the function $t \mapsto g_t(u)$ is upper semicontinuous on Q. Granting these and applying Theorem 2.1, we immediately obtain the following proposition.

PROPOSITION 5.1. *Let $\mathcal{P}$ be a finite-dimensional subspace of $C(Q)$, $p_0 \in \mathcal{P}$, and let $\{\Omega_t : t \in Q\}$ be a family of closed convex subsets of $\mathbb{C}$ such that* (D1) *and* (D2) *are satisfied. Then the system* (5.13) *satisfies the BCQ relative to $\mathcal{P}$ at any point p of $\mathcal{P} \cap \hat{S}$.*

For each $t \in Q$, e_t denotes the point-valued functional on $C(Q)$ defined by

$$\langle e_t, u\rangle = u(t) \quad \text{for all } u \in C(Q). \tag{5.16}$$

LEMMA 5.2. *The function $\hat{G}$ and the set $\{g_t : t \in Q\}$ defined above have the following properties:*

(i) *for each $u \in C(Q)$, the function $t \mapsto g_t(u)$ is upper semicontinuous;*

(ii) *the sup-function $\hat{G}(\cdot) = \sup_{t\in Q} g_t(\cdot)$ is continuous;*

(iii) *for each* $u \in C(Q)$, $t \in Q$,

$$\partial g_t(u) = \{\overline{\alpha}\, e_t \in C(Q)^* : \ \alpha \in \partial q_t(u(t))\}. \tag{5.17}$$

Proof. (i) Let $u \in C(Q)$ and let $t_0 \in Q$. By (5.7),

$$g_t(u) = [q_t(u(t)) - q_t(u(t_0))] + q_t(u(t_0)) - 1 \le \gamma|u(t) - u(t_0)| + q_t(u(t_0)) - 1.$$

Then, by Lemma 5.1(ii), we have that

$$\limsup_{t\to t_0} g_t(u) \le \limsup_{t\to t_0}[q_t(u(t_0)) - 1] = q_{t_0}(u(t_0)) - 1 = g_{t_0}(u). \tag{5.18}$$

Thus (i) is proved.

(ii) This follows from the inequalities

$$|\hat{G}(u) - \hat{G}(v)| \le \sup_{t\in Q} |g_t(u) - g_t(v)| \le \sup_{t\in Q} |q_t(u(t)) - q_t(v(t))| \le \gamma\|u - v\| \tag{5.19}$$

for any $u, v \in C(Q)$, where the last inequality holds because of (5.7).

(iii) It is easy to check that $\partial g_t(u)$ contains the set on the right-hand side of (5.17). To show the reverse inclusion, let $u^* \in \partial g_t(u)$. Then $u^* \in C(Q)^*$ and there exists a complex Radon measure μ with bounded variation on Q such that

$$\langle u^*, v\rangle = \int_Q \overline{v}\, \mathrm{d}\mu \quad \text{ for all } v \in C(Q) \tag{5.20}$$

(cf. [39, p. 350]). Write $Q_t = Q \setminus \{t\}$ and $\mu = \mu_R + \mathrm{i}\mu_I$, where $\mu_R,\ \mu_I$ are real Radon measures on Q. Let $E_i \subseteq Q_t,\ i = 1, 2$, be such that $E_1 \cup E_2 = Q_t$, $E_1 \cap E_2 = \emptyset$, μ_R is nonnegative on E_1, and μ_R is nonpositive on E_2. Then $|\mu_R|(Q_t) = \mu_R(E_1) - \mu_R(E_2)$. For any $\varepsilon > 0$, let $F_i \subseteq E_i,\ i = 1, 2$, be closed and satisfy $|\mu_R|(E_i \setminus F_i) < \frac{\varepsilon}{4},\ i = 1, 2$. By Urysohn's lemma, there exists a real continuous function w on Q satisfying $\|w\| \le 1$ and

$$w(s) = \begin{cases} 1, & s \in F_1, \\ -1, & s \in F_2, \\ 0, & s = t. \end{cases}$$

Define $v = w + u$. Since $w = 0$ at t, $g_t(w + u) = g_t(u)$, and hence

$$0 = g_t(v) - g_t(u) \ge \operatorname{Re}\langle u^*, v - u\rangle = \operatorname{Re}\int_Q (\overline{v - u})\, \mathrm{d}\mu = \int_Q (v - u)\, \mathrm{d}\mu_R.$$

This implies that

$$\mu_R(F_1) - \mu_R(F_2) < \frac{\varepsilon}{2}. \tag{5.21}$$

Consequently,

$$|\mu_R|(Q_t) = \mu_R(E_1) - \mu_R(E_2) \le \mu_R(F_1) - \mu_R(F_2) + \frac{\varepsilon}{2} < \varepsilon. \tag{5.22}$$

Hence, $|\mu_R|(Q_t) = 0$. Similarly, we have $|\mu_I|(Q_t) = 0$. Therefore μ must be a point-measure and hence $u^* = \overline{\alpha}\, e_t$ with some $\alpha \in \mathbb{C}$. Since $u^* \in \partial g_t(u),\ \alpha \in \partial q_t(u(t))$ and (5.17) is proved. □

Let $F(\cdot,\cdot)$ be any fixed function on $\mathbb{C}\times Q$ satisfying (C1)–(C3). Let $f \in C(Q)$, $p^* \in \mathcal{P}_\Omega$. Following [21, 33, 34], define

$$M(f) = \{t \in Q : \ |f(t)| = \|f\|\}, \quad B(p^*) = \{t \in Q : \ p^*(t) \in \operatorname{bd}\Omega_t\}, \tag{5.23}$$

$$\sigma(t) = f(t) - p^*(t) \quad \text{for all } t \in Q, \tag{5.24}$$

and, for each $t \in B(p^*)$, let $-\tau(t)$ denote the subdifferential of the convex function $F(\cdot,t)$ at $p^*(t)$, that is,

$$\tau(t) = -\partial F(\cdot,t)|_{p^*(t)} \quad \text{for all } t \in B(p^*). \tag{5.25}$$

Thus $\sigma(t) \in \mathbb{C}$ and $\tau(t) \subseteq \mathbb{C}$. Note that

$$B(p^*) = \{t \in Q : \ g_t(p^*) = \hat{G}(p^*) = 0\}; \tag{5.26}$$

that is, $B(p^*)$ is exactly the active index set for p^* with respect to the system (5.13). Furthermore, assume that $\mathcal{P}$ has dimension n and is spanned by, say, $\phi_1, \phi_2, \ldots, \phi_n$. For each $t \in Q$, by abuse of notation, let $\mathbf{c}(t) \subseteq \mathbb{C}^n$ be defined by

$$\mathbf{c}(t) = (\overline{\phi_1(t)}, \ldots, \overline{\phi_n(t)})\tau(t);$$

more precisely,

$$\mathbf{c}(t) = \{(\eta\overline{\phi_1(t)}, \ldots, \eta\overline{\phi_n(t)}) : \ \eta \in \tau(t)\}.$$

Similarly, we define $\mathbf{d}(t) \in \mathbb{C}^n$ by

$$\mathbf{d}(t) = (\overline{\phi_1(t)}, \ldots, \overline{\phi_n(t)})\sigma(t).$$

Define

$$\mathcal{U} = \mathcal{U}_1 \bigcup \mathcal{U}_2,$$

where

$$\mathcal{U}_1 = \{\mathbf{d}(t) : \ t \in M(f - p^*)\} \quad \text{and} \quad \mathcal{U}_2 = \bigcup_{t \in B(p^*)} \mathbf{c}(t).$$

Note that, by continuity and compactness, $\mathcal{U}_1$ is compact. Furthermore, we have the following lemma. We assume that F in (5.25) is the function F given in (5.4).

LEMMA 5.3. *$\mathcal{U}$ is compact in $\mathbb{C}^n$.*

Proof. Note first that $t \in B(p^*)$ if and only if $q_t(p^*(t)) = 1$, where q_t is given by (5.6). Let $\{t_k\} \subseteq B(p^*)$ be a convergent sequence with limit t_0. By Lemma 5.2(i), we have that

$$q_{t_0}(p^*(t_0)) \geq \limsup_k q_{t_k}(p^*(t_k)) = 1. \tag{5.27}$$

Since $p^*(t_0) \in \Omega_{t_0}$, it follows that $q_{t_0}(p^*(t_0)) = 1$. Hence $t_0 \in B(p^*)$ and $B(p^*)$ is closed. By assumption,

$$F(z,t) = q_t(z) - 1 \quad \text{for all } z \in \mathbb{C},\ t \in Q. \tag{5.28}$$

Then, by Lemma 5.1(i), one can show (as in [21, Theorem 3.1]) that $\mathcal{U}_2$ is compact and so is $\mathcal{U}$. □

Now we are ready to give the main theorem of this section, which gives characterizations of the best restricted range approximation in $C(Q)$. The properties stated in (ii)–(iv) are standard and well known in approximation theory; see, e.g., [4, 5, 20]. Note that, by Remark 1.1, for any function $F(\cdot,\cdot)$ on $\mathbb{C}\times Q$ satisfying (C1)–(C3), we have that

$$\operatorname{cone}\partial F(\cdot,t)|_{p^*(t)} = N_{\Omega_t}(p^*(t)) \quad \text{for all } t\in B(p^*). \tag{5.29}$$

THEOREM 5.1. *Let $f\in C(Q)$, $p^*\in\mathcal{P}_\Omega$. Then the following four statements are equivalent:*

(i) *p^* is a best restricted range approximation to f from $\mathcal{P}$ with respect to $\{\Omega_t\}$;*

(ii)

$$\max\left\{\max_{t\in M(f-p^*)}\operatorname{Re}\big(p(t)\overline{\sigma(t)}\big),\ \max_{t\in B(p^*)}\max_{\tau\in\tau(t)}\operatorname{Re}\big(p(t)\overline{\tau}\big)\right\}\ge 0 \quad \textit{for all } p\in\mathcal{P}; \tag{5.30}$$

(iii) *the origin of the space $\mathbb{C}^n$ belongs to the convex hull of the set $\mathcal{U}$;*

(iv) *there exist sets $\{t_1,\dots,t_k\}\subseteq M(f-p^*)$, $\{t'_1,\dots,t'_m\}\subseteq B(p^*)$, $\tau'_j\in\tau(t'_j)$, $i=1,\dots,m$ $(1+m\le k+m\le 2n+1)$, and positive constants $\lambda_1,\dots,\lambda_k$, $\lambda'_1,\dots,\lambda'_m$ such that the following condition holds:*

$$\sum_{i=1}^{k}\lambda_i p(t_i)\overline{\sigma(t_i)}+\sum_{j=1}^{m}\lambda'_j p(t'_j)\overline{\tau'}_j=0 \quad \textit{for all } p\in\mathcal{P}. \tag{5.31}$$

Proof. Since the result is trivial in the case when $f\in P_\Omega$, we assume that $f\ne p^*$. By (5.29), we may assume, without loss of generality, that F in (5.25) is simply the function given by (5.28). Let $t\in B(p^*)$ and $\eta\in\tau(t)$. Then $q_t(p^*(t))=1$; hence

$$-\eta\in\partial q_t(p^*(t)) \text{ and } -\operatorname{Re}\big(p_0(t)-p^*(t)\big)\overline{\eta}\le q_t(p_0(t))-q_t(p^*(t))=-1. \tag{5.32}$$

Therefore, the case when $k=0$ will not occur in (5.31) because otherwise (5.31) would entail that

$$\sum_{j=1}^{m}\lambda'_j(p_0-p^*)(t'_j)\overline{\tau'}_j=0 \tag{5.33}$$

(with p replaced by p_0-p^* as $\mathcal{P}$ is a vector subspace) and (5.33) contradicts (5.32) (applied to $t=t'_j$ and $\eta=\tau'_j$) as each $\lambda'_j>0$. Thus (iii)$\Longleftrightarrow$(iv) by Carathéodory's theorem (cf. [4] and [30, p. 73]). Also, since $\mathcal{P}$ is spanned by $\phi_1,\dots,\phi_n$, it is easy to verify that (ii) does not hold if and only if there exists $z=(\gamma_1,\dots,\gamma_n)\in\mathbb{C}^n$ such that $\operatorname{Re}\langle u,z\rangle<0$ for all $u\in\mathcal{U}$. Thus, as $\mathcal{U}$ is compact by Lemma 5.3, (ii)$\Longleftrightarrow$(iii) by the linear inequality theorem (see [4]). To show that (i)$\Longleftrightarrow$(iv), note that $\mathcal{P}_\Omega=\mathcal{P}\cap\hat{S}$,

where $\hat{S}$ denotes the solution set of the convex inequality system in $C(Q)$ defined by (5.13). By Proposition 5.1, this system satisfies the BCQ relative to $\mathcal{P}$ at p^*. By the implication (i)$\Longrightarrow$(iv) in Theorem 3.1 and the fact that $\mathcal{P}$ is a vector subspace containing p^* (so $N_{\mathcal{P}}(p^*)|_{\mathcal{P}} = 0$), the following statements are equivalent:

(i*) p^* is a best approximation to f from $\mathcal{P} \cap \hat{S}$;

(iv*) $J(f - p^*)|_{\mathcal{P}} \bigcap N'(p^*)|_{\mathcal{P}} \neq \emptyset$.

Since (i) and (i*) are the same, it remains to show (iv)$\Longleftrightarrow$(iv*).

(iv)$\Longrightarrow$(iv*). Suppose that (iv) holds. Without loss of generality, assume that $\sum_{i=1}^{k} \lambda_i = 1$ in (5.31). Then $\sum_{i=1}^{k} \lambda_i \overline{\sigma(t_i)} e_{t_i} \in J(f - p^*)$. By (5.17) and (5.32), each $-\overline{\tau'}_j e_{t'_j} \in \partial g_{t'_j}(p^*)$ and so, by (5.26), $\sum_{j=1}^{m} \lambda'_j\big(-\overline{\tau'}_j e_{t'_j}\big) \in N'(p^*)$. Therefore (5.31) implies (iv*).

(iv*)$\Longrightarrow$(iv). Suppose that (iv*) holds. Then there exist $v^* \in J(f - p^*)$, $w_j^* \in \partial g_{t'_j}(p^*)$, and $\lambda'_j > 0$, $j = 1, 2, \ldots, m$, with each $t'_j \in B(p^*)$ such that

$$\langle v^*, p\rangle = \sum_{j=1}^{m} \lambda'_j \langle w_j^*, p\rangle \quad \text{for all } p \in \mathcal{P}. \tag{5.34}$$

Set $u^* = v^*/\|v^*\|$. Applying [31, Lemma 1.3, p. 169] to the real linear span of $\mathcal{P} \cup \{f\}$, there exist a positive integer l (with $1 \leq l \leq 2n + 2$), l extreme points $u_1^*, \ldots, u_l^*$ of the unit ball Σ^* of $C(Q)^*$, and positive constants β_i, $i = 1, 2, \ldots, l$, with $\sum_{i=1}^{l} \beta_i = 1$ such that

$$\langle u^*, p\rangle = \sum_{i=1}^{l} \beta_i \langle u_i^*, p\rangle \quad \text{for all } p \in \mathcal{P} \cup \{f\}. \tag{5.35}$$

By a well-known representation of the extreme points of Σ^* (cf. [31, p. 69]), there exist some $\alpha_i \in \mathbb{C}$ with $|\alpha_i| = 1$ and $t_i \in Q$ such that

$$u_i^* = \alpha_i e_{t_i}, \quad i = 1, 2, \ldots, l. \tag{5.36}$$

By the definition of u^*, $\|u^*\| = 1$ and $\langle u^*, f - p^*\rangle = \|f - p^*\|$; hence, by (5.35), $t_i \in M(f - p^*)$ and $\alpha_i = \overline{(f - p^*)(t_i)}/\|f - p^*\|$. Furthermore, by (5.17), for each j, there exists $\alpha'_j \in \partial q_{t'_j}(p^*(t'_j))$ such that $w_j^* = \overline{\alpha'}_j e_{t'_j}$. Therefore, (5.34) becomes

$$\sum_{i=1}^{l} \beta'_i \overline{\sigma(t_i)} \langle e_{t_i}, p\rangle = \sum_{j=1}^{m} \lambda'_j \overline{\alpha}'_j \langle e_{t'_j}, p\rangle \quad \text{for all } p \in \mathcal{P}, \tag{5.37}$$

where $\beta'_i = \|v^*\|\beta_i/\|f - p^*\|$ for each $i = 1, \ldots, l$. This implies that (iii) holds and so (iv) holds by (iii)$\Longleftrightarrow$(iv). The proof is complete. □

REFERENCES

[1] A. Baken, F. Deutsch, and W. Li, *Strong CHIP, Normality, and Linear Regularity of Convex Sets*, available online from http://www.math.odu.edu/~wuli, May, 2002.

[2] H. Bauschke, J. Borwein, and W. Li, *Strong conical hull intersection property, bounded linear regularity, Jameson's property (G), and error bounds in convex optimization*, Math. Program. Ser. A, 86 (1999), pp. 135–160.

[3] D. Braess, *Nonlinear Approximation Theory*, Springer-Verlag, New York, 1986.

[4] E. W. Cheney, *Introduction to Approximation Theory*, McGraw-Hill, New York, 1966.
[5] B. L. Chalmers and G. D. Taylor, *Uniform approximation with constraints*, Jahresber. Deutsch. Math.-Verein., 81 (1978/1979), pp. 49–86.
[6] C. Chui, F. Deutsch, and J. Ward, *Constrained best approximation in Hilbert space*, Constr. Approx., 6 (1990), pp. 35–64.
[7] C. Chui, F. Deutsch, and J. Ward, *Constrained best approximation in Hilbert space* II, J. Approx. Theory, 71 (1992), pp. 231–238.
[8] F. Clarke, *Optimization and Nonsmooth Analysis*, John Wiley, New York, 1983.
[9] F. Deutsch, *The role of the strong conical hull intersection property in convex optimization and approximation*, in Approximation Theory IX, Vol. I: Theoretical Aspects, C. Chui and L. Schumaker, eds., Vanderbilt University Press, Nashville, TN, 1998, pp. 105–112.
[10] F. Deutsch, *Some Applications of Functional Analysis to Approximation Theory*, Doctoral dissertation, Brown University, Providence, RI, 1965.
[11] F. Deutsch, *Best Approximation in Inner Product Spaces*, Springer-Verlag, New York, 2001.
[12] F. Deutsch, W. Li, and J. Swetits, *Fenchel duality and the strong conical intersection property*, J. Optim. Theory Appl., 102 (1997), pp. 681–695.
[13] F. Deutsch, W. Li, and J. Ward, *A dual approach to constrained interpolation from a convex subset of Hilbert space*, J. Approx. Theory, 90 (1997), pp. 385–444.
[14] F. Deutsch, W. Li, and J. D. Ward, *Best approximation from the intersection of a closed convex set and a polyhedron in Hilbert space, weak Slater conditions, and the strong conical hull intersection property*, SIAM J. Optim., 10 (1999), pp. 252–268.
[15] F. Deutsch, V. Ubhaya, J. Ward, and Y. Xu, *Constrained best approximation in Hilbert space* III: *Application to n-convex functions*, Constr. Approx., 12 (1996), pp. 361–384.
[16] J. Diestel, *Geometry of Banach Spaces—Selected Topics*, Lecture Notes in Math. 485, Springer-Verlag, New York, 1975.
[17] M. A. Goberna and M. A. Lopez, *Linear Semi-infinite Optimization*, Wiley, Chichester, UK, 1998.
[18] J. Hiriart-Urruty and C. Lemarechal, *Convex Analysis and Minimization Algorithms* I, Grundlehren Math. Wiss. 305, Springer-Verlag, New York, 1993.
[19] R. C. James, *Characterization of reflexivity*, Studia Math., 23 (1964), pp. 205–216.
[20] A. Kroo and D. Schmidt, *A Haar-type theory of best uniform approximation with constraints*, Acta Math. Hungar., 58 (1991), pp. 351–374.
[21] C. Li, *On best uniform restricted range approximation in complex-valued continuous function spaces*, J. Approx. Theory, 120 (2003), pp. 71–84.
[22] C. Li and X.-Q. Jin, *Nonlinearly constrained best approximation in Hilbert spaces: The strong CHIP and the basic constraint qualification*, SIAM J. Optim., 13 (2002), pp. 228–239.
[23] C. Li and K. F. Ng, *On best approximation by nonconvex sets and perturbation of nonconvex inequality systems in Hilbert spaces*, SIAM J. Optim., 13 (2002), pp. 726–744.
[24] W. Li, C. Nahak, and I. Singer, *Constraint qualification for semi-infinite systems of convex inequalities*, SIAM J. Optim., 11 (2000), pp. 31–52.
[25] O. Mangasarian, *Nonlinear Programming*, McGraw-Hill, New York, 1969.
[26] C. Micchelli, P. Smith, J. Swetits, and J. Ward, *Constrained L_p-approximation*, Constr. Approx., 1 (1985), pp. 93–102.
[27] C. A. Micchelli and F. I. Utreras, *Smoothing and interpolation in a convex subset of a Hilbert space*, SIAM J. Sci. Statist. Comput., 9 (1988), pp. 728–746.
[28] R. R. Phelps, *Convex Functions, Monotone Operators and Differentiability*, Lecture Notes in Math. 1364, Springer-Verlag, New York, 1989.
[29] G. Sh. Rubenstein, *On an extremal problem in a normed linear space*, Sibirskii Mat. Zh., 6 (1965), pp. 711–714 (in Russian).
[30] W. Rudin, *Functional Analysis*, McGraw-Hill, New York, 1981.
[31] I. Singer, *Best Approximation in Normed Linear Spaces by Elements of Linear Subspaces*, Springer-Verlag, Berlin, Heidelberg, New York, 1970.
[32] I. Singer, *Duality for optimization and best approximation over finite intersection*, Numer. Funct. Anal. Optim., 19 (1998), pp. 903–915.
[33] G. S. Smirnov and R. G. Smirnov, *Best uniform restricted range approximation of complex-valued functions*, C. R. Math. Rep. Acad. Sci. Canada, 19 (2) (1997), pp. 58–63.
[34] G. S. Smirnov and R. G. Smirnov, *Best uniform approximation of complex-valued functions by generalized polynomials having restricted range*, J. Approx. Theory, 100 (1999), pp. 284–303.
[35] G. S. Smirnov and R. G. Smirnov, *Kolmogorov-type theory of best restricted approximation*, East J. Approx., 6 (3) (2000), pp. 309–329.
[36] G. S. Smirnov and R. G. Smirnov, *Best uniform restricted ranges approximation.* II, C. R. Acad. Sci. Paris Sér. I Math., 330 (2000), pp. 1059–1064.

[37] G. S. Smirnov and R. G. Smirnov, *Theory of best restricted ranges approximation revisited: A characterization theorem*, in Approximation Theory and Its Applications, Pr. Inst. Mat. Nats. Akad. Nauk Ukr. Mat. Zastos 31, Natsīonal Akad. Nauk Ukraïni, Īnst. Mat., Kiev, 2000, pp. 436–445.
[38] Y. Yuan and W. Sun, *Optimization Theory and Methods*, Science Press, Beijing, 1997 (in Chinese).
[39] A. C. Zaanen, *Integration*, North-Holland, Amsterdam, 1967.

SIAM J. OPTIM.
Vol. 14, No. 2, pp. 608–619

EXTENDED ACTIVE CONSTRAINTS IN LINEAR OPTIMIZATION WITH APPLICATIONS*

M. A. GOBERNA†, M. A. LÓPEZ†, AND M. I. TODOROV‡

Abstract. In this paper we introduce different relaxations of the concept of active constraint at a given point with respect to a certain linear inequality system with an arbitrary number of constraints. We show that these concepts provide useful local information in linear optimization, for instance conditions for a given feasible solution to be a unique, extreme point, optimal solution or a strongly unique optimal solution.

Key words. linear programming, linear semi-infinite programming, active constraints

AMS subject classifications. 90C05, 90C34, 15A39

DOI. S1052623402401579

1. Introduction.

Given an optimization problem

$$\text{(P)} \quad \operatorname{Inf} c'x \text{ such that } x \in F,$$

where $c \in \mathbb{R}^n$ and F is a nonempty closed convex set in $\mathbb{R}^n$, and given $\overline{x} \in F$, the positive polar of the cone of feasible directions at $\overline{x}$, $D(F,\overline{x})^0$ allows us to check (see Proposition 3.1) whether

(Q1) $\overline{x}$ is the unique feasible solution (i.e., $F = \{\overline{x}\}$),

(Q2) $\overline{x}$ is an optimal solution of (P), and

(Q3) $\overline{x}$ is a strongly unique solution of (P) (i.e., there exists $k > 0$ such that $c'x \geq c'\overline{x} + k\|x - \overline{x}\|$ for all $x \in F$).

Moreover, $D(F,\overline{x})^0$ also provides a sufficient condition for

(Q4) $\overline{x} \in \operatorname{extr} F$ (the set of extreme points of F).

In linear optimization, F is represented through a certain linear inequality system $\sigma = \{a_t'x \geq b_t, t \in T\}$, where T is an arbitrary (possibly infinite) index set, $a_\cdot : T \to \mathbb{R}^n$, and $b_\cdot : T \to \mathbb{R}$. Then (P) is a linear programming (LP) problem if $|T| < \infty$ and a linear semi-infinite programming (LSIP) problem otherwise.

Although there exist known formulae that express $D(F,\overline{x})^0$ in terms of certain convex cones associated with σ, these expressions are frequently too complex in practical situations in order to obtain a useful description of $D(F,\overline{x})^0$ allowing us to determine whether $\overline{x}$ satisfies properties (Q1)–(Q4) or not (see the discussion in section 3). There are two alternative issues in order to check these properties.

The first approach consists of determining the largest class of linear systems for which $D(F,\overline{x})^0 = A(\overline{x})$ for all $\overline{x} \in F$, where $A(\overline{x})$ is the convex conical hull of $\{a_t, t \in T(\overline{x})\}$, with $T(\overline{x}) := \{t \in T \mid a_t'\overline{x} = b_t\}$. The last two sets are called the set of active constraints and the set of active indices at $\overline{x}$, respectively, whereas $A(\overline{x})$ is the active cone at $\overline{x}$. The computation of $A(\overline{x})$ requires the calculation of all the

*Received by the editors January 28, 2002; accepted for publication (in revised form) May 22, 2003; published electronically November 6, 2003. This work was partially supported by MCYT of Spain, CONACyT of Mexico and FEDER of EU, grant BMF2002-04114-C02-01.
http://www.siam.org/journals/siopt/14-2/40157.html

†Department of Statistics and Operations Research, Faculty of Sciences, University of Alicante, 03071 Alicante, Spain (mgoberna@ua.es, marco.antonio@ua.es).

‡UDLA, Cholula, Puebla, Mexico. On leave from IMI-BAS, Sofia, Bulgaria (mtodorov@mail.udlap.mx).

zeros of the slack function at $\overline{x}$, $a_t'\overline{x}-b_t$, followed by the formation of the nonnegative linear combinations of the set of active constraints, but this is usually easier than the calculation of $D(F,\overline{x})^0$ by means of the mentioned formulae.

The second approach consists of replacing the set of active constraints by another related set being able to give a sensible answer to questions (Q1)–(Q4). This is the main purpose of this paper.

The paper is organized as follows. In section 3 we review the role played by $D(F,\overline{x})^0$ and $A(\overline{x})$ in linear optimization. Section 4 introduces a family of extended constraints sets depending on a positive parameter. Finally, section 5 analyzes the existing relationships between three different sets of extended constraints and provides conditions for the convex conical hulls of these sets to coincide with $D(F,\overline{x})^0$.

2. Preliminaries. Given a nonempty set X of a certain Euclidean space, by $\operatorname{aff} X$, $\operatorname{span} X$, $\operatorname{conv} X$, $\operatorname{cone} X$, and $\dim X$ we denote the affine hull, the linear hull, the convex hull, the convex conical hull, and the dimension of $\operatorname{aff} X$, respectively. Moreover, we define $\operatorname{cone}\emptyset=\{0_n\}$. We denote by X^0 the positive polar of a given convex cone X and by $X^\perp$ the orthogonal subspace to a given linear subspace X. From the topological side, $\operatorname{int} X$, $\operatorname{cl} X$, and $\operatorname{bd} X$ represent the interior, the closure, and the boundary of X, respectively. The null-vector, the open unit ball, and the canonical basis in $\mathbb{R}^n$ will be denoted by 0_n, B_n, and $\{e_1,\ldots,e_n\}$, respectively. Finally, $\lim_r$ should be interpreted as $\lim_{r\to\infty}$.

Throughout the paper we shall consider a given linear optimization problem (P) with feasible set $F\subset\mathbb{R}^n$ described by means of $\sigma=\{a_t'x\geq b_t, t\in T\}$. The optimal set of (P) (possibly empty even though the problem has a finite value) will be denoted by F^*. If different systems arise in the same context, they will be distinguished by subscripts, and the same subscripts will distinguish the associated elements.

$\overline{x}\in\mathbb{R}^n$ is a strong Slater (SS) point for σ if there exists $\varepsilon>0$ such that $a_t'\overline{x}\geq b_t+\varepsilon$ for all $t\in T$. In such a case $a_t'\overline{x}>b_t$ for all $t\in T$; i.e., $\overline{x}$ is a Slater point for σ. If T is a compact Hausdorff space and the functions $a_\cdot$ and $b_\cdot$ are continuous, then σ is said to be continuous. In particular, any ordinary inequality system (i.e., such that $|T|<\infty$) is continuous if we consider T equipped with the discrete topology. Any Slater point is an SS point for a continuous system.

If there exists an SS point for σ, we say that σ is SS. Analogously, if there exists a positive lower (upper) bound for $\{\|a_t\|, t\in T\}$, then we say that σ is LB (UB, respectively). These three properties are related as follows.

PROPOSITION 2.1. *The following statements are true:*

(i) *If σ is SS and UB, then* $\dim F=n$.

(ii) *If* $\dim F=n$ *and σ is LB, then σ is SS.*

(iii) *If σ is LB and UB, then σ is SS if and only if* $\dim F=n$.

Proof. (i) Let $\overline{x}\in\mathbb{R}^n$, $\varepsilon>0$, and $\mu>0$ such that $a_t'\overline{x}\geq b_t+\varepsilon$ and $\|a_t\|\leq\mu$ for all $t\in T$. Given $x\in\overline{x}+\varepsilon\mu^{-1}B_n$, we can write $x=\overline{x}+\varepsilon\mu^{-1}u$, with $\|u\|<1$. Then we have

$$a_t'x=a_t'\overline{x}+\varepsilon\mu^{-1}(a_t'u)\geq b_t+\varepsilon-\varepsilon\mu^{-1}|a_t'u|\geq b_t$$

for all $t\in T$ so that $x\in F$. Hence $\overline{x}\in\operatorname{int}F$.

(ii) Let $\overline{x}\in\mathbb{R}^n$, $\varepsilon>0$, and $\eta>0$ such that $\overline{x}+\varepsilon\operatorname{cl}B_n\subset F$ and $\|a_t\|>\eta$ for all $t\in T$. Given $t\in T$, we have $a_t'(\overline{x}-\varepsilon\|a_t\|^{-1}a_t)\geq b_t$ so that $a_t'\overline{x}\geq b_t+\varepsilon\|a_t\|>b_t+\varepsilon\eta$. Hence $\overline{x}$ is an SS point for σ.

(iii) It is a straightforward consequence of (i) and (ii). ∎

We associate with σ the so-called second moment cone, $N = \operatorname{cone}\{\binom{a_t}{b_t}, t \in T\}$, and the characteristic cone, $K = N + \operatorname{cone}\{\binom{0_n}{-1}\}$. The well-known nonhomogeneous Farkas lemma establishes that, given $a \in \mathbb{R}^n$ and $b \in \mathbb{R}$, $a'x \geq b$ for all $x \in F$ if and only if $\binom{a}{b} \in \operatorname{cl} K$. Thus, $\operatorname{cl} K$ is the same for all the linear representations of F.

3. Active constraints: A review. The following result collects those tests for questions (Q1)–(Q4) which are based upon $D(F,\overline{x})^0 := \{z \in \mathbb{R}^n \mid y'z \geq 0 \text{ for all } y \in D(F,\overline{x})\}$, where $y \in D(F,\overline{x})$ if and only if there exists $\varepsilon > 0$ such that $\overline{x} + \varepsilon y \in F$.

PROPOSITION 3.1. *Given $\overline{x} \in F$, the following statements hold:*

(i) $F = \{\overline{x}\}$ *if and only if* $0_n \in \operatorname{int} D(F,\overline{x})^0$.

(ii) $\overline{x} \in F^*$ *if and only if* $c \in D(F,\overline{x})^0$.

(iii) $\overline{x}$ *is a strongly unique solution of* (P) *if and only if* $c \in \operatorname{int} D(F,\overline{x})^0$.

(iv) *If* $\dim D(F,\overline{x})^0 = n$, *then* $\overline{x} \in \operatorname{extr} F$.

Proof. (i) $F = \{\overline{x}\}$ if and only if $D(F,\overline{x}) = \{0_n\}$, i.e., if and only if $D(F,\overline{x})^0 = \mathbb{R}^n$, which is also equivalent to $0_n \in \operatorname{int} D(F,\overline{x})^0$.

(ii) $\overline{x} \in F^*$ if and only if $c'(x - \overline{x}) \geq 0$ for all $x \in F$; i.e., $c \in D(F,\overline{x})^0$.

(iii) It is a part of Theorem 3.1 in [3] (extending the LP version due to Mangasarian [7]).

(iv) According to statement (iii), if $z \in \operatorname{int} D(F,\overline{x})^0$, then $\overline{x}$ turns out to be the unique optimal solution of $\operatorname{Inf} z'x$ such that $x \in F$ so that $\{\overline{x}\}$ is an exposed face of F and $\overline{x} \in \operatorname{extr} F$. ∎

In order to apply Proposition 3.1 in practical situations it is necessary to express $D(F,\overline{x})^0$ in terms of the coefficients of σ. Actually, given $\overline{x} \in F$, we have

$$D(F,\overline{x})^0 = \left\{a \in \mathbb{R}^n \mid \begin{pmatrix} a \\ a'\overline{x} \end{pmatrix} \in \operatorname{cl} N\right\} = \left\{a \in \mathbb{R}^n \mid \begin{pmatrix} a \\ a'\overline{x} \end{pmatrix} \in \operatorname{cl} K\right\} \tag{3.1}$$

(see Theorem 5.2 and Exercise 5.3 in [2]), but the complex expressions of $\operatorname{cl} N$ and $\operatorname{cl} K$ (which involve the calculation of limits of sequences of nonnegative linear combinations of certain subsets of $\mathbb{R}^{n+1}$) make (3.1) seldom useful in practice.

In general, given $\overline{x} \in F$, $D(F,\overline{x}) \subset A(\overline{x})^0$ so that $\operatorname{cl} A(\overline{x}) \subset D(F,\overline{x})^0$, but the reverse inclusion can fail. Fortunately, there exists a large class of linear semi-infinite systems for which $D(F,\overline{x})^0 = A(\overline{x})$ for all $\overline{x} \in F$. These systems are called locally Farkas–Minkowski (LFM). In [8] it has been proved that σ is LFM if and only if every linear inequality $a'x \geq b$ which is a consequence of such σ (i.e., $a'x \geq b$ for all $x \in F$), and such that $a'x = b$ is a supporting hyperplane to F, is also the consequence of a finite subsystem of σ. This property holds, in particular, if either σ is continuous and has a Slater point or $D(F,\overline{x}) = A(\overline{x})^0$ for all $\overline{x} \in F$ (the last property always holds if $|T| < \infty$). In the last case, σ is said to be locally polyhedral (LOP). This class of systems captures the most relevant properties of the ordinary linear systems (see [1]).

COROLLARY 3.2. *If σ is LFM and $\overline{x} \in F$, then the following statements hold:*

(i) $F = \{\overline{x}\}$ *if and only if* $0_n \in \operatorname{int} A(\overline{x})$.

(ii) $\overline{x}$ *is an optimal solution of* (P) *if and only if* $c \in A(\overline{x})$.

(iii) $\overline{x}$ *is a strongly unique solution of* (P) *if and only if* $c \in \operatorname{int} A(\overline{x})$.

(iv) *If* $\dim A(\overline{x}) = n$, *then* $\overline{x} \in \operatorname{extr} F$. *The converse statement holds if σ is LOP.*

Proof. (i), (ii), and (iii) and the direct statement in (iv) are reformulations of the corresponding statements in Proposition 3.1. The converse statement in (iv) is Theorem 4.3 in [1]. ∎

Remark 1. The optimality condition (ii) in Corollary 3.2, $c \in A(\overline{x})$, is known as the KKT condition in linear optimization. This is always a sufficient condition for $\overline{x} \in F$ to be an optimal solution of (P). If σ is not LFM, there exists $\overline{x} \in F$ and $z \in D(F, \overline{x})^0 \setminus A(\overline{x})$ so that $\overline{x}$ is an optimal solution for $\operatorname{Inf} z'x$ such that $x \in F$, whereas the KKT condition fails at $\overline{x}$ (by Proposition 3.1, part (ii)). This means that σ is LFM if and only if the KKT condition characterizes the optimality of the feasible solutions for every possible linear optimization problem with constraint system σ. Thus the LFM property can be seen as the weakest global constraint qualification in linear optimization.

Remark 2. The uniqueness condition (i) in Corollary 3.2 can be reformulated in terms of $0_n \in \operatorname{int} \operatorname{conv} \{a_t, t \in T(\overline{x})\}$. In fact, assume that $0_n \in \operatorname{int} A(\overline{x})$. If $\varepsilon \operatorname{cl} B_n \subset A(\overline{x})$ for a certain $\varepsilon > 0$, and $u \in \mathbb{R}^n$ satisfies $\|u\| = 1$, then there exists $\alpha > 0$ such that $\alpha\varepsilon u \in \operatorname{conv} \{a_t, t \in T(\overline{x})\}$. Taking, in particular, as u the elements of the canonical basis of $\mathbb{R}^n$ and their opposite vectors, there will exist positive scalars μ_i and δ_i such that $\mu_i e_i, -\delta_i e_i \in \operatorname{conv} \{a_t, t \in T(\overline{x})\}$, $i = 1, \ldots, n$, and we get

$$\begin{aligned} 0_n \in \operatorname{int} \operatorname{conv} \{\mu_i e_i, i = 1, \ldots, n; -\delta_i e_i, i = 1, \ldots, n\} \\ \subset \operatorname{int} \operatorname{conv} \{a_t, t \in T(\overline{x})\}. \end{aligned} \tag{3.2}$$

Next we show that the so-called Haar's condition, $0_n \in \operatorname{conv} \{a_t, t \in T(\overline{x})\}$, which is obviously weaker than (3.2), is either sufficient or necessary for the uniqueness of $\overline{x}$, provided that σ belongs to certain classes of linear semi-infinite systems arising in approximation problems.

PROPOSITION 3.3. *Given $\overline{x} \in F$ such that $T(\overline{x}) \neq \emptyset$, the following statements hold:*

(i) *If $F = \{\overline{x}\}$ and σ is either LFM or continuous, then $0_n \in \operatorname{conv} \{a_t, t \in T(\overline{x})\}$.*

(ii) *If $0_n \in \operatorname{conv} \{a_t, t \in T(\overline{x})\}$ and $\{a_t, t \in S\}$ is linearly independent for every set $S \subset T(\overline{x})$ such that $|S| \leq n$, then $F = \{\overline{x}\}$.*

Proof. (i) Assume that $F = \{\overline{x}\}$. If σ is LFM, then the statement is a straightforward consequence of Remark 2. So we assume that σ is continuous. If σ contains the trivial inequality, $0_n'x \geq 0$, this inequality is active at $\overline{x}$ so that $0_n \in \{a_t, t \in T(\overline{x})\}$ and the conclusion is straightforward. Therefore, we shall assume that σ is continuous and it does not contain the trivial inequality, and Corollary 5.9.1 in [2] entails $T(\overline{x}) \neq \emptyset$. Next we apply Theorem 7.2 in [2] to conclude the inconsistency of the system $\{a_t'y > 0, t \in T(\overline{x})\}$, and this is equivalent to $0_n \in \operatorname{conv} \{a_t, t \in T(\overline{x})\}$ by Theorem 3.2 in [2] (Gordan's alternative theorem).

(ii) Assume that $0_n \in \operatorname{conv} \{a_t, t \in T(\overline{x})\}$. According to Carathéodory's theorem, there exists a set $S \subset T(\overline{x})$ and positive scalars $\overline{\lambda}_t, t \in S$, such that $0_n = \sum_{t \in S} \overline{\lambda}_t a_t$, $\sum_{t \in S} \overline{\lambda}_t = 1$, and $\{a_t, t \in S\}$ is affinely independent. The linear dependence of $\{a_t, t \in S\}$ entails (recalling the hypothesis) $|S| \geq n + 1$, and the equality holds due to the linear independence of $\{\binom{a_t}{1}, t \in S\}$. Then the linear system

$$\left\{ \sum_{t \in S} \lambda_t \begin{pmatrix} a_t \\ 1 \end{pmatrix} = \begin{pmatrix} 0_n \\ 1 \end{pmatrix} \right\} \tag{3.3}$$

has a square regular coefficient matrix. Since the unique solution of (3.3) is a positive vector of $\mathbb{R}^{n+1}$ (with components $\overline{\lambda}_t, t \in S$), the same will happen if we replace the right-hand side vector $\binom{0_n}{1}$ by an arbitrary vector $\binom{a}{1}$ such that $a \in \varepsilon B_n$, where ε is a sufficiently small positive scalar. Then

$$\varepsilon B_n \subset \operatorname{conv} \{a_t, t \in S\} \subset \operatorname{conv} \{a_t, t \in T(\overline{x})\}$$

so that

$$0_n \in \operatorname{int} \operatorname{conv} \{a_t, t \in T(\overline{x})\} \subset \operatorname{int} A(\overline{x}) \subset \operatorname{int} D(F, \overline{x})^0$$

and $F = \{\overline{x}\}$ by Proposition 3.1, part (i). □

Now let us consider the following problem which arises in functional approximation: Given a compact Hausdorff space T, three functions $f, g, h \in \mathcal{C}(T)$, where g and h are positive on T, and a polynomial of degree less than n, $p(t)$ such that $-g(t) \leq p(t) - f(t) \leq h(t)$ for all $t \in T$, decide whether there exists another polynomial satisfying the same conditions. Writing $p(t) = \sum_{i=1}^{n} \overline{x}_i t^{i-1}$, the problem consists of determining whether $\overline{x}$ is the unique solution of the linear semi-infinite system

$$\left\{ f(t) - g(t) \leq \sum_{i=1}^{n} x_i t^{i-1} \leq f(t) + h(t), t \in T \right\}. \tag{3.4}$$

The following result is Theorem 2.1 in [5], which extends a unicity theorem of Guerra and Jiménez [4] for the system in (3.4).

COROLLARY 3.4. *Let σ be a continuous system, and let $\overline{x} \in F$ such that $T(\overline{x}) \neq \emptyset$ and $\{a_t, t \in S\}$ is linearly independent for all $S \subset T(\overline{x})$ such that $|S| \leq n$. Then the following statements are equivalent to each other:*

(i) $F = \{\overline{x}\}$.

(ii) *There exists $S \subset T(\overline{x})$, with $|S| = n + 1$, such that $\{a_t' x > 0, t \in S\}$ is inconsistent.*

(iii) $0_n \in \operatorname{conv} \{a_t, t \in T(\overline{x})\}$.

Proof. (i)$\Longleftrightarrow$(iii) follows from Proposition 3.3, and (ii)$\Longleftrightarrow$(iii) follows from Gordan's alternative theorem (Theorem 3.2 in [2]) and Carathéodory's theorem. □

Example 1. Consider the system $\sigma = \{t_1 x_1 + t_2 x_2 \geq t_3, t \in T\}$, where

$$T = \left\{ t \in \mathbb{R}^3 \mid \left\| \begin{pmatrix} t_1 - 1 \\ t_2 \\ t_3 + 1 \end{pmatrix} \right\| \leq 1 \right\} \cup \left\{ \begin{pmatrix} -1 \\ 0 \\ 0 \end{pmatrix} \right\}.$$

Obviously, σ is a continuous system with

$$\operatorname{cl} K = \operatorname{cone} \left\{ \pm \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \pm \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ -1 \end{pmatrix} \right\}$$

so that $F = \{0_2\}$. Observe that $t \in T(0_2)$ if and only if $t_3 = (t_1, t_2)\, 0_2 = 0$ so that the isolated index in T, $(-1, 0, 0)'$, is active at 0_2. In order to obtain the remaining elements of $T(0_2)$ we replace $t_3 = 0$ in $\left\| (t_1 - 1, t_2, t_3 + 1)' \right\| \leq 1$, which yields $(t_1 - 1)^2 + t_2^2 \leq 0$, i.e., $t_1 = 1$ and $t_2 = 0$. Consequently,

$$T(0_2) = \left\{ \pm \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} \right\},$$

with $\{a_t, t \in T(0_2)\}$ linearly dependent. It can be easily seen that the statements (i) and (iii) in Corollary 3.4 hold, but (ii) fails (as well as the linear independence assumption). Observe also that

$$A(0_2) = \operatorname{span} \left\{ \begin{pmatrix} 1 \\ 0 \end{pmatrix} \right\} \subsetneq D(F, 0_2)^0 = \mathbb{R}^2$$

so that σ is not LFM, and four statements in Corollary 3.2 fail at 0_2.

4. γ-active constraints. Given $\overline{x} \in F$ and $\gamma > 0$, we define the set of γ -active constraints at $\overline{x}$ as

$$W(\overline{x}, \gamma) := \{a_t \mid t \in T \text{ and } a_t' y = b_t \text{ for a certain } y \in \overline{x} + \gamma B_n\}.$$

In other words, if $a_t \neq 0_n$, then $a_t \in W(\overline{x}, \gamma)$ if and only if $a_t' \overline{x} < b_t + \gamma \|a_t\|$. Obviously, $\{a_t,\, t \in T(\overline{x})\} \subset W(\overline{x}, \gamma)$. Moreover, if $\overline{x} \in \operatorname{int} F$ there will exist $\gamma_0 > 0$ sufficiently small such that $W(\overline{x}, \gamma) \setminus \{0_n\} = \emptyset$ for all γ such that $0 < \gamma < \gamma_0$.

LEMMA 4.1. *Given $\overline{x} \in \operatorname{bd} F$, the following statements hold:*

(i) *$W(\overline{x}, \gamma)$ contains at least a nonzero vector for all $\gamma > 0$.*

(ii) *If $T(\overline{x}) = \emptyset$, then $W(\overline{x}, \gamma)$ is an infinite set for all $\gamma > 0$.*

(iii) *If $|T| < \infty$, then $W(\overline{x}, \gamma) = \{a_t,\, t \in T(\overline{x})\}$ for $\gamma > 0$ sufficiently small.*

Proof. (i) Given an arbitrary $z \in (\overline{x} + \gamma B_n) \setminus F$, there exists $s \in T$ such that $a_s' z < b_s$. Since $a_s' \overline{x} \geq b_s$ there must exist $y \in [\overline{x}, z[\subset \overline{x} + \gamma B_n$ such that $a_s' y = b_s$. Then $0_n \neq a_s \in W(\overline{x}, \gamma)$.

(ii) Let $T(\overline{x}) = \emptyset$, and assume that $W(\overline{x}, \gamma) \setminus \{0_n\} = \{a_t,\, t \in S\}$, with $|S| < \infty$. Then we have

$$0 < \varepsilon := \min \left\{ \frac{a_t' \overline{x} - b_t}{\|a_t\|} \mid t \in S \right\} < \gamma.$$

Hence $a_t = 0_n$ for all $a_t \in W(\overline{x}, \varepsilon)$. This contradicts (i).

(iii) Assume $|T| < \infty$. By (ii), $T(\overline{x}) \neq \emptyset$. If either $T(\overline{x}) = T$ or $\{a_t, t \in T \backslash T(\overline{x})\} = \{0_n\}$, then

$$W(\overline{x}, \gamma) = \{a_t,\, t \in T(\overline{x})\} \tag{4.1}$$

for all $\gamma > 0$. Otherwise, (4.1) holds for all $\gamma > 0$ such that

$$\gamma < \min \left\{ \frac{a_t' \overline{x} - b_t}{\|a_t\|} \mid a_t \neq 0_n,\, t \in T \backslash T(\overline{x}) \right\}. \qquad \square$$

Next we show that the γ-active constraints at $\overline{x}$ allow us to check the feasibilty of given points of the ball $\overline{x} + \gamma B_n$ and of given directions at $\overline{x}$.

LEMMA 4.2. *Let $\overline{x} \in F$ and $y \in \overline{x} + \gamma B_n$, $\gamma > 0$. Then $y \in F$ if and only if $a_t' y \geq b_t$ for all $a_t \in W(\overline{x}, \gamma)$.*

Proof. In order to prove the nontrivial part of the statement, assume $y \notin F$. Then there exists $s \in T$ such that $a_s' y < b_s$. Since $a_s' \overline{x} \geq b_s$, there will exist $z \in [\overline{x}, y[\subset \overline{x} + \gamma B_n$ such that $a_s' z = b_s$. Then $a_s \in W(\overline{x}, \gamma)$, and we get a contradiction. $\square$

LEMMA 4.3. *Let $\overline{x} \in F$ and $d \in \mathbb{R}^n$. The following statements are true:*

(i) *If for a certain $\gamma > 0$ we have $a_t' d \geq 0$ for all $a_t \in W(\overline{x}, \gamma)$, then $d \in D(F, \overline{x})$. So $D(F, \overline{x})^0 \subset \operatorname{cl} \operatorname{cone} W(\overline{x}, \gamma)$ for all $\gamma > 0$.*

(ii) *If $d \in D(F, \overline{x})$ and $|T| < \infty$, then there exists some $\gamma_0 > 0$ such that $a_t' d \geq 0$ for all $a_t \in W(\overline{x}, \gamma)$ and all positive $\gamma < \gamma_0$. In such a case, $D(F, \overline{x})^0 = \operatorname{cone} W(\overline{x}, \gamma)$.*

Proof. (i) Assume that for some $\gamma > 0$, one has $a_t' d \geq 0$ for all $a_t \in W(\overline{x}, \gamma)$. We can assume $\|d\| = 1$. Taking an arbitrary ε such that $0 < \varepsilon < \gamma$, we have $\overline{x} + \varepsilon d \in \overline{x} + \gamma B_n$ and $a_t'(\overline{x} + \varepsilon d) = a_t' \overline{x} + \varepsilon a_t' d \geq b_t$ for all $a_t \in W(\overline{x}, \gamma)$ so that $\overline{x} + \varepsilon d \in F$ by Lemma 4.2. Hence $d \in D(F, \overline{x})$.

Since $[\operatorname{cone} W(\overline{x}, \gamma)]^0 \subset D(F, \overline{x})$, we have $D(F, \overline{x})^0 \subset \operatorname{cl} \operatorname{cone} W(\overline{x}, \gamma)$.

(ii) If $\overline{x} \in \operatorname{int} F$, the statement is true if $\gamma_0 > 0$ is such that $W(\overline{x}, \gamma_0) \setminus \{0_n\} = \emptyset$. Let $\gamma_0 > 0$ such that $W(\overline{x}, \gamma) = \{a_t,\, t \in T(\overline{x})\}$ for all $\gamma > 0$ such that $0 < \gamma < \gamma_0$ (recall statement (iii) in Lemma 4.1). Under the assumptions,

$$D(F, \overline{x}) = A(\overline{x})^0 = \{d \in \mathbb{R}^n \mid a_t' d \geq 0 \text{ for all } a_t \in W(\overline{x}, \gamma)\}.$$

Finally, observe that σ is LFM (since $|T| < \infty$) so that

$$\operatorname{cone} W(\overline{x}, \gamma) = A(\overline{x}) = D(F, \overline{x})^0. \qquad \square$$

Example 2. Consider $\sigma_1 = \{x_1 - t x_2 \geq 0, t \in T_1\}$, where $T_1 = [1, +\infty[$, $\overline{x} = (1,0)'$, and $d = (-1,0)'$. It can be easily realized that

$$F_1 = \left\{x \in \mathbb{R}^2 \mid x_1 - x_2 \geq 0, -x_2 \geq 0\right\}$$

so that $\overline{x} \in F_1$ and $d \in D(F_1, \overline{x}) = \left\{z \in \mathbb{R}^2 \mid z_2 \leq 0\right\}$. Nevertheless, $a_t' d = -1$ for all $t \in T_1$. Hence, the converse of statement (i) in Lemma 4.3 is not true. On the other hand, since $D(F_1, \overline{x})^0 = \operatorname{cone}\{\left(\begin{smallmatrix} 0 \\ -1 \end{smallmatrix}\right)\}$ and

$$W_1(\overline{x}, \gamma) = \begin{cases} \left\{\begin{pmatrix} 1 \\ -t \end{pmatrix} \mid t \geq 1\right\} \text{ if} \gamma \geq \frac{1}{\sqrt{2}}, \\ \left\{\begin{pmatrix} 1 \\ -t \end{pmatrix} \mid t > \sqrt{\gamma^{-2} - 1}\right\} \text{ if } 0 < \gamma < \frac{1}{\sqrt{2}}, \end{cases}$$

we have $D(F_1, \overline{x})^0 \cap \operatorname{cone} W_1(\overline{x}, \gamma) = \{0_2\}$ for all $\gamma > 0$, and statement (ii) fails for infinite systems. This example also shows that it is not possible to replace $\operatorname{cl}\operatorname{cone} W(\overline{x}, \gamma)$ by just $\operatorname{cone} W(\overline{x}, \gamma)$ in statement (i).

We consider now the LOP system obtained by aggregating the inequality $-x_2 \geq 0$ to σ_1; i.e., let $\sigma_2 = \{a_t' x \geq b_t, t \in T_2\}$, with $T_2 = T_1 \cup \{0\}$, $a_0 = \left(\begin{smallmatrix} 0 \\ -1 \end{smallmatrix}\right)$, and $b_0 = 0$. Obviously, $\overline{x} = (1,0)' \in F_2 = F_1$ and $W_2(\overline{x}, \gamma) = W_1(\overline{x}, \gamma) \cup \{a_0\}$ so that

$$D(F_2, \overline{x})^0 = D(F_1, \overline{x})^0 \subsetneqq \operatorname{cone} W_2(\overline{x}, \gamma).$$

Hence, statement (ii) fails even for LOP systems.

PROPOSITION 4.4. *Given $\overline{x} \in F$ and $\gamma > 0$, the following statements hold:*

(i) *If $F = \{\overline{x}\}$, then $0_n \in \operatorname{int}\operatorname{cone} W(\overline{x}, \gamma)$.*

(ii) *If $\overline{x} \in F^*$, then $c \in \operatorname{cl}\operatorname{cone} W(\overline{x}, \gamma)$.*

(iii) *If $\overline{x}$ is a strongly unique solution of* (P)*, then $c \in \operatorname{int}\operatorname{cone} W(\overline{x}, \gamma)$.*

(iv) *If $\overline{x} \in \operatorname{extr} F$, then $\dim \operatorname{cone} W(\overline{x}, \gamma) = n$.*

Proof. Lemma 4.3, part (i), stated $D(F, \overline{x})^0 \subset \operatorname{cl}\operatorname{cone} W(\overline{x}, \gamma)$ for all $\gamma > 0$. Then

$$\operatorname{int} D(F, \overline{x})^0 \subset \operatorname{int}\operatorname{cl}\operatorname{cone} W(\overline{x}, \gamma) = \operatorname{int}\operatorname{cone} W(\overline{x}, \gamma)$$

for all $\gamma > 0$. Now statements (i), (ii), and (iii) follow straightforwardly from Proposition 3.1, parts (i), (ii), and (iii), respectively.

(iv) Assume that $\dim \operatorname{cone} W(\overline{x}, \gamma) < n$. Let $d \in \mathbb{R}^n$, $d \neq 0_n$, such that $d \in [\operatorname{span} W(\overline{x}, \gamma)]^{\perp}$, i.e., $a_t' d = 0$ for all $a_t \in W(\overline{x}, \gamma)$. Then $\pm d \in D(F, \overline{x})$, again by Lemma 4.3, part (i), so that $\overline{x} \notin \operatorname{extr} F$. $\square$

The sets of γ-active constraints are too large in order to guarantee the converse statements in Proposition 4.4 separately. Next we show that all these sets together provide sufficient conditions for (Q1)–(Q4) in LP (but not in LSIP).

PROPOSITION 4.5. *Let $\overline{x} \in F$ and $|T| < \infty$. The following statements hold:*

(i) *If $0_n \in \operatorname{int} \operatorname{cone} W(\overline{x}, \gamma)$ for all $\gamma > 0$, then $F = \{\overline{x}\}$.*

(ii) *If $c \in \operatorname{cl} \operatorname{cone} W(\overline{x}, \gamma)$ for all $\gamma > 0$, then $\overline{x} \in F^*$.*

(iii) *If $c \in \operatorname{int} \operatorname{cone} W(\overline{x}, \gamma)$ for all $\gamma > 0$, then $\overline{x}$ is a strongly unique solution of* (P).

(iv) *If $\dim \operatorname{cone} W(\overline{x}, \gamma) = n$ for all $\gamma > 0$, then $\overline{x} \in \operatorname{extr} F$.*

Proof. (i) Assume that $F \neq \{\overline{x}\}$, and let $d \in D(F, \overline{x})$. According to Lemma 4.3, part (ii), there exists $\gamma > 0$ such that $d'z \geq 0$ for all $z \in \operatorname{conv} W(\overline{x}, \gamma)$. Then $0_n \notin \operatorname{int} \operatorname{cone} W(\overline{x}, \gamma)$.

The proofs of statements (ii)–(iv) are based upon Lemma 4.1, part (iii). In fact, let $\gamma > 0$ such that $W(\overline{x}, \gamma) = \{a_t, t \in T(\overline{x})\}$. Since $\operatorname{cone} W(\overline{x}, \gamma) = A(\overline{x})$ is closed, the conclusion follows from Corollary 3.2. □

The next example shows that it is not possible to replace in Proposition 4.5 the finiteness of σ by the (weaker) LOP property.

Example 3. Let $\sigma_0 = \{x_2 \leq 1; tx_1 + x_2 \leq 1, t \in \mathbb{N}\}$, let $\sigma_1 = \sigma_0 \cup \{x_1 \leq 0\}$, and observe that σ_1 is LOP. Obviously, $0_2 \in F_1 =]-\infty, 0] \times]-\infty, 1]$ and $0_2 \notin \operatorname{extr} F_1$ even though $\dim \operatorname{cone} W_1(0_2, \gamma) = 2$ for all $\gamma > 0$. Consider also $\sigma_2 = \sigma_1 \cup \{x_1 \geq 0, x_2 \geq 0\}$, which is also LOP, with $F_2 = \{0\} \times [0, 1]$. We have $0_2 \in \operatorname{int} \operatorname{cone} W_2(0_2, \gamma)$ for all $\gamma > 0$, but $F_2 \neq \{0_2\}$. Finally, taking $c = (0, -1)'$, we have $c \in \operatorname{int} \operatorname{cone} W_2(0_2, \gamma)$ for all $\gamma > 0$, although $0_2 \notin F_2^* = \{(0, 1)'\}$.

We have seen that each set $W(\overline{x}, \gamma)$, $\gamma > 0$, is too large for our purpose of giving a sensible answer to the questions (Q1)–(Q4). An alternative choice could be $\bigcap_{\gamma > 0} W(\overline{x}, \gamma)$, but this set is too small, since

$$\bigcap_{\gamma > 0} W(\overline{x}, \gamma) = \{a_t, t \in T(\overline{x})\} \tag{4.2}$$

can be empty even when $\overline{x} \in \operatorname{bd} F$. In fact, if $a_t \in \bigcap_{\gamma > 0} W(\overline{x}, \gamma)$, then $a_t \in W\left(\overline{x}, k^{-1}\right)$ for all $k \in \mathbb{N}$. Let $y^k \in \overline{x} + k^{-1} B_n$ such that $a_t' y^k = b_t$. Then $\lim_k y^k = \overline{x}$ and $a_t' \overline{x} = b_t$; i.e., $t \in T(\overline{x})$.

5. Extended active constraints. We discuss in this section three different sets of extended active constraints at $\overline{x} \in F$ which are motivated and defined as follows:

(a) We can replace N or K in (3.1) by the simpler (and smaller) set, $D = \{\binom{a_t}{b_t}, t \in T\}$. This suggests the following definition:

$$D(\overline{x}) := \left\{ a \in \mathbb{R}^n \mid \begin{pmatrix} a \\ a'\overline{x} \end{pmatrix} \in \operatorname{cl} D \right\}. \tag{5.1}$$

(b) We can replace in (4.2) each set of γ-active constraints, $W(\overline{x}, \gamma)$, by its closure, obtaining the set

$$W(\overline{x}) := \bigcap_{\gamma > 0} \operatorname{cl} W(\overline{x}, \gamma).$$

Since $\{W(\overline{x}, \gamma) \mid \gamma > 0\}$ is a contractive family of nonempty sets, when $\gamma \searrow 0$, it is possible to write

$$W(\overline{x}) = \lim_k W(\overline{x}, \gamma_k) \tag{5.2}$$

(limit in the Painlevé–Kuratowski sense) for any arbitrary sequence of positive real numbers, $\{\gamma_k\}_{k=1}^{\infty}$, such that $\lim_k \gamma_k = 0$.

(c) Finally, we can take the set of active constraints at $\overline{x} \in F$ of a suitable representation of F. According to Theorem 5.12 in [2], if G is the set of accumulation points of

$$\left\{ \left\| \begin{pmatrix} a_t \\ b_t \end{pmatrix} \right\|^{-1} \begin{pmatrix} a_t \\ b_t \end{pmatrix} \mid \begin{pmatrix} a_t \\ b_t \end{pmatrix} \neq 0_n, t \in T \right\},$$

then $\sigma \cup \{a'x \geq b, \binom{a}{b} \in G\}$ is another representation of F which is LFM if $\dim F = n$. Therefore, we consider the set

$$G(\overline{x}) := \left\{ a \in \mathbb{R}^n \mid \begin{pmatrix} a \\ a'\overline{x} \end{pmatrix} \in D \cup G \right\}. \tag{5.3}$$

PROPOSITION 5.1. *If $\overline{x} \in \operatorname{bd} F$ and σ is UB, then $W(\overline{x}) \neq \emptyset$.*

Proof. Let $\{\gamma_k\}_{k=1}^{\infty}$ be a sequence of positive scalars such that $\lim_k \gamma_k = 0$. By Lemma 4.1, part (i), there exists $a_{t_k} \in W(\overline{x}, \gamma_k)$, $k = 1, 2, \ldots$. Since $\{a_{t_k}\}_{k=1}^{\infty}$ is bounded, we can assume $\lim_k a_{t_k} = a \in \mathbb{R}^n$. Then, recalling (5.2), $a \in W(\overline{x})$. ∎

Now we show that $D(\overline{x})$ and $W(\overline{x})$ are essentially the same set.

LEMMA 5.2. *For any fixed $\overline{x} \in F$, one has $D(\overline{x}) \setminus \{0_n\} \subset W(\overline{x}) \subset D(\overline{x})$. Moreover, if σ is LB, then $0_n \notin D(\overline{x})$ (and so $W(\overline{x}) = D(\overline{x})$).*

Proof. Let $a \in D(\overline{x})$, $a \neq 0_n$. We shall prove that $a \in W(\overline{x})$.

According to (5.1), there exists $\{t_k\}_{k=1}^{\infty}$, sequence contained in T, such that

$$\lim_k \begin{pmatrix} a_{t_k} \\ b_{t_k} \end{pmatrix} = \begin{pmatrix} a \\ a'\overline{x} \end{pmatrix}. \tag{5.4}$$

The scalar product of both members of (5.4) by $\binom{\overline{x}}{-1}$ gives $\lim_k \left(a'_{t_k}\overline{x} - b_{t_k}\right) = 0$, with $\beta_k := a'_{t_k}\overline{x} - b_{t_k} \geq 0$ for all $k \in \mathbb{N}$. Since $\lim_k a_{t_k} = a \neq 0_n$ (again from (5.4)), we can assume $a_{t_k} \neq 0_n$ for all $k \in \mathbb{N}$.

Let $y^k := \overline{x} - \beta_k \|a_{t_k}\|^{-2} a_{t_k}$, $k = 1, 2, \ldots$. Since $a'_{t_k} y^k = a'_{t_k}\overline{x} - \beta_k = b_{t_k}$, and $\|y^k - \overline{x}\| = \frac{\beta_k}{\|a_{t_k}\|} < \frac{\beta_k + k^{-1}}{\|a_{t_k}\|}$, we have $a_{t_k} \in W(\overline{x}, \frac{\beta_k + k^{-1}}{\|a_{t_k}\|})$ for all $k \in \mathbb{N}$.

Since $\lim_k \frac{\beta_k + k^{-1}}{\|a_{t_k}\|} = 0$ we get, from (5.2), $a = \lim_k a_{t_k} \in W(\overline{x})$.

Conversely, assume $a \in W(\overline{x})$. Let $\{t_k\}_{k=1}^{\infty}$ be a sequence in T such that $a_{t_k} \in W\left(\overline{x}, k^{-1}\right)$ for all $k \in \mathbb{N}$ and $\lim_k a_{t_k} = a$. Then, for each $k \in \mathbb{N}$, we can take $y^k \in \overline{x} + k^{-1}B_n$ such that $a'_{t_k} y^k = b_{t_k}$. Since $\lim_k y^k = \overline{x}$, we have

$$\lim_k b_{t_k} = \lim_k a'_{t_k} y^k = a'\overline{x}.$$

Then $\binom{a}{a'\overline{x}} \in \operatorname{cl} D$ and, recalling (5.1), we get $a \in D(\overline{x})$.

Finally, if $0_n \in D(\overline{x})$, there exists a sequence in T, say $\{t_k\}_{k=1}^{\infty}$, such that $\lim_k \binom{a_{t_k}}{b_{t_k}} = 0_{n+1}$ and σ cannot be LB. Hence, if σ is LB, $W(\overline{x}) = D(\overline{x})$. ∎

Example 4. Let $n = 1$ and $\sigma = \left\{t^{-1}x \geq -t^{-1}, t \in \mathbb{N}\right\}$. Since $W(0, \gamma) = \emptyset$ for all γ such that $0 < \gamma < 1$, $W(0) = \emptyset$. Nevertheless, $0_2 \in \operatorname{cl} D$ so that $D(0) = \{0\}$.

LEMMA 5.3. *Given $\overline{x} \in F$, one has $\mathbb{R}_+ [D(\overline{x}) \setminus \{0_n\}] \subset \mathbb{R}_+ [G(\overline{x}) \setminus \{0_n\}]$. Further, if σ is LB and UB, then $\mathbb{R}_+ D(\overline{x}) = \mathbb{R}_+ G(\overline{x})$.*

Proof. Given $a \in D(\overline{x})$, $a \neq 0_n$, two cases are possible (according to (5.1)).

If $\binom{a}{a'\overline{x}} \in D$, we have $a \in G(\overline{x})$ (recall (5.3)).

Otherwise, $\binom{a}{a'\overline{x}} \in (\operatorname{cl} D)\backslash D$, and there exists $\{t_k\}_{k=1}^{\infty}$, sequence in T, such that $\binom{a_{t_k}}{b_{t_k}} \neq \binom{a}{a'\overline{x}}$, for all $k \in \mathbb{N}$, and

$$\lim_k \begin{pmatrix} a_{t_k} \\ b_{t_k} \end{pmatrix} = \begin{pmatrix} a \\ a'\overline{x} \end{pmatrix}. \tag{5.5}$$

From (5.5) and $\lim_k a_{t_k} = a \neq 0_n$, we can assert that no subsequence of $\{\binom{a_{t_k}}{b_{t_k}}\}_{k=1}^{\infty}$ tends to 0_{n+1}. Let $\varepsilon > 0$ such that $\|\binom{a_{t_k}}{b_{t_k}}\| > \varepsilon$ for all $k \in \mathbb{N}$. Since $0 < \|\binom{a_{t_k}}{b_{t_k}}\|^{-1} < \varepsilon^{-1}$ for all $k \in \mathbb{N}$, we can assume without loss of generality the existence of a scalar μ such that

$$\lim_k \left\| \begin{pmatrix} a_{t_k} \\ b_{t_k} \end{pmatrix} \right\|^{-1} = \mu \geq 0. \tag{5.6}$$

From (5.5) and (5.6) we get

$$\mu \begin{pmatrix} a \\ a'\overline{x} \end{pmatrix} = \lim_k \left\| \begin{pmatrix} a_{t_k} \\ b_{t_k} \end{pmatrix} \right\|^{-1} \begin{pmatrix} a_{t_k} \\ b_{t_k} \end{pmatrix}. \tag{5.7}$$

Since the right-hand side vector in (5.7) is unitary, we obtain $\mu > 0$. Two cases are possible in (5.7).

If there exists $k \in \mathbb{N}$ such that

$$\mu \begin{pmatrix} a \\ a'\overline{x} \end{pmatrix} = \left\| \begin{pmatrix} a_{t_k} \\ b_{t_k} \end{pmatrix} \right\|^{-1} \begin{pmatrix} a_{t_k} \\ b_{t_k} \end{pmatrix}, \tag{5.8}$$

then $a'_{t_k}\overline{x} - b_{t_k} = 0$, i.e., $t_k \in T(\overline{x})$, and so $a_{t_k} \in G(\overline{x})$. Then from (5.8) we get

$$a = \mu^{-1} \left\| \begin{pmatrix} a_{t_k} \\ b_{t_k} \end{pmatrix} \right\|^{-1} a_{t_k} \in \mathbb{R}_+ \left[G(\overline{x}) \setminus \{0_n\} \right].$$

Alternatively, if (5.8) fails for each $k \in \mathbb{N}$, from (5.7) we get $\mu\binom{a}{a'\overline{x}} \in G$ so that $\mu a \in G(\overline{x})$, and we get again $a \in \mathbb{R}_+ [G(\overline{x}) \setminus \{0_n\}]$.

Now we assume that σ is LB and UB. In such a case we have shown in Lemma 5.2 that $0_n \notin D(\overline{x})$. In the same way, $0_{n+1} \notin D \cup G$ so that $0_n \notin G(\overline{x})$. Hence $\mathbb{R}_+ D(\overline{x}) \subset \mathbb{R}_+ G(\overline{x})$. It remains to prove the reverse inclusion.

Given $a \in G(\overline{x})$, two cases can arise (again from (5.3)).

If $\binom{a}{a'\overline{x}} \in D$, then $\binom{a}{a'\overline{x}} \in \operatorname{cl} D$ and so $a \in D(\overline{x})$.

Thus we can assume that $\binom{a}{a'\overline{x}} \in G$. Then there exists a sequence in T, $\{t_k\}_{k=1}^{\infty}$, such that $0_{n+1} \neq \binom{a_{t_k}}{b_{t_k}} \neq \binom{a}{a'\overline{x}}$, for all $k \in \mathbb{N}$, and

$$\lim_k \left\| \begin{pmatrix} a_{t_k} \\ b_{t_k} \end{pmatrix} \right\|^{-1} \begin{pmatrix} a_{t_k} \\ b_{t_k} \end{pmatrix} = \begin{pmatrix} a \\ a'\overline{x} \end{pmatrix}. \tag{5.9}$$

Let $M_1 > 0$ such that $\|a_t\| > M_1$ for all $t \in T$.

On the other hand, from (5.9), we get

$$\lim_k \left\| \begin{pmatrix} a_{t_k} \\ b_{t_k} \end{pmatrix} \right\|^{-1} a_{t_k} = a \neq 0_n \tag{5.10}$$

so that the sequence in (5.10) is bounded, and there exists $M_2 > 0$ such that $\|\binom{a_{t_k}}{b_{t_k}}\|^{-1} \|a_{t_k}\| < M_2$ for all $k \in \mathbb{N}$.

Since $\|\binom{a_{t_k}}{b_{t_k}}\|^{-1} < \frac{M_2}{M_1}$, we can assume without loss of generality the existence of δ such that $\lim_k \|\binom{a_{t_k}}{b_{t_k}}\|^{-1} = \delta \geq 0$. From (5.10) and the UB property of σ we get $\delta > 0$. Hence (5.9) yields

$$\begin{pmatrix} \delta^{-1}a \\ \left(\delta^{-1}a\right)' \overline{x} \end{pmatrix} = \lim_k \begin{pmatrix} a_{t_k} \\ b_{t_k} \end{pmatrix} \in \operatorname{cl} D$$

so that $\delta^{-1}a \in D(\overline{x})$ and $a \in \mathbb{R}_+ D(\overline{x})$. □

Concerning the boundedness assumptions in Lemma 5.3, Example 4 (where $G(0) = \emptyset$ and $D(0) = \{0\}$) shows that the inclusion $\mathbb{R}_+ D(\overline{x}) \subset \mathbb{R}_+ G(\overline{x})$ can fail if σ is UB but not LB. Similarly, the system (in $\mathbb{R}$) $\{tx \geq -t^{-1}, t \in \mathbb{N}\}$ is LB (but not UB), $G(0) = \{1\}$, and $D(0) = \emptyset$. Hence both the LB and the UB properties are necessary for the equation $\mathbb{R}_+ D(\overline{x}) = \mathbb{R}_+ G(\overline{x})$.

We can now establish the main result in this paper.

PROPOSITION 5.4. *Given $\overline{x} \in F$, the following statements hold:*

(i) $A(\overline{x}) \subset \operatorname{cone} W(\overline{x}) = \operatorname{cone} D(\overline{x}) \subset \operatorname{cone} G(\overline{x}) \subset D(F, \overline{x})^0$.

(ii) *If σ is LFM, the five cones in* (i) *coincide.*

(iii) *If* $\dim F = n$, *then* $\operatorname{cone} G(\overline{x}) = D(F, \overline{x})^0$.

(iv) *If σ is UB and SS, then*

$$\operatorname{cone} D(\overline{x}) = \operatorname{cone} G(\overline{x}) = D(F, \overline{x})^0. \tag{5.11}$$

(v) *If D is closed, then* $A(\overline{x}) = \operatorname{cone} W(\overline{x}) = \operatorname{cone} D(\overline{x})$.

(vi) *If σ is continuous and LB, then* $A(\overline{x}) = \operatorname{cone} W(\overline{x}) = \operatorname{cone} D(\overline{x}) = \operatorname{cone} G(\overline{x})$.

Proof. (i) From (4.2), $\{a_t, t \in T(\overline{x})\} \subset W(\overline{x})$ so that $A(\overline{x}) \subset \operatorname{cone} W(\overline{x})$. On the other hand, from Lemmas 5.2 and 5.3 we obtain $\operatorname{cone} W(\overline{x}) = \operatorname{cone} D(\overline{x}) \subset \operatorname{cone} G(\overline{x})$. Finally, the inclusion $\operatorname{cone} G(\overline{x}) \subset D(F, \overline{x})^0$ is the consequence of the definition of $G(\overline{x})$ as the set of active constraints of a certain linear representation of F, $\sigma_1 = \sigma \cup \{a'x \geq b, \binom{a}{b} \in G\}$, taking into account the known relationship between the active cone and the positive polar of the cone of feasible directions at any feasible point.

(ii) It is a straightforward consequence of (i) and the equation $A(\overline{x}) = D(F, \overline{x})^0$ for LFM systems.

(iii) If $\dim F = n$, then σ_1 is an LFM representation of F (Theorem 5.12 in [2]) so that $\operatorname{cone} G(\overline{x}) = A_1(\overline{x}) = D(F, \overline{x})^0$.

(iv) If σ is UB and SS, then $\operatorname{cone} D(\overline{x}) = D(F, \overline{x})^0$ by Lemma 3.3 in [6], and the conclusion follows from (i).

(v) Under the assumption,

$$D(\overline{x}) = \left\{ a \in \mathbb{R}^n \mid \exists t \in T \text{ such that } \begin{pmatrix} a \\ a'\overline{x} \end{pmatrix} = \begin{pmatrix} a_t \\ b_t \end{pmatrix} \right\} = \{a_t, t \in T(\overline{x})\}$$

so that $\operatorname{cone} D(\overline{x}) = A(\overline{x})$.

(vi) It follows from (v) and Lemma 5.3. □

The non-LFM system σ_0 in Example 3 shows that we may have $\operatorname{cone} D(\overline{x}) \subsetneq \operatorname{cone} G(\overline{x})$ even though the full dimensional assumption in (iii) is fulfilled. In fact, it can be easily realized that $T(0_2) = \emptyset$, $G_0 = \{(-1, 0, 0)'\}$, $\operatorname{cone} G_0(0_2) = D(F, 0_2)^0 = \operatorname{cone}\{(-1, 0)'\}$, $D_0 = \{(t, -1, -1)', t = 0, 1, \dots\}$, and $A_0(0_2) = \operatorname{cone} W_0(0_2) = \operatorname{cone} D_0(0_2) = \{0_2\}$. Hence the assumption "$\sigma$ is UB" in statement (iii) is not superfluous.

Now let us consider again the pathological Example 1. It can be easily realized that $W(0_2) = D(0_2) = G(0_2) = \{\pm(\begin{smallmatrix}1\\0\end{smallmatrix})\}$ so that

$$A(0_2) = \operatorname{cone} W(0_2) = \operatorname{cone} D(0_2) = \operatorname{cone} G(0_2) = \operatorname{span}\left\{\begin{pmatrix}1\\0\end{pmatrix}\right\} \neq \mathbb{R}^2 = D(F, 0_2)^0.$$

This example shows that (5.11) may fail even though σ satisfies the LB and the UB properties. The following result is an immediate consequence of Proposition 5.4.

COROLLARY 5.5. *Let $\overline{x} \in F$, and let $U(\overline{x})$ be one of the three sets of extended active constraints ($W(\overline{x})$, $D(\overline{x})$, or $G(\overline{x})$). Then the following statements hold:*

(i) *If $0_n \in \operatorname{int}\operatorname{cone} U(\overline{x})$, then $F = \{\overline{x}\}$.*

(ii) *If $c \in \operatorname{cone} U(\overline{x})$, then $\overline{x}$ is an optimal solution of* (P).

(iii) *If $c \in \operatorname{int}\operatorname{cone} U(\overline{x})$, then $\overline{x}$ is a strongly unique solution of* (P).

(iv) *If $\dim \operatorname{cone} U(\overline{x}) = n$, then $\overline{x} \in \operatorname{extr} F$.*

The converse statements of (i)–(iii) are true if $\operatorname{cone} U(\overline{x}) = D(F, \overline{x})^0$(sufficient conditions are given in Proposition 5.4). Example 1 shows again the necessity of these assumptions. Taking $\overline{x} = 0_2$ and $c = (0,1)'$, we observe that the converse statements of (i)–(iv) fail for the three sets of extended active constraints. In fact, σ is not LFM, since $D(F, 0_2)^0 = \{0_2\}^0 = \mathbb{R}^2 \neq A(0_2)$.

Finally, observe that, for general systems, Proposition 4.4 and Corollary 5.5 can be seen as providing necessary conditions and sufficient conditions for (Q1)–(Q4), based upon different sets of extended active constraints.

REFERENCES

[1] E. J. ANDERSON, M. A. GOBERNA, AND M. A. LÓPEZ, *Locally polyhedral linear semi-infinite systems*, Linear Algebra Appl., 103 (1998), pp. 95–119.

[2] M. A. GOBERNA AND M. A. LÓPEZ, *Linear Semi-Infinite Optimization*, John Wiley and Sons., Chichester, UK, 1998.

[3] M. A. GOBERNA, M. A. LÓPEZ, AND M. I. TODOROV, *Unicity in linear optimization*, J. Optim. Theory Appl., 86 (1995), pp. 37–56.

[4] F. GUERRA AND M. A. JIMÉNEZ, *On feasible sets defined through Chebyshev approximation*, Math. Methods Oper. Res., 47 (1998), pp. 255–264.

[5] A. HASSOUNI AND W. OETTLI, *On regularity and optimality in nonlinear semi-infinite programming*, in Semi-Infinite Programming. Recent Advances, M. A. Goberna and M. A. López, eds., Kluwer, Dordrecht, The Netherlands, 2001, pp. 59–74.

[6] S. HELBIG AND M. I. TODOROV, *Unicity results for general linear semi-infinite optimization problems using a new concept of active constraints*, Appl. Math. Optim., 38 (1998), pp. 21–43.

[7] O. MANGASARIAN, *Uniqueness of solutions in linear programming*, Linear Algebra Appl., 25 (1979), pp. 151–162.

[8] R. PUENTE AND V. VERA DE SERIO, *Locally Farkas-Minkowski linear inequality systems*, Top, 7 (1999), pp. 103–121.

SIAM J. OPTIM.
Vol. 14, No. 3, pp. 621–645

FIRST-ORDER AND SECOND-ORDER CONDITIONS FOR ERROR BOUNDS*

ZILI WU† AND JANE J. YE†

Abstract. For a lower semicontinuous function f on a Banach space X, we study the existence of a positive scalar μ such that the distance function d_S associated with the solution set S of $f(x) \leq 0$ satisfies

$$d_S(x) \leq \mu \max\{f(x), 0\}$$

for each point x in a neighborhood of some point x_0 in X with $f(x) < \epsilon$ for some $0 < \epsilon \leq +\infty$. We give several sufficient conditions for this in terms of an abstract subdifferential and the Dini derivatives of f. In a Hilbert space we further present some second-order conditions. We also establish the corresponding results for a system of inequalities, equalities, and an abstract constraint set.

Key words. error bounds, existence of solutions, inequality systems, lower Dini derivatives, abstract subdifferentials, first-order conditions, second-order conditions

AMS subject classifications. 49J52, 90C26, 90C31

DOI. 10.1137/S1052623402412982

1. Introduction.

Let (X, d) be a metric space. For a proper and lower semicontinuous (l.s.c.) function $f : X \to (-\infty, \infty]$, denote the solution set of the inequality system $f(x) \leq 0$ by

$$S := \{x \in X : f(x) \leq 0\}$$

and the distance from a point $x \in X$ to the set S by

$$d_S(x) := \inf\{d(x, s) : s \in S\}$$

if S is nonempty. By convention, $d_S(x) = +\infty$ if S is empty.

Let T be a nonempty subset of X and let γ be a positive scalar. We say that the inequality system $f(x) \leq 0$ has an *error bound* of the pair (S, T) with exponent γ if the set S is nonempty and there exists a scalar $\mu > 0$ such that

$$d_S(x) \leq \mu[f(x)_+]^\gamma \quad \text{for all } x \in T,$$

where $f(x)_+ := \max\{f(x), 0\}$. For the case $\gamma = 1$, if

$$T = f^{-1}(0, \epsilon) := \{x \in X : 0 < f(x) < \epsilon\}$$

for some $0 < \epsilon < +\infty$ ($\epsilon = +\infty$), we simply say that the system $f(x) \leq 0$ (or the set S) has a *local* (*global*) *error bound*; if

$$T = B(x_0, \delta) := \{x \in X : d(x, x_0) < \delta\}$$

for some $x_0 \in S$ and $0 < \delta$, the set S is said to be *metrically regular* at x_0.

*Received by the editors August 12, 2002; accepted for publication (in revised form) May 29, 2003; published electronically December 19, 2003.
http://www.siam.org/journals/siopt/14-3/41298.html

†Department of Mathematics and Statistics, University of Victoria, Victoria, BC, Canada V8W 3P4 (ziliwu@email.com, janeye@math.uvic.ca). The research of the second author was supported by an NSERC grant.

Error bounds have important applications in sensitivity analysis of mathematical programming and in convergence analysis of some algorithms. In his seminal paper [8], Hoffman showed that a linear inequality system has a global error bound. For nonlinear inequality systems, the existence of error bounds usually requires some conditions. Most earlier results about error bounds are related to a continuous or convex system on R^n. The reader is referred to the recent survey papers [11, 14] and the references therein for a summary of the theory and applications of error bounds.

Recently Ng and Zheng [15, 16] and Wu and Ye [21, 22] studied l.s.c. inequality systems and presented several sufficient conditions for error bounds in terms of the lower Dini derivative and an abstract subdifferential. These results are mainly established for the case $T = f^{-1}(0, \epsilon)$ $(0 < \epsilon \leq +\infty)$. The first purpose of this paper is to extend and develop the above first-order conditions to the case $T = B(x_0, \delta) \cap f^{-1}(0, \epsilon)$, where $x_0 \in X$, $0 < \epsilon \leq +\infty$ and $0 < \delta \leq +\infty$. We do not assume that x_0 lies in the solution set S nor that δ is $+\infty$. However, our results are applicable to the cases $x_0 \in S$ and $\delta = +\infty$; that is, they serve as sufficient conditions not only for regularity (when $x_0 \in S$) but also for error bounds (when $\delta = +\infty$). The second purpose is to present a second-order sufficient condition for the existence of error bounds with exponents $1/2$ in a Hilbert space from which we can further obtain sufficient conditions for nonconvex quadratic systems. Our third purpose is to specify the first-order and second-order conditions for the following system of inequalities, equalities, and an abstract set:

$$\begin{aligned} g_i(x) &\leq 0 \text{ for all } i \in I := \{1, \ldots, m\}, \\ h_j(x) &= 0 \text{ for all } j \in J := \{1, \ldots, n\}, \\ x &\in C, \end{aligned}$$

where g_i and $|h_j|$ are l.s.c. and C is a nonempty closed subset of X.

It is worth pointing out that, unlike other error bound results, the nonemptiness of the solution set of an inequality system in ours comes as a conclusion instead of an assumption. Therefore, we can also use them as sufficient conditions for the existence of its solutions.

Apart from the above notation, the following concepts on nonsmooth analysis also are needed in this paper (see, e.g., [3, 4, 17]):

Let X be a normed linear space, let x and v be in X, and let $f : X \to (-\infty, +\infty]$ be finite at x.

- The lower Dini derivative of f at x in the direction v is

$$f^-(x; v) := \liminf_{\substack{u \to v \\ t \to 0^+}} \frac{f(x + tu) - f(x)}{t}.$$

- The upper Dini derivative of f at x in the direction v is

$$f^+(x; v) := \limsup_{\substack{u \to v \\ t \to 0^+}} \frac{f(x + tu) - f(x)}{t}.$$

- The *Clarke derivative* of f at x in the direction v is

$$f^\circ(x; v) := \limsup_{\substack{y \to x \\ t \to 0^+}} \frac{f(y + tv) - f(y)}{t}.$$

- The *Clarke subdifferential* of f at x is

$$\partial^\circ f(x) := \{\xi \in X^* : \langle \xi, v\rangle \le f^\circ(x; v) \text{ for all } v \in X\}.$$

When X is a Hilbert space, we say that a vector $\xi \in X$ is a *proximal subgradient* of f at x provided that there exist positive scalars M and δ such that

$$f(y) \ge f(x) + \langle \xi, y - x\rangle - M\|y - x\|^2 \text{ for all } y \in B(x, \delta).$$

The set of all such ξ, denoted by $\partial^\pi f(x)$, is referred to as the *proximal subdifferential* of f at x.

For each $\xi \in \partial^\pi f(x)$, we define the following second-order subderivatives:

$$d_L^2 f(x|\xi)(u) := \liminf_{t\to 0^+} \frac{f(x+tu) - f(x) - t\langle \xi, u\rangle}{t^2},$$
$$d_-^2 f(x|\xi)(u) := \liminf_{\substack{v\to u\\ t\to 0^+}} \frac{f(x+tv) - f(x) - t\langle \xi, v\rangle}{t^2},$$
$$d_+^2 f(x|\xi)(u) := \limsup_{\substack{v\to u\\ t\to 0^+}} \frac{f(x+tv) - f(x) - t\langle \xi, v\rangle}{t^2}.$$

Usually, for $u \in X$ and $\xi \in \partial^\pi f(x)$, we have

$$d_-^2 f(x|\xi)(u) \le d_L^2 f(x|\xi)(u) \le d_+^2 f(x|\xi)(u).$$

If f is a C^2 function with its first-order and second-order derivatives at x denoted by $\nabla f(x)$ and $\nabla^2 f(x)$, respectively, then, since $\partial^\pi f(x) = \{\nabla f(x)\}$, these second-order subderivatives coincide with each other and satisfy

$$d_L^2 f(x|\nabla f(x))(u) = d_-^2 f(x|\nabla f(x))(u) = d_+^2 f(x|\nabla f(x))(u) = \frac{1}{2}\langle \nabla^2 f(x)u, u\rangle.$$

For other second-order subderivatives, the reader is referred to [7, 17] and the references therein.

For a nonempty set C in a normed linear space X, ψ_C denotes the indicator function associated with the set C defined as below:

$$\psi_C(x) = \begin{cases} 0 & \text{if } x \in C,\\ +\infty & \text{otherwise.}\end{cases}$$

2. Sufficient conditions in terms of subdifferentials. We recall the concept of an abstract subdifferential introduced in [21].

DEFINITION 2.1. *Let X be a Banach space, and let $f : X \to (-\infty, +\infty]$ be l.s.c. at $x \in X$ with $f(x) < +\infty$. A subset of X^*, denoted by $\partial_\omega f(x)$, is called a* ∂_ω-subdifferential of f *at x if it has the following properties:*

- (ω_1) $\partial_\omega g(x) = \partial_\omega f(x)$ *if $g = f$ near x.*
- (ω_2) $0 \in \partial_\omega f(x)$ *when f attains a local minimum at x.*
- (ω_3) $\partial_\omega f(x) \subseteq L\overline{B}^*$ *if f is convex and Lipschitz of L near x.*
- (ω_4) *If $g : X \to (-\infty, +\infty]$ is Lipschitz near x, then for each $\xi \in \partial_\omega(f+g)(x)$ and each $\delta > 0$ there exist $x_1, x_2 \in B(x, \delta)$ such that*

$$-\delta < f(x_1) - f(x) < \delta, \quad -\delta < g(x_2) - g(x) < \delta, \quad \text{and } \xi \in \partial_\omega f(x_1) + \partial_\omega g(x_2) + \delta B^*,$$

where B^* is the open unit ball in X^* and $\overline{B}^*$ is its closure.

As indicated in [21], ∂_ω-subdifferentials include the Clarke subdifferential and the Michel–Penot subdifferential in a Banach space, the Fréchet subdifferential in an Asplund space, the proximal subdifferential in a Hilbert space, and the lower Dini subdifferential in R^n. Thus these subdifferentials can be taken as ∂_ω-subdifferentials in our main result of this section below whose proof is based on Ioffe's technique [9].

THEOREM 2.2. *Let X be a Banach space and let $f : X \to (-\infty, +\infty]$ be l.s.c. Suppose that, for some $x_0 \in X$, $0 < \delta \le +\infty$, $0 < \mu < +\infty$, and $0 < \epsilon \le \delta/(2\mu)$, the set $B(x_0, \delta/2) \cap f^{-1}(-\infty, \epsilon)$ is nonempty and*

$$\|\xi\|_* \ge \mu^{-1} \text{ for all } \xi \in \partial_\omega f(x) \text{ and each } x \in B(x_0, \delta) \cap f^{-1}(0, \epsilon).$$

Then $S := \{x \in X : f(x) \le 0\}$ is nonempty and

$$d_S(x) \le \mu f(x)_+ \text{ for all } x \in B\left(x_0, \frac{\delta}{2}\right) \cap f^{-1}(-\infty, \epsilon).$$

Moreover, if $x_0 \in S$, then the condition $0 < \epsilon \le \delta/(2\mu)$ can be replaced with $0 < \epsilon \le +\infty$.

Proof. Obviously it suffices to prove that

$$d_S(x) \le \mu f(x)_+ \text{ for all } x \in B\left(x_0, \frac{\delta}{2}\right) \cap f^{-1}(-\infty, \epsilon)$$

since this together with the nonemptiness of the set $B(x_0, \delta/2) \cap f^{-1}(-\infty, \epsilon)$ implies the nonemptiness of S.

Suppose that there were $u \in B(x_0, \delta/2) \cap f^{-1}(-\infty, \epsilon)$ such that $d_S(u) > \mu f(u)_+$. Then $u \notin S$ and hence $0 < f(u) < \epsilon$. In addition, we can choose $t > 1$ and $\alpha > 0$ such that

$$d_S(u) > t\mu f(u) := \gamma \text{ and } \begin{cases} \max\{\gamma, \|u - x_0\|\} \le \frac{\delta}{2+\alpha} & \text{for } 0 < \epsilon \le \delta/(2\mu); \\ \|u - x_0\| \le \frac{\delta}{2+\alpha} & \text{for } x_0 \in S \text{ and } 0 < \epsilon \le +\infty. \end{cases} \tag{1}$$

Thus $f(u)_+ = f(u) = \gamma(t\mu)^{-1}$ and hence

$$f(u)_+ \le \inf_{v \in X} f(v)_+ + \gamma(t\mu)^{-1}.$$

Note that the function $f(\cdot)_+$ is l.s.c. and bounded below. Applying Ekeland's variational principle [5] to $f(\cdot)_+$ with $\sigma = \gamma(t\mu)^{-1}$ and $\lambda = \gamma$, we find $x \in X$ satisfying

$$f(x)_+ \le f(u)_+, \tag{2}$$

$$\|x - u\| \le \gamma, \tag{3}$$

$$f(v)_+ + (t\mu)^{-1}h(v) \ge f(x)_+ \text{ for all } v \in X, \tag{4}$$

where $h(v) := \|v - x\|$. It follows from (1), (2), and (3) that $0 < f(x) < \epsilon$.

On the other hand, (4) implies that the function $f(v)_+ + (t\mu)^{-1}h(v)$ attains its minimum on X at x. Hence, by property (ω_2) in Definition 2.1,

$$0 \in \partial_\omega[f(x)_+ + (t\mu)^{-1}h(x)]. \tag{5}$$

Since f is l.s.c. and $0 < f(x)$, there exists $\delta_1 > 0$ such that

$$0 < f(y) \text{ for all } y \in B(x, \delta_1).$$

Thus, by property (ω_1) in Definition 2.1 and (5),

$$0 \in \partial_\omega (f + (t\mu)^{-1} h)(x). \tag{6}$$

Let $\epsilon_1 := \min\{f(x), (1-t^{-1})\mu^{-1}, \delta_1, \epsilon - f(u), \alpha\delta(2+\alpha)^{-1}\} > 0$. Then by property (ω_4) in Definition 2.1 and (6) there exist x_1 and x_2 both in $B(x, \epsilon_1)$ such that

$$f(x) - \epsilon_1 < f(x_1) < f(x) + \epsilon_1$$

and

$$0 \in \partial_\omega f(x_1) + \partial_\omega((t\mu)^{-1} h)(x_2) + \epsilon_1 B^*.$$

These inequalities with (2) mean that $x_1 \in B(x, \epsilon_1) \cap f^{-1}(0, \epsilon)$. The inclusion, by property (ω_3) in Definition 2.1, implies that there exists $\xi \in \partial_\omega f(x_1)$ such that

$$\|\xi\|_* < (t\mu)^{-1} + \epsilon_1 \le (t\mu)^{-1} + (1 - t^{-1})\mu^{-1} = \mu^{-1},$$

which contradicts the assumption since $x_1 \in f^{-1}(0, \epsilon)$ and, by the triangle inequality and (1),

$$\|x_1 - x_0\| \le \|x_1 - x\| + \|x - u\| + \|u - x_0\| < \epsilon_1 + \gamma + \frac{\delta}{2+\alpha}$$
$$\le \begin{cases} \frac{\alpha\delta}{2+\alpha} + \frac{2\delta}{2+\alpha} = \delta & \text{for } 0 < \epsilon \le \delta/(2\mu); \\ \frac{(1+\alpha)\delta}{2+\alpha} + d_S(u) \le \frac{(1+\alpha)\delta}{2+\alpha} + \|u - x_0\| \le \delta & \text{for } x_0 \in S \text{ and } 0 < \epsilon \le +\infty. \end{cases}$$

□

Remark 2.1. Note that the nonemptiness of S in Theorem 2.2 is a natural result of the inequality for error bounds and the nonemptiness of the set $B(x_0, \delta/2) \cap f^{-1}(-\infty, \epsilon)$. It is worth comparing Theorem 2.2 with [22, Theorem 4], in which the nonemptiness of S can follow from an existence theorem of minimum in [18]. When f is regular, $f^-(x; v) = f^\circ(x; v)$ holds for each $x \in X$ and $v \in X$. The condition that $f^-(x; h_x) \le -\mu^{-1}$ for some $\mu > 0$, each $x \in f^{-1}(0, \epsilon)$, and corresponding h_x in [22, Theorem 4] turns into $f^\circ(x; h_x) \le -\mu^{-1}$, which implies that $\|\xi\|_* \ge \mu^{-1}$ for each $\xi \in \partial^\circ f(x)$. So the corresponding result of [22, Theorem 4] can be obtained from Theorem 2.2 by taking $\delta = +\infty$ and $\partial_\omega = \partial^\circ$. Hence Theorem 2.2 provides a weaker condition for the existence of solutions for an inequality system than that in [22, Theorem 4].

Theorem 2.2 is an extension of [21, Theorem 3.1] in that $B(x_0, \delta) \cap f^{-1}(0, \epsilon)$, not just $B(x_0, \delta)$ or $f^{-1}(0, \epsilon)$, can be taken as a test set T. In particular, for the case where the test set $T = f^{-1}(0, \epsilon)$, Theorem 2.2 is a refinement of [21, Theorem 3.1], in which the nonemptiness of S is a part of the assumption, not of the conclusion. In addition, the inequality $d_S(x) \le \mu f(x)_+$ in [21, Theorem 3.1] holds only for all $x \in X$ with $f(x) < \epsilon/2$ instead of for all $x \in X$ with $f(x) < \epsilon$, as in Theorem 2.2. We thank Dr. Qiji Jim Zhu for his help in the proof of this improvement.

For an l.s.c. function f on a Hilbert space X, the *limiting subdifferential* $\partial_L f(x)$ of f at $x \in dom\, f$ is a set defined by

$$\partial_L f(x) := \{w\text{-}\lim \xi_i : \xi_i \in \partial^\pi f(x_i), x_i \to x, f(x_i) \to f(x)\}.$$

That is, $\partial_L f(x)$ consists of all vectors, each of which is the weak limit (that is what $w\text{-}\lim \xi_i$ signifies) of a weak convergent sequence $\{\xi_i\}$, where $\xi_i \in \partial^\pi f(x_i)$ with $x_i \to x$

and $f(x_i) \to f(x)$. It is easy to check that the limiting subdifferential satisfies (ω_1)–(ω_3) in Definition 2.1. In addition, if at least one of functions f and g is Lipschitz near x, then

$$\partial_L(f+g)(x) \subseteq \partial_L f(x) + \partial_L g(x)$$

[4, Proposition 10.1, p. 62]; that is, the sum rule holds. So the limiting subdifferential is a ∂_ω-subdifferential and Theorem 2.2 is applicable to it. The following is a version of Theorem 2.2 with $\partial_\omega = \partial_L$ and f replaced with $f + \psi_C$.

COROLLARY 2.3. *Let X be a Hilbert space, let C be a closed subset of X, and let $f_i : X \to R$ be locally Lipschitz continuous for each $i \in I$. Denote*

$$f(x) = \max\{f_i(x) : i \in I\} \quad \text{and} \quad I(x) := \{i \in I : f_i(x) = f(x)\} \text{ for } x \in X.$$

Suppose that, for some $x_0 \in X$, $0 < \delta \le +\infty$, $0 < \mu < +\infty$, and $0 < \epsilon \le \delta/(2\mu)$, the set $C \cap B(x_0, \delta/2) \cap f^{-1}(-\infty, \epsilon)$ is nonempty and

$$\|\xi\|_* \ge \mu^{-1} \text{ for all } \xi \in co\{\partial_L f_i(x) : i \in I(x)\} + N_C^L(x) \text{ for all } x \in C \cap B(x_0, \delta) \cap f^{-1}(0, \epsilon),$$

where $co\, A$ denotes the convex hull of a set A and $N_C^L(x) := \partial_L \psi_C(x)$. Then $S := \{x \in C : f(x) \le 0\}$ is nonempty and

$$d_S(x) \le \mu f(x)_+ \text{ for all } x \in C \cap B\left(x_0, \frac{\delta}{2}\right) \cap f^{-1}(-\infty, \epsilon).$$

Moreover, if $x_0 \in S$, then the condition $0 < \epsilon \le \delta/(2\mu)$ can be replaced with $0 < \epsilon \le +\infty$.

Proof. For each $\xi \in \partial_L f(x)$, by the conclusion in [4, Problem 11.17, p. 65] and the sum rule, there exist $\gamma_i \ge 0\, (i \in I(x))$ with $\sum_{i \in I(x)} \gamma_i = 1$ such that

$$\xi \in \partial_L \left(\sum_{i \in I(x)} \gamma_i f_i \right)(x) \subseteq co\{\partial_L f_i(x) : i \in I(x)\}.$$

Hence applying Theorem 2.2 to $\partial_\omega = \partial_L$ with f replaced with $f + \psi_C$ completes the proof. ∎

Next we use Theorem 2.2 to prove a result about the regularity of a set at a point.

THEOREM 2.4. *Let X be a separable Hilbert space, C a closed subset of X, and $x_0 \in C$. Suppose that $g : X \to R^m$ and $h : X \to R^n$ are Lipschitz near x_0 and*

$$f(x) = \max_{i,\, j}\{g_i(x), |h_j(x)|\}.$$

If the constraint qualification

$$\left.\begin{aligned} 0 \le \gamma \in R^m,\ \gamma_i[g_i(x_0) - f(x_0)] = 0,\ i \in I \\ \lambda \in R^n,\ \lambda_j[|h_j(x_0)| - f(x_0)] = 0,\ j \in J \\ 0 \in \partial_L[\langle \gamma, g\rangle + \langle \lambda, h\rangle](x_0) + N_C^L(x_0) \end{aligned}\right\} \Rightarrow \gamma = 0 \text{ and } \lambda = 0$$

(where $N_C^L(x_0) := \partial_L \psi_C(x_0)$) is satisfied at x_0, then there exist $0 < \delta < +\infty$ and $0 < \mu < +\infty$ such that

$$\|\xi\| \ge \mu^{-1} \text{ for all } \xi \in \partial^\pi(f + \psi_C)(x) \text{ and each } x \in C \cap B(x_0, \delta) \cap f^{-1}(0, +\infty).$$

Moreover, if the set $C \cap B(x_0, \delta/2) \cap f^{-1}[0, \epsilon)$ *is nonempty for some* $0 < \epsilon \le \delta/(2\mu)$, *then the set* $S := \{x \in C : g(x) \le 0,\ h(x) = 0\}$ *is nonempty and*

$$d_S(x) \le \mu f(x) \text{ for all } x \in C \cap B\left(x_0, \frac{\delta}{2}\right) \cap f^{-1}(0, \epsilon).$$

In particular, if $x_0 \in S$, *then* S *is metrically regular at* x_0. *If* x_0 *lies in the interior of* C, *then the above conclusions hold in every Hilbert space* X.

Proof. Suppose that there did not exist $0 < \delta < +\infty$ and $0 < \mu < +\infty$ such that

$$\|\xi\| \ge \mu^{-1} \text{ for all } \xi \in \partial^\pi(f + \psi_C)(x) \text{ and each } x \in C \cap B(x_0, \delta) \cap f^{-1}(0, +\infty).$$

Then there would exist sequences

$$C \ni x_k \to x_0, \quad f(x_k) > 0, \quad \xi_k \in \partial^\pi(f + \psi_C)(x_k), \quad \|\xi_k\| \to 0.$$

If x_0 is in the interior of C, then $\xi_k \in \partial^\pi f(x_k)$ and $\|\xi_k\| \to 0$ imply that $0 \in \partial_L f(x_0)$. Thus there exist $0 \le \gamma \in R^m$ and $\lambda \in R^n$ such that

$$\gamma_i[g_i(x_0) - f(x_0)] = 0 \text{ for } i \in I, \tag{7}$$

$$\lambda_j[|h_j(x_0)| - f(x_0)] = 0 \text{ for } j \in J, \tag{8}$$

$$\sum_{i=1}^{m} \gamma_i + \sum_{j=1}^{n} |\lambda_j| = 1, \tag{9}$$

$$0 \in \partial_L[\langle \gamma, g\rangle + \langle \lambda, h\rangle](x_0)$$

(see [4, Problem 1.11.17, p. 65]), which contradicts the assumption.

If x_0 is not in the interior of C, then since f is Lipschitz near x_k when k is large enough, by [4, Theorem 1.8.3, p. 56], there exist $y_k \to x_0$, $C \ni z_k \to x_0$, $\eta_k \in \partial^\pi f(y_k)$, and $\zeta_k \in \partial^\pi \psi_C(z_k)$ such that $f(y_k) > 0$ and

$$\xi_k \in \eta_k + \zeta_k + B\left(x_0, \frac{1}{k}\right). \tag{10}$$

Since $\partial^\pi f(y_k) \subseteq \partial_L f(y_k)$, for k large enough so that y_k enters some prescribed neighborhood of x_0 on which f is Lipschitz, there exist $0 \le \gamma^k \in R^m$ and $\lambda^k \in R^n$ such that

$$\gamma_i^k[g_i(y_k) - f(y_k)] = 0 \text{ for } i \in I,$$

$$\lambda_j^k[|h_j(y_k)| - f(y_k)] = 0 \text{ for } j \in J,$$

$$\sum_{i=1}^{m} \gamma_i^k + \sum_{j=1}^{n} |\lambda_j^k| = 1,$$

$$\eta_k \in \partial_L[\langle \gamma^k, g\rangle + \langle \lambda^k, h\rangle](y_k).$$

By extracting convergent subsequences of $\{\gamma^k\}$ and $\{\lambda^k\}$ (we do not relabel them) and taking the limit of (γ^k, λ^k), we obtain a nonzero $(\gamma, \lambda) \in R^m \times R^n$ satisfying (7)–(9).

Note that

$$\partial_L[\langle \gamma^k, g\rangle + \langle \lambda^k, h\rangle](y_k) \subseteq \partial^\circ[\langle \gamma^k, g\rangle + \langle \lambda^k, h\rangle](y_k)$$

and the set on the right-hand side is contained in a ball of the form $L\overline{B}_*$ (for some positive L) which is weak* compact when k is large enough. There is a weakly convergent subsequence of $\{\eta_k\}$ (without relabeling) corresponding to (γ^k, λ^k) whose weak limit lies in

$$\partial_L[\langle\gamma, g\rangle + \langle\lambda, h\rangle](x_0)$$

since X is a separable Hilbert space (see [4, Problem 1.11.21, p. 66]).

In addition, corresponding to η_k, by (10), the sequence $\{\zeta_k\}$ contains a weakly convergent subsequence with its limit belonging to $N_C^L(x_0)$. Therefore we have

$$0 \in \partial_L[\langle\gamma, g\rangle + \langle\lambda, h\rangle](x_0) + N_C^L(x_0),$$

but (γ, λ) is nonzero. This is again a contradiction.

The rest follows immediately from the conclusion shown above and from Theorem 2.2. □

Remark 2.2. Theorem 2.4 is a refinement of [4, Theorem 3.8, p. 131] in that x_0 may not be in S and an abstract constraint set is allowed. In a general Banach space, one relevant result about metrical regularity in terms of Clarke subdifferentials can be found in [3, Theorem 6.6.1]. However, in Hilbert space where limiting subdifferential is applicable, our constraint qualification is weaker than that in [3, Theorem 6.6.1].

If $x_0 \in S$, $g_1, \ldots, g_m, h_1, h_2, \ldots, h_n$ are all C^1 functions and $C = X$, the constraint qualification in Theorem 2.4 is equivalent to the Mangasarian–Fromovitz constraint qualification in mathematical programming. In particular, if

$$\nabla g_1(x_0), \ldots, \nabla g_m(x_0), \nabla h_1(x_0), \ldots, \nabla h_n(x_0)$$

are linearly independent, then the Mangasarian–Fromovitz constraint qualification is satisfied at x_0.

Example 2.1. For $x \in R^3$, let

$$f_1(x) := ax_1 + g_1(x_2, x_3), \quad f_2(x) = bx_2 + g_2(x_3), \quad f_3(x) = cx_3,$$

where a, b, and c are nonzero constants while g_1 and g_2 are locally Lipschitz continuous. Since, for any point $x_0 \in R^3$, $\nabla f_1(x_0), \nabla f_2(x_0), \nabla f_3(x_0)$ are linearly independent, by Theorem 2.4, the system $S = \{x \in R^3 : f(x) \le 0\}$ with $f(x) := \max\{f_1(x), f_2(x), f_3(x)\}$ is metrically regular at any $x_0 \in S$.

Note that for an l.s.c. convex function f on a Banach space X the Clarke subdifferential of f at $x \in X$ reduces to the subdifferential of f at x in the sense of convex analysis given by

$$\partial f(x) := \{\xi \in X^* : \langle\xi, y - x\rangle \le f(y) - f(x) \text{ for all } y \in X\}.$$

It has been shown in [22] that for a convex inequality system a global error bound exists iff a local error bound does, and many first-order sufficient conditions for the existence of error bounds become necessary as well. In the following result, we use $\partial f(x)$ to develop the sufficient condition stated in Theorem 2.2 into a necessary one for a convex system.

THEOREM 2.5. *Let X be a Banach space, let $f : X \to (-\infty, +\infty]$ be l.s.c. and convex, and let $S := \{x \in X : f(x) \le 0\}$. Then for some $x_0 \in X$ and $0 < \mu < +\infty$ the following are equivalent:*

(i) *For some* $0 < \delta \le +\infty$, *each* $0 < \epsilon \le \delta/(2\mu)$, *and each* $\delta' \in (0, \delta)$ *the set* $B(x_0, \delta') \cap f^{-1}(-\infty, \epsilon)$ *is nonempty and*

$$\|\xi\|_* \ge \mu^{-1} \text{ for all } \xi \in \partial f(x) \text{ and each } x \in B(x_0, \delta) \cap f^{-1}(0, \epsilon).$$

(ii) *For some* $0 < \delta \le +\infty$, *each* $0 < \epsilon \le \delta/(2\mu)$, *and each* $\delta' \in (0, \delta)$ *the set* $B(x_0, \delta') \cap f^{-1}(-\infty, \epsilon)$ *is nonempty and*

$$d_S(x) \le \mu f(x)_+ \text{ for all } x \in B(x_0, \delta) \cap f^{-1}(-\infty, \epsilon).$$

In particular, if $x_0 \in S$, *then* (i) *and* (ii) *are equivalent to each other with "each* $0 < \epsilon \le \delta/(2\mu)$*" in both replaced by "some* $0 < \epsilon \le +\infty$*."*

Proof. (i) $\Rightarrow$ (ii) This is immediate from Theorem 2.2 by taking $\partial_\omega f(x) = \partial f(x)$.

(ii) $\Rightarrow$ (i) Let $x \in B(x_0, \delta) \cap f^{-1}(0, \epsilon)$. Then $d_S(x) > 0$ and for any $\xi \in \partial f(x)$ we have

$$\|\xi\|_* \cdot \|y - x\| \ge -\langle \xi, y - x\rangle \ge -[f(y) - f(x)] \ge f(x) \text{ for all } y \in S.$$

This implies $\|\xi\|_* \cdot d_S(x) \ge f(x)$, from which we have

$$\|\xi\|_* \ge \frac{f(x)}{d_S(x)} \ge \mu^{-1}.$$

Therefore the desired inequality follows. □

3. Second-order conditions. In mathematical programming, it is known that a second-order sufficient condition implies strict local minimum of order 2. This idea can be applied to error bounds. For a nonnegative function $f : R^n \to R$, consider the inequality system $S = \{x \in R^n : f(x) \le 0\}$. If $x_0 \in S$, f is twice continuously differentiable near x_0, and there exist $\mu > 0$ and $\delta > 0$ such that

$$\langle \nabla^2 f(x')u, u\rangle \ge \mu^{-1} \quad \text{for each unit vector } u \in R^n \text{ and } x' \in B(x_0, \delta), \tag{11}$$

then for each $x \in B(x_0, \delta)$, by the Taylor expansion, there exists $x' \in [x_0, x]$ such that

$$f(x) = \frac{1}{2}\langle \nabla^2 f(x')(x - x_0), x - x_0\rangle,$$

which along with (11) implies that

$$f(x) \ge \frac{1}{2\mu}\|x - x_0\|^2.$$

Thus

$$d_S^2(x) \le 2\mu f_+(x) \text{ for all } x \in B(x_0, \delta).$$

Note that under the above assumption, S must be a singleton. In studying weak sharp minima, several authors, including Bonnans and Ioffe [1, 2] and Ward [20] have extended the above result to include the case where f is not twice continuously differentiable and the solution set S is not a singleton by using certain second-order subderivatives. In the following main result in this section, we present a second-order sufficient condition for the existence of error bound with exponent 1/2. Note that if f

is nonnegative and twice continuously differentiable, then our second-order condition in Theorem 3.1 amounts to

$$\langle \nabla^2 f(x)u_x, u_x\rangle \le -4\mu^{-1} \quad \text{for some unit vector } u_x \in X \text{ and each } x \notin S.$$

Hence, unlike the second-order condition of type (11), which requires certain convexity, our second-order condition is suitable for nonconvex systems.

THEOREM 3.1. *Let X be a Hilbert space and let $f : X \to (-\infty, +\infty]$ be l.s.c. Suppose that, for some $x_0 \in X$, $0 < \delta \le +\infty$, $0 < \mu$, and $0 < \epsilon \le (2\mu)^{-1}(\delta/2)^2$, the set $B(x_0, \delta/2) \cap f^{-1}(-\infty, \epsilon)$ is nonempty and that, for each $x \in B(x_0, \delta) \cap f^{-1}(0, \epsilon)$, one of the following is satisfied for each $\xi \in \partial^\pi f(x)$ with $\|\xi\| \le \min\{2\sqrt{2\epsilon}\,\mu^{-1/2}, \delta\mu^{-1}\}$:*

(i) *There exists a unit vector u_x such that $d^2_- f(x|\xi)(u_x) \le -2\mu^{-1}$.*

(ii) *There exist sequences $t_n \to 0^+$ in R and $\{u_n\}$ in X such that $\lim_{n\to+\infty} \|u_n\| = 1$ and*

$$\lim_{n\to+\infty} \frac{f(x + t_n u_n)_+ - f(x)_+ - t_n\langle \xi, u_n\rangle}{t_n^2} \le -2\mu^{-1}.$$

Then $S := \{x \in X : f(x) \le 0\}$ is nonempty and

$$d_S^2(x) \le 2\mu f(x)_+ \text{ for all } x \in B\left(x_0, \frac{\delta}{2}\right) \cap f^{-1}(0, \epsilon).$$

Moreover, if $x_0 \in S$, then the condition $0 < \epsilon \le (2\mu)^{-1}(\delta/2)^2$ can be replaced with $0 < \epsilon \le +\infty$.

Proof. By the definition of the second-order subderivative, condition (i) implies condition (ii). Hence it suffices to prove the theorem under condition (ii).

We now prove the theorem by contradiction. Suppose that there were $u \in B(x_0, \delta/2) \cap f^{-1}(0, \epsilon)$ such that $d_S^2(u) > 2\mu f(u)_+$. We choose $t > 1$ such that

$$4\gamma := 2t\mu f(u) < \begin{cases} \min\{d_S^2(u), (\frac{\delta}{2})^2\} & \text{for } 0 < \epsilon \le (2\mu)^{-1}(\frac{\delta}{2})^2; \\ d_S^2(u) & \text{for } x_0 \in S \text{ and } 0 < \epsilon \le +\infty. \end{cases} \tag{12}$$

Thus $f(u) = 2\gamma(t\mu)^{-1}$ and hence

$$f(u)_+ \le \inf_{v\in X} f(v)_+ + 2\gamma(t\mu)^{-1}.$$

Note that the function $f(\cdot)_+$ is l.s.c. and bounded below. Applying smooth variational principle [4, Theorem 4.2, p. 43] to $f(\cdot)_+$ with $\sigma = 2\gamma(t\mu)^{-1}$ and $\lambda = \sqrt{\gamma}$, we find $x, y \in X$ satisfying

$$\|y - u\| < \lambda, \quad \|x - y\| < \lambda, \quad f(x)_+ \le f(u)_+$$

and

$$f(v)_+ + 2(t\mu)^{-1}h(v) \ge f(x)_+ + 2(t\mu)^{-1}h(x) \text{ for all } v \in X, \tag{13}$$

where $h(v) := \|v - y\|^2$. Thus

$$\|x - u\| \le \|x - y\| + \|y - u\| < 2\lambda = 2\sqrt{\gamma} < d_S(u)$$

and, by the triangle inequality and (12),

$$\begin{aligned} \|x - x_0\| &\le \|x - u\| + \|u - x_0\| \\ &< \begin{cases} 2\sqrt{\gamma} + \frac{\delta}{2} < \min\{d_S(u), \frac{\delta}{2}\} + \frac{\delta}{2} \le \delta & \text{for } 0 < \epsilon \le (2\mu)^{-1}(\frac{\delta}{2})^2; \\ d_S(u) + \|u - x_0\| < \delta & \text{for } x_0 \in S \text{ and } 0 < \epsilon \le +\infty \end{cases} \end{aligned}$$

and hence $x \in B(x_0, \delta) \cap f^{-1}(0, \epsilon)$.

On the other hand, from (13) and [4, Proposition 1.2.11, p. 38], we have

$$0 \in \partial^\pi(f(x) + 2(t\mu)^{-1}h(x)) = \partial^\pi f(x) + 2(t\mu)^{-1}\{2(x-y)\}. \tag{14}$$

This implies that $\xi := 4(t\mu)^{-1}(y-x) \in \partial^\pi f(x)$ and hence, by (12),

$$\begin{aligned}\|\xi\| &\le 4\|y-x\|(t\mu)^{-1} < 4\lambda(t\mu)^{-1} = 4\sqrt{\gamma}(t\mu)^{-1} = 2\sqrt{2t\mu f(u)}(t\mu)^{-1}\\ &\le 2\min\left\{\sqrt{2t\mu\epsilon}, d_S(u), \frac{\delta}{2}\right\}(t\mu)^{-1} < \min\{2\sqrt{2\epsilon}\,\mu^{-1/2}, \delta\mu^{-1}\}.\end{aligned}$$

So for the sequences $\{t_n\}$ and $\{u_n\}$ given in condition (ii) corresponding to ξ, by (13), we have

$$\begin{aligned}&\lim_{n\to+\infty}\frac{f(x+t_nu_n)_+ - f(x)_+ - t_n\langle\xi, u_n\rangle}{t_n^2}\\ &= \lim_{n\to+\infty}\frac{f(x+t_nu_n)_+ + 2(t\mu)^{-1}h(x+t_nu_n) - f(x)_+ - 2(t\mu)^{-1}h(x)}{t_n^2} - 2(t\mu)^{-1}\\ &\ge -2(t\mu)^{-1} > -2\mu^{-1},\end{aligned}$$

which contradicts condition (ii). ∎

To put first-order and second-order conditions together, we will use the following relation between a global error bound and a local error bound.

PROPOSITION 3.2. *Let (X, d) be a metric space, let $f : X \to (-\infty, +\infty]$ be proper, and let $S := f^{-1}(-\infty, 0]$. Then the following are equivalent:*

(i) *There exist $0 < \epsilon_1 < \epsilon_2 \le +\infty$ and $0 < \mu_1, \mu_2 < +\infty$ such that*

$$\begin{aligned}d_S(x) &\le \mu_1 f(x)_+ \quad \text{for all } x \in f^{-1}(0, \epsilon_1) \text{ and}\\ d_{S_1}(x) &\le \mu_2 f(x)_+ \quad \text{for all } x \in f^{-1}[\epsilon_1, \epsilon_2),\end{aligned}$$

where $S_1 := f^{-1}(-\infty, \epsilon_1)$.

(ii) *There exist $0 < \epsilon \le +\infty$ and $0 < \mu < +\infty$ such that*

$$d_S(x) \le \mu f(x)_+ \quad \text{for all } x \in f^{-1}(0, \epsilon).$$

Proof. The implication (ii) ⇒ (i) is immediate. We only need to show (i) ⇒ (ii). Let $0 < \epsilon_1 < \epsilon_2 \le +\infty$ and $0 < \mu_1, \mu_2 < +\infty$ satisfy

$$\begin{aligned}d_S(x) &\le \mu_1 f(x)_+ \quad \text{for all } x \in f^{-1}(0, \epsilon_1) \text{ and}\\ d_{S_1}(x) &\le \mu_2 f(x)_+ \quad \text{for all } x \in f^{-1}[\epsilon_1, \epsilon_2),\end{aligned}$$

where $S_1 := f^{-1}(-\infty, \epsilon_1)$. Note that for any fixed $x \in f^{-1}[\epsilon_1, \epsilon_2)$ and each $y \in f^{-1}(-\infty, \epsilon_1)$ we have

$$d_S(x) \le d_S(y) + d(x, y) \le \mu_1 f(y)_+ + d(x, y) \le \mu_1\epsilon_1 + d(x, y).$$

Taking the inferior of the right-hand side expression in the above inequalities for y over $f^{-1}(-\infty, \epsilon_1)$ yields $d_S(x) \le \mu_1\epsilon_1 + d_{S_1}(x)$. And hence

$$d_S(x) \le \mu_1\epsilon_1 + \mu_2 f(x)_+ \le (\mu_1 + \mu_2)f(x)_+ = \mu f(x)_+$$

for $\mu := \mu_1 + \mu_2$. Therefore, (ii) holds for $\epsilon = \epsilon_2$. ∎

Remark 3.1. When X is a normed linear space and f is convex, we can prove that the nonemptiness of S and the first inequality in (i) of Proposition 3.2 imply the second inequality in it. So Proposition 3.2 is an extension of [22, Proposition 2], which states that for a convex system a local error bound implies a global error bound.

Next, we use Proposition 3.2 and Theorems 2.2 and 3.1 to give a mixed condition.

THEOREM 3.3. *Let X be a Hilbert space, and let $f : X \to (-\infty, +\infty]$ be continuous. Denote*

$$D(\mu) := \{x \in X : 0 < f(x) \text{ and } \|\xi\| \le \mu^{-1} \text{ for some } \xi \in \partial^\pi f(x)\} \text{ for } \mu > 0.$$

Suppose that there exist $0 < \epsilon_1 < \epsilon_2 \le +\infty$ and $0 < \mu_1, \mu_2$ such that the set $f^{-1}(-\infty, \epsilon_1)$ is nonempty and the following conditions hold:

(i) $D(\mu_1) \subseteq f^{-1}(\epsilon_1, \epsilon_2)$.

(ii) *For each $x \in f^{-1}(\epsilon_1, \epsilon_2)$ there exists a unit vector u_x such that*

$$d^2_- f(x|\xi)(u_x) \le -2\mu_2^{-1} \text{ for all } \xi \in \partial^\pi f(x) \text{ with } \|\xi\| \le 2\sqrt{2(\epsilon_2 - \epsilon_1)/\mu_2}.$$

Then $S := \{x \in X : f(x) \le 0\}$ is nonempty and

$$d_S(x) \le \mu f(x) \text{ for all } x \in f^{-1}(0, \epsilon_2),$$

where $\mu = \mu_1 + (2\mu_2/\epsilon_1)^{1/2}$.

Proof. Since condition (i) implies that

$$\|\xi\| > \mu_1^{-1} \text{ for all } \xi \in \partial^\pi f(x) \text{ and each } x \in f^{-1}(0, \epsilon_1),$$

applying Theorem 2.2 to the function f with $\partial_\omega = \partial^\pi$, we obtain that S is nonempty and

$$d_S(x) \le \mu_1 f(x)_+ \text{ for all } x \in f^{-1}(-\infty, \epsilon_1).$$

This also holds for all $x \in X$ satisfying $f(x) = \epsilon_1$ by the continuity of f and d_S.

Next, by applying Theorem 3.1 to the function $f(\cdot) - \epsilon_1$, we have

$$d_{f^{-1}(-\infty,\epsilon_1]}(x) \le \sqrt{2\mu_2[f(x) - \epsilon_1]} < \sqrt{\frac{2\mu_2}{\epsilon_1}}\, f(x) \text{ for all } x \in f^{-1}(\epsilon_1, \epsilon_2).$$

Thus, by Proposition 3.2, for $\mu = \mu_1 + (2\mu_2/\epsilon_1)^{1/2}$ we have

$$d_S(x) \le \mu f(x)_+ \text{ for all } x \in f^{-1}(0, \epsilon_2). \qquad \square$$

Remark 3.2. Only in a Hilbert space is Theorem 3.3 established, unlike [7, Theorem 3.2], which is given in a Banach space for the case $\epsilon_2 = +\infty$. However, the function f in [7, Theorem 3.2] needs to be not only continuous but also Gâteaux differentiable, while the inequality $d^2_- f(x|\xi)(u_x) \le -2\mu_2^{-1}$ in (ii) is required to hold for each x in $D(\mu_1) \backslash f^{-1}(-\infty, \epsilon_1]$ and for all points in the corresponding interval $(x, x + Tu_x)$ for some $T > 0$. In Theorem 3.3, we do not restrict ϵ_2 to equal $+\infty$ nor require the condition $d^2_- f(x|\xi)(u_x) \le -2\mu_2^{-1}$ to be satisfied in the interval $(x, x + Tu_x)$ for each $x \in f^{-1}(\epsilon_1, \epsilon_2]$.

In what follows, we use Theorem 3.1 to develop sufficient conditions for a system of inequalities, equalities, and an abstract constraint to have error bounds in terms of the second-order subderivatives of the functions involved and certain tangent cones to the abstract constraint set.

We first review some concepts about tangent cone and contingent cone briefly. For a closed subset C in a Banach space X and $x \in C$, the *tangent cone* to C at x, denoted $T_C(x)$, is defined as

$$T_C(x) := \{v \in X : d_C^\circ(x; v) = 0\},$$

and the *contingent* (or the *Bouligand tangent*) *cone* to C at x, denoted $K_C(x)$, is given by

$$K_C(x) := \{v \in X : d_C^-(x; v) = 0\}.$$

It is well known that $v \in T_C(x)$ iff, for every sequence x_n in C converging to x and sequence t_n in $(0, +\infty)$ decreasing to 0, there is a sequence v_n in X converging to v such that $x_n + t_n v_n \in C$ for all n and that $v \in K_C(x)$ iff there exist $v_n \to v$ and $t_n \to 0^+$ such that $x + t_n v_n \in C$. Therefore we have the inclusive relation $T_C(x) \subseteq K_C(x)$.

We also recall that a vector v is *hypertangent* to the set C at the point x in C if there exists $0 < \epsilon$ such that

$$y + tw \in C \quad \text{for all } y \in B(x, \epsilon) \cap C,\ w \in B(v, \epsilon),\ t \in (0, \epsilon).$$

[3, Theorem 2.4.8] states that if the set of hypertangents to the set C at x is nonempty, then it coincides with $\operatorname{int} T_C(x)$, the interior of $T_C(x)$.

The above concepts turn out to be important for us to use Theorem 3.1 to give sufficient conditions for an inequality system with an abstract constraint set to have error bounds.

THEOREM 3.4. *Let X be a Hilbert space, let C be a nonempty closed set in X, and let $f : X \to (-\infty, +\infty]$ be an l.s.c. function. Suppose that, for some $x_0 \in X$, $0 < \delta \le +\infty$, $0 < \mu$, and $0 < \epsilon \le (2\mu)^{-1}(\delta/2)^2$, the set $C \cap B(x_0, \delta/2) \cap f^{-1}(-\infty, \epsilon)$ is nonempty and that, for each $x \in C \cap B(x_0, \delta) \cap f^{-1}(0, \epsilon)$, there exists a unit vector $u_x \in X$ such that*

(i) *u_x is hypertangent to C at x and satisfies*

$$d_-^2 f(x|\xi)(u_x) \le -2\mu^{-1}$$

for each $\xi \in \partial^\pi(f + \psi_C)(x)$ with $\|\xi\| \le \min\{2\sqrt{2\epsilon}\,\mu^{-1/2}, \delta\mu^{-1}\}$; or

(ii) *$u_x \in K_C(x)$ and satisfies*

$$d_+^2 f(x|\xi)(u_x) \le -2\mu^{-1}$$

for all $\xi \in X$ with $\|\xi\| \le \min\{2\sqrt{2\epsilon}\,\mu^{-1/2}, \delta\mu^{-1}\}$ and $\langle \xi, u_x \rangle \le f^+(x; u_x)$.

Then $S := \{x \in C : f(x) \le 0\}$ is nonempty and

$$d_S^2(x) \le 2\mu f(x)_+ \text{ for all } x \in C \cap B\left(x_0, \frac{\delta}{2}\right) \cap f^{-1}(0, \epsilon).$$

Moreover, if $x_0 \in S$, then the condition $0 < \epsilon \le (2\mu)^{-1}(\delta/2)^2$ can be replaced with $0 < \epsilon \le +\infty$.

Proof. Let $x \in C \cap B(x_0, \delta) \cap f^{-1}(0, \epsilon)$. Based on Theorem 3.1, it suffices to show that there exists a unit vector $u_x \in X$ such that

$$d_-^2 (f + \psi_C)(x|\xi)(u_x) \le -2\mu^{-1}$$

for each $\xi \in \partial^\pi(f + \psi_C)(x)$ with $\|\xi\| \le \min\{2\sqrt{2\epsilon}\,\mu^{-1/2}, \delta\mu^{-1}\}$.

Now if u_x is a unit hypertangent vector in (i), then, for each $\xi \in \partial^\pi(f+\psi_C)(x)$ with $\|\xi\| \le \min\{2\sqrt{2\epsilon}\,\mu^{-1/2}, \delta\mu^{-1}\}$, we have sequences $u_n \to u_x$ and $t_n \to 0^+$ such that $x + t_n u_n \in C$ and

$$\begin{aligned} d^2_-(f+\psi_C)(x|\xi)(u_x) &\le \lim_{n\to+\infty} \frac{f(x+t_nu_n) - f(x) - t_n\langle \xi, u_n\rangle}{t_n^2} \\ &= d^2_- f(x|\xi)(u_x) \le -2\mu^{-1}. \end{aligned}$$

If $u_x \in K_C(x)$ is a unit vector in (ii), then there exist sequences $u_n \to u_x$ and $t_n \to 0^+$ such that $x + t_n u_n \in C$. It follows that for each $\xi \in \partial^\pi(f+\psi_C)(x)$ there exists some $M > 0$ such that

$$f(x+t_nu_n) - f(x) \ge t_n\langle \xi, u_n\rangle - Mt_n^2\|u_n\|^2$$

for sufficiently large n. This implies that $\langle \xi, u_x\rangle \le f^+(x; u_x)$ for each $\xi \in \partial^\pi(f + \psi_C)(x)$, that is,

$$\partial^\pi(f+\psi_C)(x) \subseteq \{\xi \in X : \langle \xi, u_x\rangle \le f^+(x; u_x)\}.$$

Thus for each $\xi \in \partial^\pi(f+\psi_C)(x)$ with $\|\xi\| \le \min\{2\sqrt{2\epsilon}\,\mu^{-1/2}, \delta\mu^{-1}\}$ we have

$$\begin{aligned} d^2_-(f+\psi_C)(x|\xi)(u_x) &\le \limsup_{n\to+\infty} \frac{f(x+t_nu_n) - f(x) - t_n\langle \xi, u_n\rangle}{t_n^2} \\ &\le d^2_+ f(x|\xi)(u_x) \le -2\mu^{-1}. \end{aligned}$$

The proof is therefore complete. □

Remark 3.3. From the above proof we see that Theorem 3.4 is a direct result of Theorem 3.1. Note that if x is an interior point of a closed subset C of X, then the set of hypertangents to the set C at x is just X. In particular, when $C = X$, each unit vector u_x is hypertangent to C at $x \in X$. In this case Theorem 3.4 reduces to Theorem 3.1. So they are in fact equivalent.

To apply Theorem 3.1 to a system of inequalities, we first give a result about the proximal subdifferential of the pointwise maxima function of several functions.

PROPOSITION 3.5. *Let $f_i : X \to R$ be Lipschitz near x for each $i \in I$. Denote*

$$f(x) = \max\{f_i(x) : i \in I\} \quad \textit{and} \quad I(x) := \{i \in I : f_i(x) = f(x)\} \textit{ for } x \in X.$$

Suppose that $\partial^\pi f_i(x) = \partial^\circ f_i(x)$ for each $i \in I(x)$. Then

$$\partial^\pi f(x) = co\{\partial^\pi f_i(x) : i \in I(x)\} = \partial^\circ f(x),$$

where co A is the convex hull of a set A.

Proof. Since $\partial^\pi f_i(x) = \partial^\circ f_i(x)$ for each $i \in I(x)$, by [3, Proposition 2.3.12], we have

$$\partial^\pi f(x) \subseteq \partial^\circ f(x) \subseteq co\{\partial^\circ f_i(x) : i \in I(x)\} = co\{\partial^\pi f_i(x) : i \in I(x)\}.$$

So it suffices to show that $co\{\partial^\pi f_i(x) : i \in I(x)\} \subseteq \partial^\pi f(x)$.

For any fixed $i \in I(x)$ and $\xi_i \in \partial^\pi f_i(x)$, there exist $M > 0$ and $\delta > 0$ such that

$$f_i(y) - f_i(x) + M\|y - x\|^2 \ge \langle \xi_i, y - x\rangle \text{ for all } y \in B(x, \delta).$$

It follows that

$$f(y) - f(x) + M\|y - x\|^2 \geq \langle \xi_i, y - x\rangle \text{ for all } y \in B(x, \delta),$$

which implies that $\xi_i \in \partial^\pi f(x)$. Since i and ξ_i are arbitrary, $\partial^\pi f_i(x) \subseteq \partial^\pi f(x)$ for each $i \in I(x)$. In addition, $\partial^\pi f(x)$ is convex, so for any $\lambda_i \geq 0$ with $\sum_{i\in I(x)} \lambda_i = 1$,

$$\sum_{i\in I(x)} \lambda_i \xi_i \in \partial^\pi f(x).$$

This is what we need to prove. □

THEOREM 3.6. *Let X be a Hilbert space, and let $f_i : X \to R$ be an l.s.c. function for each $i \in I$. Denote*

$$f(x) = \max\{f_i(x) : i \in I\} \quad \textit{and} \quad I(x) := \{i \in I : f_i(x) = f(x)\} \textit{ for } x \in X.$$

Suppose that, for some $x_0 \in X$, $0 < \delta \leq +\infty$, $0 < \mu$, and $0 < \epsilon \leq (2\mu)^{-1}(\delta/2)^2$, the set $B(x_0, \delta/2) \cap f^{-1}(-\infty, \epsilon)$ is nonempty and that, for each $x \in B(x_0, \delta) \cap f^{-1}(0, \epsilon)$ and each $i \in I(x)$,

(i) *f_i is Lipschitz near x and $\partial^\pi f_i(x) = \partial^\circ f_i(x)$; and*

(ii) *there exists a unit vector u_x such that $d_L^2 f_j(x|\xi_j)(u_x) \leq -2\mu^{-1}$ and*

$$\lim_{t\to 0^+} \frac{f_i(x + tu_x) - [f_j(x + tu_x) - t\langle \xi_j, u_x\rangle] - t\langle \xi_k, u_x\rangle}{t^2} = 0$$

for some $j \in I(x)$ and $\xi_j \in \partial^\pi f_j(x)$, each $i \in I(x)$ and $\xi_i \in \partial^\pi f_i(x)$, and each $k \in I(x)$ and $\xi_k \in \partial^\pi f_k(x)$.

Then $S := \{x \in X : f(x) \leq 0\}$ is nonempty and

$$d_S^2(x) \leq 2\mu f(x)_+ \textit{ for all } x \in B\left(x_0, \frac{\delta}{2}\right) \cap f^{-1}(0, \epsilon).$$

Moreover, if $x_0 \in S$, then the condition $0 < \epsilon \leq (2\mu)^{-1}(\delta/2)^2$ can be replaced with $0 < \epsilon \leq +\infty$.

Proof. Let $x_0 \in X$, $0 < \delta \leq +\infty$, $0 < \mu$ and let the set $B(x_0, \delta/2) \cap f^{-1}(-\infty, \epsilon)$ be nonempty for some $0 < \epsilon < (2\mu)^{-1}(\delta/2)^2$. If, for $x \in B(x_0, \delta) \cap f^{-1}(0, \epsilon)$, $\partial^\pi f_i(x) = \partial^\circ f_i(x)$ for each $i \in I(x)$, then, for $\xi \in \partial^\pi f(x)$, by Proposition 3.5, $\xi = \sum_{i\in I(x)} \lambda_i \xi_i$ for some $\lambda_i \geq 0$ and $\xi_i \in \partial^\pi f_i(x)$ with $i \in I(x)$ and $\sum_{i\in I(x)} \lambda_i = 1$.

If u_x is the unit vector stated in the assumption, then

$$\begin{aligned}
d_L^2 f(x|\xi)(u_x) &= \liminf_{t\to 0^+} \frac{\max\{f_i(x + tu_x) : i \in I(x)\} - f(x) - t\langle \xi, u_x\rangle}{t^2} \\
&= \liminf_{t\to 0^+} \frac{\max\{f_i(x + tu_x) : i \in I(x)\} - f(x) - t\sum_{i\in I(x)} \lambda_i \langle \xi_i, u_x\rangle}{t^2} \\
&= \liminf_{t\to 0^+} \sum_{i\in I(x)} \lambda_i \frac{\max\{f_i(x + tu_x) : i \in I(x)\} - f(x) - t\langle \xi_i, u_x\rangle}{t^2} \\
&\leq \liminf_{t\to 0^+} \frac{f_j(x + tu_x) - f_j(x) - t\langle \xi_j, u_x\rangle}{t^2} \\
&\quad + \lim_{t\to 0^+} \sum_{i\in I(x)} \frac{|f_i(x + tu_x) - [f_j(x + tu_x) - t\langle \xi_j, u_x\rangle] - t\langle \xi_k, u_x\rangle|}{t^2} \\
&= d_L^2 f_j(x|\xi_j)(u_x) \leq -2\mu^{-1}.
\end{aligned}$$

Thus the conclusion follows from Theorem 3.1. ∎

Remark 3.4. From the proof of Theorem 3.6 we see that condition (i) can be replaced with the condition that f_i be continuous at x and $\partial^\pi f(x) = co\{\partial^\pi f_i(x) : i \in I(x)\}$.

THEOREM 3.7. *Let X be a Hilbert space and, for each $i \in I$, let $f_i : X \to R$ be C^1 and satisfy $\partial^\pi f_i(x) = \partial^\circ f_i(x)$ for $x \in X$. Denote*

$$f(x) = \max\{f_i(x) : i \in I\} \quad \text{and} \quad I(x) := \{i \in I : f_i(x) = f(x)\} \text{ for } x \in X.$$

Suppose that, for some $x_0 \in X$, $0 < \delta \le +\infty$, $0 < \mu$, and $0 < \epsilon \le (2\mu)^{-1}(\delta/2)^2$, the set $B(x_0, \delta/2) \cap f^{-1}(-\infty, \epsilon)$ is nonempty and that, for each $x \in B(x_0, \delta) \cap f^{-1}(0, \epsilon)$, there exists a unit vector u_x such that

$$\langle \nabla f_i(x), u_x\rangle = \langle \nabla f_j(x), u_x\rangle \quad \text{and} \quad d_+^2 f_i(x|\nabla f_i(x))(u_x) \le -2\mu^{-1} \text{ for all } i, j \in I(x).$$

Then $S := \{x \in X : f(x) \le 0\}$ is nonempty and

$$d_S^2(x) \le 2\mu f(x)_+ \text{ for all } x \in B\left(x_0, \frac{\delta}{2}\right) \cap f^{-1}(0, \epsilon).$$

Moreover, if $x_0 \in S$, then the condition $0 < \epsilon \le (2\mu)^{-1}(\delta/2)^2$ can be replaced with $0 < \epsilon \le +\infty$.

Proof. Let x_0, δ, μ, and ϵ be given as in the assumption. For each $x \in B(x_0, \delta) \cap f^{-1}(0, \epsilon)$ and $i \in I(x)$, f_i is C^1 and $\partial^\pi f_i(x) = \{\nabla f_i(x)\} = \partial^\circ f_i(x)$, so for $\xi \in \partial^\pi f(x)$, by Proposition 3.5, $\xi = \sum_{i\in I(x)} \lambda_i \nabla f_i(x)$ for some $\lambda_i \ge 0$ with $i \in I(x)$ and $\sum_{i\in I(x)} \lambda_i = 1$.

If u_x is the vector in the assumption, then there exists $t_n \to 0$ such that

$$\begin{aligned}
d_L^2 f(x|\xi)(u_x) &= \lim_{n\to+\infty} \frac{\max\{f_i(x + t_n u_x) : i \in I(x)\} - f(x) - t_n\langle \xi, u_x\rangle}{t_n^2} \\
&= \lim_{n\to+\infty} \frac{\max\{f_i(x + t_n u_x) : i \in I(x)\} - f(x) - t_n \sum_{i\in I(x)} \lambda_i \langle \nabla f_i(x), u_x\rangle}{t_n^2} \\
&= \lim_{n\to+\infty} \sum_{i\in I(x)} \lambda_i \frac{\max\{f_i(x + t_n u_x) : i \in I(x)\} - f(x) - t_n\langle \nabla f_i(x), u_x\rangle}{t_n^2} \\
&\le \limsup_{n\to+\infty} \max\left\{ \frac{f_i(x + t_n u_x) - f_i(x) - t_n\langle \nabla f_i(x), u_x\rangle}{t_n^2} : i \in I(x)\right\} \\
&\le \max\{d_+^2 f_i(x|\nabla f_i(x))(u_x) : i \in I(x)\} \le -2\mu^{-1};
\end{aligned}$$

that is, we have $d_L^2 f(x|\xi)(u_x) \le -2\mu^{-1}$. Therefore, upon using Theorem 3.1 to f, the conclusion follows. ∎

We now consider a system of quadratic inequalities

$$S = \{x \in R^n : f_1(x) \le 0, \ldots, f_m(x) \le 0\},$$

where $f_i(x) = x^t Q_i x + b_i^t x + c_i$, Q_i is a real $n \times n$ symmetric matrix, $b_i \in R^n$, and $c_i \in R$ for each $i \in I$ with x^t denoting the transpose of x. For the convex quadratic system, i.e., when each Q_i is positive semidefinite, Luo and Luo [12] and Wang and Pang [19] show that the nonemptiness of S implies the existence of a positive integer $d \le n + 1$ and a positive scalar μ such that

$$d_S(x) \le \mu \left[f(x)_+ + f(x)_+^{\frac{1}{2^d}} \right] \text{ for all } x \in R^n, \tag{15}$$

where $f(x) = \max\{f_i(x) : i \in I\}$. Furthermore, if S contains an interior point, then $d = 0$.

For a nonconvex quadratic system, there are very few existing error bound results. For the special case of a single quadratic function, Luo and Sturm [13] show that (15) holds with d equal to 1; Ng and Zheng [15] further prove that for a single quadratic function, global error bounds with either exponents 1 or 1/2 hold, and they also classify the cases for exponents being 1 or 1/2. In the following theorem we apply Theorem 3.7 to derive a sufficient condition for a nonconvex quadratic system. It is worth pointing out that even for the case of a single quadratic system our theorem offers something new since an error bound is explicitly given in terms of the eigenvalues of matrices.

COROLLARY 3.8. *For each $i \in I$, let*

$$f_i(x) = x^t Q_i x + b_i^t x + c_i \quad \text{for } x \in R^n,$$

where Q_i is a real $n \times n$ symmetric matrix, $b_i \in R^n$, and $c_i \in R$. Denote

$$f(x) = \max\{f_i(x) : i \in I\} \quad \text{and} \quad I(x) := \{i \in I : f_i(x) = f(x)\} \text{ for } x \in R^n.$$

Suppose that for each $x \in f^{-1}(0, +\infty)$ and for each Q_i there exists a negative eigenvalue λ_i with a common eigenvector u and $\langle 2Q_i x + b_i, u\rangle = \langle 2Q_j x + b_j, u\rangle$ for all $i, j \in I(x)$. Then $S := \{x \in R^n : f(x) \le 0\}$ is nonempty and

$$d_S^2(x) \le -\frac{4}{\lambda} f(x)_+ \text{ for all } x \in R^n,$$

where $\lambda = \max\{\lambda_i : i \in I(x)\}$. In particular, if $I = \{1\}$ and λ_1 and λ_2 are the smallest eigenvalue and the largest eigenvalue of Q_1 with $\lambda_1 < 0 < \lambda_2$, then $S := \{x \in R^n : f_1(x) = 0\}$ is nonempty and

$$d_S^2(x) \le -\frac{4}{\lambda} |f_1(x)| \text{ for all } x \in R^n,$$

where $\lambda = \max\{\lambda_1, -\lambda_2\}$.

Proof. Let u be a common eigenvector of Q_i corresponding to an eigenvalue $\lambda_i < 0$ for all $i \in I(x)$. Then we have

$$f_i(\alpha u) = \lambda_i \alpha^2 u^t Q_i u + \alpha b_i^t u + c_i < 0$$

for sufficiently large positive scalar α. This implies that $S := \{x \in R^n : f(x) \le 0\}$ is nonempty.

Denote $u_x := \frac{u}{\|u\|}$. Then

$$\langle \nabla f_i(x), u_x\rangle = \langle 2Q_i x + b_i, u_x\rangle = \langle 2Q_j x + b_j, u_x\rangle = \langle \nabla f_j(x), u_x\rangle \text{ and}$$

$$d_+^2 f_i(x|\nabla f_i(x))(u_x) = u_x^t Q_i u_x = \lambda_i \le \lambda.$$

Thus, by Theorem 3.7,

$$d_S^2(x) \le -\frac{4}{\lambda} f(x)_+ \text{ for all } x \in R^n.$$

Now if $I = \{1\}$, we consider

$$f(x) := \max\{f_1(x), -f_1(x)\} = \begin{cases} f_1(x) & \text{if } f_1(x) \ge 0, \\ -f_1(x) & \text{if } f_1(x) < 0. \end{cases}$$

It is easy to see that $S := \{x \in R^n : f_1(x) = 0\}$ is nonempty and that $\nabla f(x) = \nabla f_1(x)$ for $x \in f^{-1}(0, +\infty)$ and $\nabla f(x) = -\nabla f_1(x)$ for $x \in f^{-1}(-\infty, 0)$. If u_1 and u_2 are unit eigenvectors corresponding to λ_1 and λ_2, respectively, then, for each $x \in f_1^{-1}(0, +\infty)$,

$$d_+^2 f(x|\nabla f(x))(u_1) = u_1^t Q_1 u_1 = \lambda_1 \leq \lambda$$

and, for each $x \in f_1^{-1}(-\infty, 0)$,

$$d_+^2 f(x|\nabla f(x))(u_2) = -u_2^t Q_1 u_2 = -\lambda_2 \leq \lambda.$$

Therefore it follows from Theorem 3.7 that

$$d_S^2(x) \leq -\frac{4}{\lambda}|f_1(x)| \text{ for all } x \in R^n. \qquad \square$$

Example 3.1. For $x \in R^2$, define

$$f_1(x) = x^t Q_1 x + b_1^t x + 1 \quad \text{and} \quad f_2(x) = x^t Q_2 x + b_2^t x,$$

where

$$Q_1 = \begin{pmatrix} -1 & 0 \\ 0 & 3 \end{pmatrix},\ b_1 = \begin{pmatrix} 3 \\ 0 \end{pmatrix},\ Q_2 = \begin{pmatrix} -1 & 0 \\ 0 & -4 \end{pmatrix},\ \text{and } b_2 = \begin{pmatrix} 3 \\ -1 \end{pmatrix}.$$

It is easy to see that $\lambda = -1$ is a common eigenvalue of Q_1 and Q_2 with a common eigenvector $u = (1, 0)^t$ and that $\langle 2Q_1 x + b_1, u\rangle = -2x_1 + 3 = \langle 2Q_2 x + b_2, u\rangle$. Therefore, by Corollary 3.8, $S := \{x \in R^2 : f_1(x) \leq 0,\ f_2(x) \leq 0\}$ is nonempty and

$$d_S^2(x) \leq 4 \max\{f_1(x), f_2(x)\}_+ \text{ for all } x \in R^2.$$

4. Sufficient conditions in lower Dini derivatives. We note that in a general Banach space the lower Dini subdifferential is not always a ∂_ω-subdifferential (see [10]). Thus Theorem 2.2 is not applicable to the lower Dini subdifferential in a general Banach space. However, in this case the lower Dini derivative $f^-(x; \cdot)$ of function f at x turns out to be more convenient for us to present a sufficient condition for error bounds to exist. For this we first prove one of the main results in this section.

THEOREM 4.1. *Let (X, d) be a metric space and let $f : X \to (-\infty, +\infty]$ be an l.s.c. function. For some $0 < \epsilon \leq +\infty$ and $0 < \mu < +\infty$ we consider the following statements:*

(i) *If the set $f^{-1}(-\infty, \epsilon)$ is nonempty and for each $x \in f^{-1}(0, \epsilon)$ there exists a point $y \in f^{-1}[0, \epsilon)$ such that*

$$0 < d(x, y) \leq \mu[f(x) - f(y)],$$

then $S := \{x \in X : f(x) \leq 0\}$ is nonempty and

$$d_S(x) \leq \mu f(x)_+ \text{ for all } x \in f^{-1}(-\infty, \epsilon).$$

(ii) *If for some $x_0 \in X$ and $0 < \delta < +\infty$ the set $B(x_0, \delta) \cap f^{-1}(-\infty, \epsilon)$ is nonempty and for some $0 < \rho < 1$ and each $x \in B(x_0, \delta) \cap f^{-1}(0, \epsilon)$ there exists a point $y \in f^{-1}[0, \epsilon)$ such that*

$$d(y, x_0) \leq \max\{\rho\delta, d(x, x_0)\} \quad \text{and} \quad 0 < d(x, y) \leq \mu[f(x) - f(y)],$$

then $S := \{x \in X : f(x) \leq 0\}$ is nonempty and

$$d_S(x) \leq \mu f(x)_+ \text{ for all } x \in B(x_0, \delta) \cap f^{-1}(-\infty, \epsilon).$$

(iii) *If for some nonempty closed subset C of X the set $C\cap f^{-1}(-\infty,\epsilon)$ is nonempty and for some $0<\mu<+\infty$ and each $x\in C\cap f^{-1}(0,\epsilon)$ there exists a sequence $\{x_n\}\subseteq C\backslash\{x\}$ such that*

$$\lim_{n\to+\infty}\frac{f(x_n)_+ - f(x)_+}{\|x_n-x\|}\le -\mu^{-1}, \tag{16}$$

then $S:=\{x\in C: f(x)\le 0\}$ is nonempty and

$$d_S(x)\le \mu f(x)_+ \text{ for all } x\in C\cap f^{-1}(-\infty,\epsilon).$$

In a metric space X, (i) $\Rightarrow$ (ii)*; if X is also complete, then both* (i) *and* (ii) *hold. In a normed space X,* (i) $\Leftrightarrow$ (iii)*; hence* (iii) *holds in a Banach space X.*

Proof. We first prove that (i) implies (ii) in a metric space X.

Let $x_0\in X$, $0<\delta<+\infty$, and $0\le\rho<1$. For each $m\in N$ (the set of natural numbers) such that $B(x_0,\rho\delta)\subseteq \overline{B}_m(x_0,\delta):=\overline{B}(x_0,(1-1/m)\delta)$ and each $x\in\overline{B}_m(x_0,\delta)\cap f^{-1}(0,\epsilon)$ there exists y with the properties stated in (ii) such that $y\in\overline{B}_m(x_0,\delta)$ since

$$d(y,x_0)\le\max\{\rho\delta,d(x,x_0)\}\le\left(1-\frac{1}{m}\right)\delta.$$

Upon applying (i) to the function $f+\psi_{\overline{B}_m(x_0,\delta)}$, we obtain that $S_m:=\overline{B}_m(x_0,\delta)\cap f^{-1}(-\infty,0]$ is nonempty and

$$d_{S_m}(x)\le\mu f(x)_+ \text{ for all } x\in\overline{B}_m(x_0,\delta)\cap f^{-1}(-\infty,\epsilon).$$

This implies that (ii) holds since for each $x\in B(x_0,\delta)\cap f^{-1}(-\infty,\epsilon)$ there exists an m stated above such that $x\in\overline{B}_m(x_0,\delta)\cap f^{-1}(-\infty,\epsilon)$ and $d_S(x)\le d_{S_m}(x)$.

Now it is known from [22, Theorem 3] that (i) holds in a complete metric space, so (ii) also holds in a complete metric space.

Next, we prove that (i) and (iii) are equivalent in a normed space X.

Suppose that (i) is true. To prove (iii) to be also true, it suffices to show that for any $\lambda>1$ and $x\in C\cap f^{-1}(0,\epsilon)$ there exists a point $y\in C\cap(f_+)^{-1}[0,\epsilon)$ such that

$$0<\|x-y\|\le\lambda\mu[f(x)_+-f(y)_+].$$

Let $\lambda>1$ be fixed. For each $x\in C\cap f^{-1}(0,\epsilon)$, by assumption, there exists a sequence $\{x_n\}\subseteq C\backslash\{x\}$ satisfying (16). Hence for sufficiently large n we have

$$\frac{f(x_n)_+-f(x)_+}{\|x_n-x\|}\le-(\lambda\mu)^{-1},$$

that is,

$$0<\|x_n-x\|\le\lambda\mu[f(x)_+-f(x_n)_+].$$

So we can take $y=x_n$ for any such an n.

Now, to prove (iii) $\Rightarrow$ (i), we suppose that for each $x\in f^{-1}(0,\epsilon)$ there exists a point $y\in f^{-1}[0,\epsilon)$ such that

$$0<\|x-y\|\le\mu[f(x)-f(y)].$$

By taking $x_n = y$ we have

$$\lim_{n\to+\infty} \frac{f(x_n)_+ - f(x)_+}{\|x_n - x\|} = \lim_{n\to+\infty} \frac{f(y) - f(x)}{\|y - x\|} \le -\mu^{-1}.$$

It follows from statement (iii) with $C = X$ that S is nonempty and satisfies

$$d_S(x) \le \mu f(x)_+ \text{ for all } x \in f^{-1}(-\infty, \epsilon).$$

Therefore (i) is valid.

As we indicated above, (i) holds in a complete metric space, so (iii) holds in a Banach space. □

Based on Theorem 4.1, we present some sufficient conditions in terms of Dini derivatives of involved functions and tangent cones to a set as below.

THEOREM 4.2. *Let X be a Banach space and let C be a nonempty closed subset in X. Suppose that $f : X \to (-\infty, +\infty]$ is an l.s.c. function and that for some $0 < \epsilon \le +\infty$ the set $C \cap f^{-1}(-\infty, \epsilon)$ is nonempty. If for some $0 < \mu$ and each $x \in C \cap f^{-1}(0, \epsilon)$ there exists a unit hypertangent vector u_x to C at x such that $f^-(x; u_x) \le -\mu^{-1}$ or $u_x \in K_C(x)$ such that $f^+(x; u_x) \le -\mu^{-1}$, then $S := \{x \in C : f(x) \le 0\}$ is nonempty and*

$$d_S(x) \le \mu f(x)_+ \text{ for all } x \in C \cap f^{-1}(-\infty, \epsilon).$$

Proof. For some $0 < \epsilon \le +\infty$, let $x \in C \cap f^{-1}(0, \epsilon)$. If u_x is a unit hypertangent vector to C at x satisfying $f^-(x; u_x) \le -\mu^{-1}$, then there exist sequences $u_n \to u_x$ and $t_n \to 0^+$ such that

$$\lim_{n\to+\infty} \frac{f(x + t_n u_n) - f(x)}{t_n} = f^-(x; u_x) \le -\mu^{-1}$$

and $x + t_n u_n \in C$. If $u_x \in K_C(x)$ and $f^+(x; u_x) \le -\mu^{-1}$, then there exist sequences $u_n \to u_x$ and $t_n \to 0^+$ such that $x + t_n u_n \in C$, for which we have

$$\liminf_{n\to+\infty} \frac{f(x + t_n u_n) - f(x)}{t_n} \le f^+(x; u_x) \le -\mu^{-1}.$$

Now for the above sequences $u_n \to u_x$ and $t_n \to 0^+$ we have $x_n := x + t_n u_n \in C \backslash \{x\}$ and

$$\liminf_{n\to+\infty} \frac{f(x_n)_+ - f(x)_+}{\|x_n - x\|} = \liminf_{n\to+\infty} \frac{f(x_n) - f(x)}{\|x_n - x\|} \le -\mu^{-1}.$$

Hence there exists a subsequence $\{x_{n_k}\}$ satisfying the condition (iii) in Theorem 4.1. Therefore the conclusion holds. □

Similar to Theorem 3.4, Theorem 4.2 has the following equivalent result.

THEOREM 4.3. *Let X be a Banach space, and let $f : X \to (-\infty, +\infty]$ be an l.s.c. function. Suppose that for some $0 < \epsilon \le +\infty$ the set $f^{-1}(-\infty, \epsilon)$ is nonempty and that for some $0 < \mu$ and each $x \in f^{-1}(0, \epsilon)$ there exists a unit vector u_x in X such that $f^-(x; u_x) \le -\mu^{-1}$. Then $S := \{x \in X : f(x) \le 0\}$ is nonempty and*

$$d_S(x) \le \mu f(x)_+ \text{ for all } x \in f^{-1}(-\infty, \epsilon).$$

In what follows we use Theorem 4.2 to establish error bounds for a system containing functions f and $g_i : X \to (-\infty, +\infty]$ $(i \in I)$ for which we denote

$$g(x) := \max\{g_i(x) : i \in I\} \quad \text{and} \quad I(x) := \{i \in I : g_i(x) = g(x)\} \text{ for } x \in X.$$

THEOREM 4.4. *Let C be a nonempty closed subset in a Banach space X, let $f : X \to (-\infty, +\infty]$ be l.s.c., and let $g_i : X \to (-\infty, +\infty)$ be locally Lipschitz for each $i \in I$. Denote*

$$C_0 := \{x \in C : g_i(x) \le 0 \text{ for each } i \in I\}.$$

Suppose that for some $0 < \epsilon \le +\infty$ the set $C_0 \cap f^{-1}(-\infty, \epsilon)$ is nonempty. If, for some $0 < \mu$ and each $x \in C_0 \cap f^{-1}(0, \epsilon)$, there exists a unit vector $u_x \in K_C(x)$ such that $f^+(x; u_x) \le -\mu^{-1}$ and, for each $x \in C_0 \cap f^{-1}(0, \epsilon)$ with $g(x) = 0$ and each $i \in I(x)$, $g_i^+(x; u_x) < 0$, then $S := \{x \in C_0 : f(x) \le 0\}$ is nonempty and

$$d_S(x) \le \mu f(x)_+ \text{ for all } x \in C_0 \cap f^{-1}(-\infty, \epsilon).$$

Proof. Let $x \in C_0 \cap f^{-1}(0, \epsilon)$ and let $u_x \in K_C(x)$ be the unit vector in the assumption. Then there exist sequences $u_n \to u_x$ and $t_n \to 0^+$ such that $x + t_n u_n \in C$. According to Theorem 4.2 , it suffices to show that $u_x \in K_{C_0}(x)$.

If $g(x) < 0$, then, by the continuity of g, $g(x + t_n u_n) \le 0$ when n is large enough. This implies that $x + t_n u_n \in C_0$ when n is large enough and hence $u_x \in K_{C_0}(x)$.

If $g(x) = 0$, then, by the definition of $g_i^+(x; u_x)$, there are $\delta > 0$ and $\epsilon > 0$ such that for each $i \in I(x)$ and all $t \in (0, \delta)$ we have

$$g_i(x + tu_x) - g_i(x) \le -\epsilon t.$$

Since g_i is Lipschitz near x, there exists a constant L_i such that

$$\begin{aligned} g_i(x + t_n u_n) &\le g_i(x + t_n u_x) + L_i t_n \|u_n - u_x\| \\ &\le g_i(x) + t_n(-\epsilon + L_i \|u_n - u_x\|) \le 0 \end{aligned}$$

for sufficiently large n. It follows that $x + t_n u_n \in C_0$ when n is large enough. Thus u_x belongs to $K_{C_0}(x)$. □

PROPOSITION 4.5. *Let x be a point in a closed subset C of a Banach space X, let $f_i : X \to (-\infty, +\infty)$ be Lipschitz near x, let $g_i : X \to (-\infty, +\infty]$ be Fréchet differentiable at x for each $i \in I$, and let $h_j : X \to (-\infty, +\infty)$ be continuous in a neighborhood of x and Fréchet differentiable at x for each $j \in J$ with the Fréchet derivative $\nabla h(x) = (\nabla h_1(x), \ldots, \nabla h_n(x))^t$ being surjective. Denote*

$$C_1 := \{x \in C : (f_i + g_i)(x) \le 0 \text{ for } i \in I \text{ and } h_j(x) = 0 \text{ for } j \in J\} \text{ and}$$

$$I(x) := \{i \in I : (f_i + g_i)(x) = 0\}.$$

Suppose that $x \in C_1$ and there exists $v^* \in X$ such that $f_i^\circ(x; v^*) + g_i'(x; v^*) < 0$ for each $i \in I(x)$ and $h_j'(x; v^*) = 0$ for each $j \in J$. If the set of hypertangents to the set C at x is nonempty, then

$$\{v \in \operatorname{int} T_C(x) : f_i^+(x; v) + g_i'(x; v) \le 0,\ i \in I(x);\ h_j'(x; v) = 0,\ j \in J\} \subseteq K_{C_1}(x).$$

Proof. First, for $v \in \operatorname{int} T_C(x)$ satisfying $f_i^+(x; v) + g_i'(x; v) < 0$ for each $i \in I(x)$ and $h_j'(x; v) = 0$ for each $j \in J$, we prove that $v \in K_{C_1}(x)$.

Since $\nabla h(x)$ is surjective, by the correction theorem of Halkin [6, Theorem F] and its proof, there exist a neighborhood U of x and a continuous mapping ξ from U into X such that $\xi(x) = 0$, $\nabla \xi(x) = 0$, and

$$h_j(y + \xi(y)) = \langle \nabla h_j(x), y - x \rangle \text{ for all } y \in U \text{ and each } j \in J.$$

By taking $y = x + sv$ we have, for $t > 0$ small enough and all $s \in (0, t)$,

$$h_j(x + sv + \xi(x + sv)) = \langle \nabla h_j(x), sv \rangle = 0 \text{ for each } j \in J.$$

Note that $\xi(x) = 0$ and $\nabla \xi(x) = 0$, so $\xi(x + tv)/t \to 0$ as $t \to 0$. By the inequality $f_i^+(x; v) + g_i'(x; v) < 0$, we can take $\epsilon > 0$ and $t > 0$ small enough such that

$$(f_i + g_i)(x + sv + \xi(x + sv)) \le (f_i + g_i)(x) - \epsilon s = -\epsilon s$$

for all $s \in (0, t)$ and each $i \in I(x)$. Also, if $v \in \operatorname{int} T_C(x)$, then v is hypertangent to C at x. Hence

$$x + sv + \xi(x + sv) = x + s\left[v + \frac{\xi(x + sv)}{s}\right] \in C \text{ for all } s \in (0, t)$$

when $t > 0$ is small enough. This implies that $v \in K_{C_1}(x)$.

Now, if $v^* \in X$ satisfies $f_i^\circ(x; v^*) + g_i'(x; v^*) < 0$ for each $i \in I(x)$ and $h_j'(x; v^*) = 0$ for each $j \in J$, then, for $v \in \operatorname{int} T_C(x)$ with $f_i^+(x; v) + g_i'(x; v) \le 0$ for each $i \in I(x)$ and $h_j'(x; v) = 0$ for each $j \in J$, we can take $t > 0$ small enough such that, for all $s \in (0, t)$, $v + sv^* \in \operatorname{int} T_C(x)$ and

$$\begin{aligned} f_i^+(x; v + sv^*) &+ g_i'(x; v + sv^*) \\ &\le f_i^+(x; v) + g_i'(x; v + sv^*) + \sup_{u \in X}[f_i^+(x; u + sv^*) - f_i^+(x; u)] \\ &\le f_i^+(x; v) + g_i'(x; v) + s[f_i^\circ(x; v^*) + g_i'(x; v^*)] < 0 \end{aligned}$$

for each $i \in I(x)$ and

$$h_j'(x; v + sv^*) = \langle \nabla h_j(x), v + sv^* \rangle = 0$$

for each $j \in J$. By the conclusion of the above paragraph, we have $v + sv^* \in K_{C_1}(x)$ for all $s > 0$ small enough. This implies that $v \in K_{C_1}(x)$ since $K_{C_1}(x)$ is closed. □

Combining Theorem 4.2 with Proposition 4.5, we obtain the following result.

THEOREM 4.6. *Let C be a nonempty closed subset in a Banach space X, let $f : X \to (-\infty, +\infty]$ be l.s.c., let $f_i : X \to (-\infty, +\infty)$ be locally Lipschitz and $g_i : X \to (-\infty, +\infty]$ Fréchet differentiable on C for each $i \in I$, and let $h_j : X \to (-\infty, +\infty]$ be continuous on C for each $j \in J$. Denote*

$$C_1 := \{x \in C : (f_i + g_i)(x) \le 0 \text{ for } i \in I \text{ and } h_j(x) = 0 \text{ for } j \in J\} \text{ and}$$

$$I(x) := \{i \in I : (f_i + g_i)(x) = 0\} \text{ for } x \in C_1.$$

Suppose that for some $0 < \epsilon \le +\infty$ the set $C_1 \cap f^{-1}(-\infty, \epsilon)$ is nonempty, that, for each $x \in C_1 \cap f^{-1}(0, \epsilon)$, h_j is Fréchet differentiable at x for each $j \in J$ with the Fréchet derivative $\nabla h(x) = (\nabla h_1(x), \ldots, \nabla h_n(x))^t$ being surjective and there exists $v_x^* \in X$ such that $f_i^\circ(x; v_x^*) + g_i(x; v_x^*) < 0$ for each $i \in I(x)$ and $h_j'(x; v_x^*) = 0$ for each $j \in J$, and that there exists a unit hypertangent vector u_x to the set C at x such that $f_i^+(x; u_x) + g_i'(x; u_x) \le 0$ for each $i \in I(x)$, $h_j'(x; u_x) = 0$ for each $j \in J$ and $f^+(x; u_x) \le -\mu^{-1}$ for some $0 < \mu$ independent of x. Then $S := \{x \in C_1 : f(x) \le 0\}$ is nonempty and

$$d_S(x) \le \mu f(x)_+ \text{ for all } x \in C_1 \cap f^{-1}(-\infty, \epsilon).$$

In what follows, we consider an inequality system determined by several inequalities.

THEOREM 4.7. *Let C be a nonempty closed subset in a Banach space X and let $f_i : X \to R$ be continuous for each $i \in I$. Denote*

$$f(x) = \max\{f_i(x) : i \in I\} \quad \text{and} \quad I(x) := \{i \in I : f_i(x) = f(x)\} \text{ for } x \in X.$$

Suppose that, for some $0 < \epsilon \le +\infty$, the set $C \cap f^{-1}(-\infty, \epsilon)$ is nonempty and that, for some $0 < \mu$, each $x \in f^{-1}(0, \epsilon)$, and $i \in I(x)$, there exists a unit vector u_x such that

(i) *u_x is hypertangent to C at x, $f_j^-(x; u_x) \le -\mu^{-1}$ for some $j \in I(x)$ and*

$$\lim_{\substack{v \to u_x \\ t \to 0^+}} \frac{f_i(x + tv) - f_j(x + tv)}{t} = 0 \text{ for each } i \in I(x); \text{ or}$$

(ii) *$u_x \in K_C(x)$ and $f_i^+(x; u_x) \le -\mu^{-1}$ for each $i \in I(x)$.*

Then $S := \{x \in C : f(x) \le 0\}$ is nonempty and

$$d_S(x) \le \mu f(x)_+ \text{ for all } x \in C \cap f^{-1}(0, \epsilon).$$

Proof. Let $0 < \mu$, $0 < \epsilon \le +\infty$, and $x \in C \cap f^{-1}(0, \epsilon)$. If u_x is a unit vector satisfying (i), then

$$\begin{aligned} f^-(x; u_x) &= \liminf_{\substack{v \to u_x \\ t \to 0^+}} \frac{\max\{f_i(x + tv) : i \in I(x)\} - f(x)}{t} \\ &\le \liminf_{\substack{v \to u_x \\ t \to 0^+}} \frac{f_j(x + tv) - f_j(x)}{t} + \lim_{\substack{v \to u_x \\ t \to 0^+}} \sum_{i \in I(x)} \frac{|f_i(x + tv) - f_j(x + tv)|}{t} \\ &= f_j^-(x; u_x) \le -\mu^{-1}, \end{aligned}$$

where the first equality is obtained by the continuity of f_i at x for each $i \in I$.

If u_x satisfies (ii), then there exist $u_n \to u_x$ and $t_n \to 0^+$ such that

$$\begin{aligned} f^+(x; u_x) &= \lim_{n \to +\infty} \frac{f(x + t_n u_n) - f(x)}{t_n} \\ &= \lim_{n \to +\infty} \max\left\{ \frac{f_i(x + t_n u_n) - f_i(x)}{t_n} : i \in I(x) \right\} \\ &\le \max\{f_i^+(x; u_x) : i \in I(x)\} \le -\mu^{-1}. \end{aligned}$$

Therefore, from Theorem 4.2, the required result follows. ∎

COROLLARY 4.8. *For each $i \in I$, let $g_i : R^n \to R$ be differentiable and let $f_i(x) := g_i(x) + b_i^t x + c_i$, where $b_i = (b_{i1}, \ldots, b_{in})^t \in R^n$ and $c_i \in R$. Suppose that for all $i \in I$ and some $j \in J$ the coordinates b_{ij} have the same sign and all g_i's are independent of the jth coordinate x_j of $x \in R^n$. Then*

$$S := \{x \in R^n : f_i(x) \le 0 \text{ for all each } i \in I\}$$

is nonempty and for some $0 < \mu$ there holds $d_S(x) \le \mu f(x)_+$ for all $x \in R^n$.

Proof. In fact, for $x \in f^{-1}(0, +\infty)$ and $i \in I(x)$ and $u_x := (0, \ldots, 0, -sgn\, b_{ij}, 0, \ldots, 0)^t$ we have

$$f_i'(x; u_x) = \langle \nabla f_i(x), u_x \rangle = \langle \nabla g_i(x) + b_i, u_x \rangle = -|b_{ij}| \text{ for each } i \in I(x).$$

Taking $\mu^{-1} = \min\{|b_{ij}| : i \in I\}$ and applying Theorem 4.7, we arrive at the conclusion. □

Example 4.1. We consider the functions

$$\begin{aligned} f_1(x) &= g_1(x_1, x_2) + 2x_1 - x_2 + 3x_3, \\ f_2(x) &= g_2(x_1, x_2) + 2x_3, \\ f_3(x) &= g_3(x_1, x_2) + 2x_1 + 6x_3 - 4, \end{aligned}$$

where g_1, g_2, and g_3 are differentiable and independent of x_3. Since the coefficients of x_3 in f_i's are all positive and their minimum is 2,

$$S := \{x \in R^3 : f_i(x) \le 0 \text{ for } i = 1, 2, 3\}$$

is nonempty and $d_S(x) \le \frac{1}{2} f(x)_+$ holds for all $x \in R^3$.

Acknowledgment. We thank two anonymous referees for their valuable comments and suggestions.

REFERENCES

[1] J. F. Bonnans and A. D. Ioffe, *Second-order sufficiency and quadratic growth for non isolated minima*, Math. Oper. Res., 20 (1995), pp. 801–817.
[2] J. F. Bonnans and A. D. Ioffe, *Quadratic growth and stability in convex programming problems with multiple solutions*, J. Convex Anal., 2 (1995), pp. 41–57.
[3] F. H. Clarke, *Optimization and Nonsmooth Analysis*, Wiley-Interscience, New York, 1983; reprinted as Classics Appl. Math. 5, SIAM, Philadelphia, PA, 1990.
[4] F. H. Clarke, Yu. S. Ledyaev, R. J. Stern, and P. R. Wolenski, *Nonsmooth Analysis and Control Theory*, Grad. Texts in Math. 178, Springer-Verlag, New York, 1998.
[5] I. Ekeland, *On the variational principle,* J. Math. Anal. Appl., 47 (1974), pp. 324–353.
[6] H. Halkin, *Implicit functions and optimization problems without continuous differentiability of the data,* SIAM J. Control, 12 (1974), pp. 229–236.
[7] L. R. Huang and K. F. Ng, *On first and second-order conditions for error bounds,* preprint.
[8] A. J. Hoffman, *On approximate solutions of systems of linear inequalities,* J. Research Nat. Bur. Standards, 49 (1952), pp. 263–265.
[9] A. D. Ioffe, *Regular points of Lipschitz functions,* Trans. Amer. Math. Soc., 251 (1979), pp. 61–69.
[10] A. D. Ioffe, *Calculus of Dini subdifferentials of functions and contingent coderivatives of set-valued maps,* Nonlinear Anal., 8 (1984), pp. 517–539.
[11] A. S. Lewis and J. S. Pang, *Error bounds for convex inequality systems,* in Generalized Convexity, Generalized Monotonicity, J.-P. Crouzeix, J.-E. Martinez-Legaz, and M. Volle, eds., Kluwer Academic, Dordrecht, The Netherlands, 1998, pp. 75–110.
[12] X.-D. Luo and Z.-Q. Luo, *Extension of Hoffman's error bound to polynomial systems*, SIAM J. Optim., 4 (1994), pp. 383–392.
[13] Z.-Q. Luo and J. F. Sturm, *Error bounds for quadratic systems,* in High Performance Optimization, Appl. Optim. 33, Kluwer Academic, Dordrecht, The Netherlands, 2000, pp. 383–404.
[14] J. S. Pang, *Error bounds in mathematical programming,* Math. Programming, 79 (1997), pp. 299–332.
[15] K. F. Ng and X. Y. Zheng, *Global error bounds with fractional exponents,* Math. Program. Ser. B, 88 (2000), pp. 357–370.
[16] K. F. Ng and X. Y. Zheng, *Error bounds for lower semicontinuous functions in normed spaces*, SIAM J. Optim., 12 (2001), pp. 1–17.
[17] R. T. Rockafellar and R. J.-B. Wets, *Variational Analysis,* Grundlehren Math. Wiss. 317, Springer-Verlag, Berlin, Heidelberg, 1998.
[18] W. Takahashi, *Existence theorems generalizing fixed point theorems for multivalued mappings,* in Fixed Point Theory and Applications, M. A. Théra and J. B. Baillon, Pitman Res. Notes Math. Ser. 252, Longman Scientific and Technical, Harlow, UK, 1991, pp. 397–406.
[19] T. Wang and J. S. Pang, *Global error bounds for convex quadratic inequality systems*, Optimization, 31 (1994), pp. 1–12.

[20] D. Ward, *Sufficient conditions for weak sharp minima of order two and directional derivatives of the value function*, in Mathematical Programming with Data Perturbations, Lecture Notes in Pure and Appl. Math. 195, Dekker, New York, 1998, pp. 419–436.

[21] Z. Wu and J. J. Ye, *Sufficient conditions for error bounds,* SIAM J. Optim., 12 (2001), pp. 421–435.

[22] Z. Wu and J. J. Ye, *On error bounds for lower semicontinuous functions,* Math. Program. Ser. A, 92 (2002), pp. 301–314.

SIAM J. OPTIM.
Vol. 14, No. 3, pp. 646–669

A GLOBALLY CONVERGENT FILTER METHOD FOR NONLINEAR PROGRAMMING*

CLÓVIS C. GONZAGA†, ELIZABETH KARAS‡, AND MÁRCIA VANTI§

Abstract. In this paper we present a filter algorithm for nonlinear programming and prove its global convergence to stationary points. Each iteration is composed of a feasibility phase, which reduces a measure of infeasibility, and an optimality phase, which reduces the objective function in a tangential approximation of the feasible set. These two phases are totally independent, and the only coupling between them is provided by the filter. The method is independent of the internal algorithms used in each iteration, as long as these algorithms satisfy reasonable assumptions on their efficiency. Under standard hypotheses, we show two results: for a filter with minimum size, the algorithm generates a stationary accumulation point; for a slightly larger filter, all accumulation points are stationary.

Key words. filter methods, nonlinear programming, global convergence

AMS subject classifications. 49M37, 65K05, 90C30

DOI. 10.1137/S1052623401399320

1. Introduction.

We shall study the nonlinear programming problem

$$\text{(P)} \qquad \begin{array}{ll} \text{minimize} & f_0(x) \\ \text{subject to} & f_{\mathcal{E}}(x) = 0, \\ & f_{\mathcal{I}}(x) \le 0, \end{array}$$

where the index sets $\mathcal{E}$ and $\mathcal{I}$ refer to the equality and inequality constraints, respectively. Let the cardinality of $\mathcal{E} \cup \mathcal{I}$ be m, and assume that the functions $f_i : \mathbb{R}^n \to \mathbb{R}$ for $i = 0, 1, \ldots, m$ are continuously differentiable. The Jacobian matrices of $f_{\mathcal{E}}$ and $f_{\mathcal{I}}$ are denoted, respectively, $A_{\mathcal{E}}(\cdot)$ and $A_{\mathcal{I}}(\cdot)$.

We define the function $f^+ : \mathbb{R}^n \to \mathbb{R}^m$ by

$$f_i^+(x) = \begin{cases} f_i(x) & \text{if } \ i \in \mathcal{E}, \\ \max\{0, f_i(x)\} & \text{if } \ i \in \mathcal{I}. \end{cases} \tag{1.1}$$

The ith constraint, $i = 1, \ldots, m$, is satisfied at $x \in \mathbb{R}^n$ if $f_i^+(x) = 0$. We consider a measure of constraint infeasibility $x \in \mathbb{R}^n \mapsto h(x)$, which is an exact penalty applied to the constraints. Usually this measure is given by

$$h(x) = \|f^+(x)\|, \tag{1.2}$$

where $\|\cdot\|$ denotes an arbitrary norm.

A nonlinear programming algorithm must deal with two conflicting criteria, f_0 and h, which must be simultaneously minimized, with preference given to the infeasibility measure h, which must be driven to zero.

*Received by the editors December 7, 2001; accepted for publication (in revised form) January 6, 2003; published electronically December 19, 2003. The work of the authors was supported by CNPq and CAPES, Brazil.

http://www.siam.org/journals/siopt/14-3/39932.html

†Department of Mathematics, Federal University of Santa Catarina, Cx. Postal 5210, 88040-970 Florianópolis, SC, Brazil (clovis@mtm.ufsc.br).

‡Department of Mathematics, Federal University of Paraná and Federal University of Santa Catarina, Cx. Postal 19081, 81531-990 Curitiba, PR, Brazil (karas@mat.ufpr.br).

§Department of Electrical Engineering, Federal University of Santa Catarina, Cx. Postal 5210, 88040-970 Florianópolis, SC, Brazil (marcia@labplan.ufsc.br).

Optimality and feasibility can be combined using penalty functions or augmented Lagrangians, or they can be treated more or less independently. The methods studied in this paper belong to the class in which f_0 and h are treated as two independent objectives. Each iteration of these methods is composed of two phases: a feasibility phase, which decreases h, followed by an optimization phase, which decreases f_0.

Such methods can be traced back to Rosen's gradient projection method [16] and Abadie and Carpentier's GRG [1]. They are surveyed in Martínez and Pilotta [11]. Combining the ideas of sequential quadratic programming and trust region algorithms for problems with equality constraints only, Celis, Dennis, and Tapia [6] started a line of research which led to the method of Byrd [3] and Omojokum [14]: each iteration of this method works in a trust region centered at the current iterate x^k and is composed of a *normal* step (feasibility step) followed by a *tangential* step (optimality step). The tangential step must follow a direction in the null space of the constraint Jacobian at x^k.

The feasibility and optimality phases become more independent in the *inexact restoration* algorithms described by Martínez [9] and by Martínez and Pilotta [10, 11], who place the trust region used in each iteration around the point obtained *after* the feasibility phase. Any method for reducing h can be used in the feasibility phase: they describe an algorithm for problems with nonlinear equality constraints and box inequality constraints. Methods for the feasibility phase can also use ideas from Byrd, Gilbert, and Nocedal [4] and from Byrd, Hribar, and Nocedal [5], who rewrite the problem using equality constraints and nonnegative slack variables.

In these algorithms, the progress is usually measured by a *merit function* $\psi = f_0 + \nu h$, where ν is a positive weight. At iteration k, the points in $\{x \in \mathbb{R}^n \mid \psi(x) \geq \psi(x^k)\}$ are *forbidden*, and the step tries to decrease the value of ψ. The choice of ν may be tricky: small values of ν may forbid the optimal solutions; large values of ν may slow down the algorithm.

As a rule, algorithms must include some procedure to increase ν when needed, increasing the importance of h in ψ. This choice of ν usually depends on both the feasibility and optimality steps, reducing their independence.

Filter algorithms. Filter algorithms define a *forbidden region* in a clever way: by memorizing the pairs $(f_0(x^k), h(x^k))$ from well-chosen former iterations and then avoiding points dominated by these by the usual Pareto domination rule:

$$\text{“}x \text{ dominates } y \quad \text{if and only if} \quad f_0(y) \geq f_0(x) \quad \text{and} \quad h(y) \geq h(x).\text{”}$$

We cannot construct the set of forbidden points, but it very easy to check whether a point belongs to it by performing a small number of comparisons in $\mathbb{R}^2$.

These methods were introduced by Fletcher and Leyffer in their important paper [8], and a global convergence proof was obtained by Fletcher et al. [7]. The approach was also applied to interior point algorithms by Ulbrich, Ulbrich, and Vicente [17]. In these papers, each feasibility phase must reduce h until a property called *compatibility* is verified, which depends on a trust region radius and on the linear model of the constraints.

Our method is an inexact restoration algorithm in the sense of Martínez and Pilotta [10], which uses a filter. The method has the following characteristics:

- Each iteration starts with a filter and its associated forbidden region.
- The feasibility and optimality phases are totally independent and may be based on any algorithms satisfying some reasonable hypotheses. The only

connection between both phases is that they are not allowed to generate forbidden points.
- Differently from the filter algorithms cited above, no compatibility is required after a feasibility step: the only requirement is that h decreases by at least a fixed ratio.
- In our first algorithm, the number of pairs $(h(x^k), f_0(x^k))$ introduced in the filter is perhaps the minimum possible to guarantee the existence of a stationary accumulation point.

Local convergence. In this paper we deal with the global convergence of filter algorithms without discussing details of the internal algorithms. Fletcher and Leyffer [8] comment that filter algorithms may suffer from the Maratos effect and propose a second order correction to remedy this shortcoming. Wächter and Biegler in their recent work [18] propose a filter method using line searches and also discuss the usage of a second order correction. In our general approach, it is easy to show that the Maratos effect will be present when the method is applied to Powell's example [15]. Although we believe that second order correction schemes can be devised for this general setting, this will not be discussed in this paper.

Structure of the paper. In this section we present some general definitions and hypotheses. Section 2 describes the main algorithm and proves that under a very general hypothesis on the behavior of a complete step of the algorithm, any sequence generated by it has a stationary accumulation point. This section also discusses how to break this general hypothesis into reasonable independent assumptions for the feasibility and optimality phases. Section 3 describes the internal algorithms and shows how to satisfy the hypotheses used in section 2. Section 4 deepens the convergence analysis, showing that the objective values always converge under the hypotheses in section 2, and presents two improvements on the algorithms: first, using a slightly larger filter, we prove that all accumulation points are stationary; second, we discuss a simplified optimality step using the Jacobian matrices already calculated in the feasibility phase. Section 5 shows a graphical example, and an appendix proves some continuity properties.

Hypotheses. We shall develop algorithms which generate sequences (x^k) and (z^k) in $\mathbb{R}^n$. Here are the general hypotheses used in this paper.

(H1) The iterates (x^k) and (z^k) remain in a convex compact domain $X \subset \mathbb{R}^n$.

(H2) All the functions $f_i(\cdot)$ for $i = 0, 1, \ldots, m$ are uniformly Lipschitz continuously differentiable in an open set containing X.

(H3) All feasible accumulation points $\bar{x} \in X$ of (x^k) satisfy the Mangasarian–Fromovitz (M-F) qualification condition, namely, the gradients $\nabla f_i(\bar{x})$ for $i \in \mathcal{E}$ are linearly independent, and there exists a direction $d \in \mathbb{R}^n$ such that $A_{\mathcal{E}}(\bar{x})d = 0$ and $A_{\bar{\mathcal{I}}}(\bar{x})d < 0$, where $\bar{\mathcal{I}} = \{i \in \mathcal{I} \mid f_i(\bar{x}) = 0\}$.

The first hypothesis is quite usual. It can be enforced by adding a large box constraint to the problem. If the set $\{x \in \mathbb{R}^n \mid h(x) \le \bar{H}\}$ is bounded for some $\bar{H} \ge h(x^0)$, then the filter may start with a pair $(-\infty, \bar{H})$ (see below for the filter structure), thus forbidding forever points x with $h(x) \ge \bar{H}$. Similarly, if $\{x \in \mathbb{R}^n \mid f_0(x) \le \bar{F}\}$ for some upper bound $\bar{F}$ for the value of an optimal solution, then the pair $(\bar{F}, -\infty)$ in the filter ensures (H1). Of course, both entries can be used if $\{x \in \mathbb{R}^n \mid h(x) \le \bar{H}, f_0(x) \le \bar{F}\}$ is bounded.

From (H2) we conclude that for $x, y \in X$ and $i = 0, 1, \ldots, m$,

$$f_i(y) = f_i(x) + \nabla f_i(x)^T (y - x) + o(x, y), \tag{1.3}$$

where $|o(x,y)| \leq M\|x-y\|^2$ and $M>0$ is a Lipschitz constant.

The linearized sets. We shall associate with each $z \in \mathbb{R}^n$ a linearization of the set $\{x \in \mathbb{R}^n \mid f_{\mathcal{E}}(x) = f_{\mathcal{E}}(z),\ f_{\mathcal{I}}(x) \leq f_{\mathcal{I}}^+(z)\}$:

$$L(z) = \{x \in \mathbb{R}^n \mid A_{\mathcal{E}}(z)\,(x-z) = 0,\ f_{\mathcal{I}}(z) + A_{\mathcal{I}}(z)\,(x-z) \leq f_{\mathcal{I}}^+(z)\}. \tag{1.4}$$

At a feasible point z, $L(z)$ is a linearization of the feasible set. The following facts are easily seen:

- The M-F condition at a feasible point z is equivalent to the following: $A_{\mathcal{E}}(z)$ has linearly independent rows and the set $L(z)$ satisfies a Slater condition; i.e., $L(z)$ has an interior point, a point $y \in L(z)$ such that $f_{\mathcal{I}}(z) + A_{\mathcal{I}}(z)(y-z) < f_{\mathcal{I}}^+(z)$.
- The Karush–Kuhn–Tucker (KKT) conditions for (P) at z coincide with the KKT conditions at z for the problem of minimizing $f_0(\cdot)$ in $L(z)$. These conditions are also equivalent to the inexistence of a feasible descent direction from z into $L(z)$.

Optimality conditions. Here we make some comments on optimality conditions and on our usage of the expression *stationary point*.

Let us define the *projected Cauchy direction* or *projected gradient direction* associated with each $z \in \mathbb{R}^n$

$$d_c(z) = P_{L(z)}(z - \nabla f_0(z)) - z, \tag{1.5}$$

where $P_\Gamma(w)$ denotes the orthogonal projection of $w \in \mathbb{R}^n$ onto the closed set $\Gamma \subset \mathbb{R}^n$.

The projected gradient direction is well known. See, for instance, Bertsekas [2]. It satisfies $d_c(z) = 0$ if and only if there exists no feasible descent direction from z into $L(z)$. We conclude from the facts above that at a feasible z, the KKT conditions are equivalent to $d_c(z) = 0$. If $d_c(z) \neq 0$, then $\nabla f_0(z)^T d_c(z) < 0$.

Actually, this direction is the main construct used by Martínez and Svaiter [12] to define an optimality condition which lies between KKT and Fritz–John in generality: a feasible point $\bar{x}$ satisfies a Martínez–Svaiter optimality condition if and only if

$$\liminf_{x \to \bar{x}} \|d_c(x)\| = 0. \tag{1.6}$$

This optimality condition is actually quite constructive: what we shall prove in this paper is that our algorithms produce feasible limit points satisfying (1.6). These points will be called *stationary*.

Here we have two possible courses of action: either we rely on their paper and do not use the M-F condition, or use it and the fact that in this case KKT and Martínez–Svaiter are equivalent conditions. We choose the second option.

For completeness, we now prove this equivalence, using continuity properties of the point to set map $L(\cdot)$ which are shown in the appendix. We keep this treatment in the paper because we believe that it may have some interest in itself.

LEMMA 1.1. *Let $\bar{x}$ be a feasible point satisfying an M-F condition. Then*

(i) *the map* (1.5) *is continuous at $\bar{x}$;*

(ii) *$\bar{x}$ satisfies the KKT conditions if and only if it satisfies the Martínez–Svaiter conditions.*

Proof. (i) follows directly from Lemmas A.1 and A.2: under an M-F condition, $z \mapsto L(z)$ is a continuous map at $\bar{x}$ by Lemma A.1, and Lemma A.2 ensures that $z \mapsto P_{L(z)}(z - \nabla f_0(z))$ is continuous because $\nabla f_0(\cdot)$ is continuous.

To prove (ii), note that for a continuous map $d_c(\cdot)$, (1.6) is equivalent to $d_c(\bar{x}) = 0$, which as we saw above is equivalent to the KKT conditions, completing the proof. ∎

Notation. Given two nonnegative functions $g_1, g_2 : X \subseteq \mathbb{R}^n \to \mathbb{R}$ we say that

- $g_1(x) = O(g_2(x))$ (or equivalently $g_2(x) = \Omega(g_1(x))$) in $\Gamma \subseteq X$ if there exists $M > 0$ such that for all $x \in \Gamma$, $g_1(x) \leq M g_2(x)$;
- $g_1(x) = o(g_2(x))$ in $\Gamma \subseteq X$ if $\lim_{g_2(x)\to 0^+} \frac{g_1(x)}{g_2(x)} = 0$.

2. The algorithm. In this section we present the method with no specification of the internal algorithms used in the feasibility and optimality steps. Afterward we state assumptions on the performance of these steps and prove that any sequence generated by the algorithm has a stationary accumulation point. The next section will show that quite usual methods for the internal steps fulfill these assumptions.

ALGORITHM 2.1. *Filter algorithm.*

Data: $x^0 \in \mathbb{R}^n$, $F_0 = \emptyset$, $\mathcal{F}_0 = \emptyset$, $\alpha \in (0, 1)$.

$k = 0$

REPEAT

$(\tilde{f}_0, \tilde{h}) = (f_0(x^k) - \alpha h(x^k),\ (1 - \alpha)h(x^k))$.

Construct the set $\bar{F}_k = F_k \bigcup \{(\tilde{f}_0, \tilde{h})\}$.

Define the set $\bar{\mathcal{F}}_k = \mathcal{F}_k \bigcup \{x \in \mathbb{R}^n \mid f_0(x) \geq \tilde{f}_0, h(x) \geq \tilde{h}\}$.

Feasibility phase:

if $h(x^k) = 0$, then set $z^k = x^k$

else compute $z^k \notin \bar{\mathcal{F}}_k$ such that $h(z^k) < (1 - \alpha)\ h(x^k)$.

if impossible, then stop without success.

Optimality phase:

if z^k is stationary, then stop with success

else compute $x^{k+1} \notin \bar{\mathcal{F}}_k$ such that $x^{k+1} \in L(z^k)$ and $f_0(x^{k+1}) \leq f_0(z^k)$.

Filter update:

if $f_0(x^{k+1}) < f_0(x^k)$, then

$F_{k+1} = F_k, \quad \mathcal{F}_{k+1} = \mathcal{F}_k \qquad (f_0\text{-iteration})$

else

$F_{k+1} = \bar{F}_k, \quad \mathcal{F}_{k+1} = \bar{\mathcal{F}}_k \qquad (h\text{-iteration})$

$k = k + 1$.

Section 5 shows a graphical example, where each step of the algorithm is depicted. The main feature of the algorithm is the construction of the filter: at the beginning of each iteration, the pair $(f_0(x^k) - \delta,\ h(x^k) - \delta)$, with $\delta = \alpha h(x^k)$, $\alpha \in (0, 1)$, is *temporarily* introduced in the filter. After the complete iteration, this entry will become permanent in the filter only if the iteration *does not* produce a decrease in f_0.

The algorithm deals with the filter and with the forbidden set associated with it. One must keep in mind that the forbidden set is never constructed, but helps the understanding of the process.

Stopping rules. The algorithm can stop in two situations:

(i) A stationary point is obtained. In this case there is nothing to prove.

(ii) The feasibility algorithm fails. This may well happen, depending on the method used. A common condition that may cause the failure is the existence of a stationary point $\bar{x}$ for $h(\cdot)$, with $h(\bar{x}) \neq 0$.

Elimination of filter entries. Whenever a new entry (f_0^j, h^j) is introduced in the filter, one can eliminate from it all entries dominated by the incoming one. This

saves comparisons when checking whether a point is forbidden. See the example in section 5.

From now on we shall assume that the algorithm generates infinite sequences (x^k) and (z^k). We also assume that the hypotheses (H1)–(H3) are satisfied, and now we state the main assumption on the performance of the algorithm at each iteration. We shall postpone the discussion of this assumption until the end of this section, where it will be thoroughly analyzed and replaced by simpler ones. In the next section we shall state methods which satisfy this assumption.

The main hypothesis. Given an iterate x^k, we start by defining the *filter slack* at x^k:

$$H_k = \min\left\{1, \min\{h^j \mid (f_0^j, h^j) \in F_k, f_0^j \leq f_0(x^k)\}\right\}, \tag{2.1}$$

illustrated in Figure 2.1. Our main hypothesis is the following:

(H4) Given a feasible nonstationary point $\bar{x} \in X$, there exists a neighborhood V of $\bar{x}$ such that for any iterate $x^k \in V$,

$$f_0(x^k) - f_0(x^{k+1}) = \Omega(\sqrt{H_k}). \tag{2.2}$$

Note that (H4) is a local condition. The relation (2.2) means that there exists $M > 0$ dependent on $\bar{x}$ such that whenever x^k is near $\bar{x}$, $f_0(x^k) - f_0(x^{k+1}) \geq M\sqrt{H_k}$.

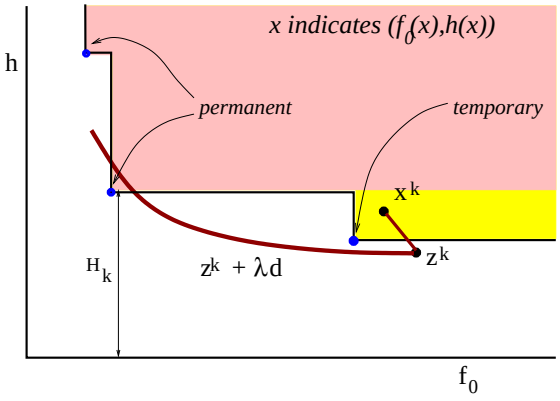


FIG. 2.1. *Example of the set $\bar{F}_k$ and of the quantity H_k.*

The following facts follow directly from the hypotheses and the construction made by the algorithm.

FACT 2.2. *Given $k \in \mathbb{N}$, $x^{k+p} \notin \mathcal{F}_{k+1}$ for all $p \geq 1$.*

FACT 2.3. *Given $k \in \mathbb{N}$, at least one of the following two situations occurs:*

(i) $h(x^{k+1}) \leq (1-\alpha)\, h(x^k)$.
(ii) $f_0(x^{k+1}) \leq f_0(x^k) - \alpha\, h(x^k)$.

FACT 2.4. *Given $k \in \mathbb{N}$, $h^j > 0$ for all $j \in \mathbb{N}$ such that $(f_0^j, h^j) \in F_k$. Consequently, $H_k > 0$ for all $k \in \mathbb{N}$.*

By Algorithm 2.1, the pair $(\tilde{f}_0, \tilde{h})$ is included in the filter at the end of the iteration if and only if that iteration is an h-iteration. If $\tilde{h} = h(x^k) = 0$, then $z^k = x^k$ and $f_0(x^{k+1}) < f_0(z^k)$, so the iteration k is an f_0-iteration, and both statements in the fact follow.

LEMMA 2.5. *Let $\bar{x} \in X$ be a nonstationary point. Then there exist $\bar{k} \in \mathbb{N}$ and a neighborhood V of $\bar{x}$ such that whenever $k > \bar{k}$ and $x^k \in V$, the iteration k is an f_0-iteration.*

Proof. If $\bar{x} \in X$ is a feasible point, then by (H4) and Fact 2.4 there exists a neighborhood V of $\bar{x}$ such that for all $x^k \in V$,

$$f_0(x^k) - f_0(x^{k+1}) = \Omega(\sqrt{H_k}) > 0,$$

and k is an f_0-iteration.

Assume that $\bar{x}$ is infeasible, i.e., $h(\bar{x}) > 0$. Assume by contradiction that there exists an infinite set $\mathcal{K} \subset \mathbb{N}$ such that $x^k \overset{\mathcal{K}}{\to} \bar{x}$ and all iterations in $\mathcal{K}$ are h-iterations. Since h and f_0 are continuous functions, we have

$$h(x^k) \overset{\mathcal{K}}{\to} h(\bar{x}) \qquad \text{and} \qquad f_0(x^k) \overset{\mathcal{K}}{\to} f_0(\bar{x}).$$

Then there must exist $k_1 \in \mathcal{K}$ such that for all $k \in \mathcal{K}$, $k \geq k_1$,

$$|h(x^k) - h(\bar{x})| < \frac{\alpha}{2} h(x^{k_1}) \qquad \text{and} \qquad |f_0(x^k) - f_0(\bar{x})| < \frac{\alpha}{2} h(x^{k_1}). \tag{2.3}$$

For any given $k_2 \in \mathcal{K}$ such that $k_2 > k_1$,

$$|h(x^{k_2}) - h(\bar{x})| < \frac{\alpha}{2} h(x^{k_1}) \qquad \text{and} \qquad |f_0(x^{k_2}) - f_0(\bar{x})| < \frac{\alpha}{2} h(x^{k_1}). \tag{2.4}$$

Using the triangle inequality, (2.3), and (2.4), we have

$$|h(x^{k_2}) - h(x^{k_1})| < \alpha\, h(x^{k_1}) \qquad \text{and} \qquad |f_0(x^{k_2}) - f_0(x^{k_1})| < \alpha\, h(x^{k_1}).$$

Therefore $x^{k_2} \in \mathcal{F}_{k_1+1}$, contradicting Fact 2.2 and completing the proof. ∎

LEMMA 2.6. *Suppose that $(x^k)_{k\in\mathbb{N}}$ has no stationary accumulation point. Then for k sufficiently large, all iterations are f_0-iterations.*

Proof. Assume by contradiction that there exists an infinity of h-iterations. Then there exists an infinite set $\mathcal{K}_1 \subset \mathbb{N}$ such that for $k \in \mathcal{K}_1$, the iteration k is an h-iteration. By hypothesis (H1), $(x^k)_{k\in\mathcal{K}_1}$ is bounded, and hence there exist $\mathcal{K}_2 \subset \mathcal{K}_1$ and $\bar{x} \in \mathbb{R}^n$ such that $x^k \overset{\mathcal{K}_2}{\to} \bar{x}$. From the previous lemma, $\bar{x}$ must be a stationary accumulation point, contradicting the hypothesis and completing the proof. ∎

THEOREM 2.7. *The sequence (x^k) has a stationary accumulation point.*

Proof. Assume by contradiction that $(x^k)_{k\in\mathbb{N}}$ has no stationary accumulation point. Then from Lemma 2.6 for k large (say, $k > k_1$), all iterations are f_0-iterations, $f_0(x^k)$ decreases, and hence

$$f_0(x^{k+1}) - f_0(x^k) \to 0. \tag{2.5}$$

For any $k \geq k_1$, $F_k = F_{k_1}$ by construction, and using Fact 2.4, $H_k \geq H_{k_1} > 0$.

The sequence (x^k) cannot have a feasible accumulation point, because by the hypothesis (H4), if there exist $\mathcal{K}_1 \in \mathbb{N}$ and a feasible $\bar{x} \in X$ such that $x^k \overset{\mathcal{K}_1}{\to} \bar{x}$, then for large $k \in \mathcal{K}_1$ (say, $k > k_2 > k_1$)

$$f_0(x^k) - f_0(x^{k+1}) = \Omega(\sqrt{H_{k_1}}) > 0,$$

contradicting (2.5).

Now we prove the following claim: for large $k \in \mathbb{N}$,

$$h(x^{k+1}) \leq (1-\alpha)\, h(x^k). \tag{2.6}$$

Assume by contradiction that in some infinite set $\mathcal{K}_2 \subset \mathbb{N}$,

$$h(x^{k+1}) > (1-\alpha)\, h(x^k).$$

Using Fact 2.3, for $k \in \mathcal{K}_2$,

$$f_0(x^{k+1}) \leq f_0(x^k) - \alpha\, h(x^k).$$

Using (2.5), we conclude that $h(x^k) \overset{\mathcal{K}_2}{\to} 0$, which contradicts the fact that (x^k) has no feasible accumulation points.

Hence (2.6) holds and $h(x^k)$ converges linearly to zero. This again contradicts the fact that (x^k) has no feasible accumulation points, completing the proof. ∎

The hypothesis (H4). This hypothesis is an assumption on each complete iteration. Although it may be difficult to check for specific algorithms, its interpretation is simple: near a feasible nonstationary point, the optimality step dominates, and the reduction of f_0 is large. The filter slack H_k indicates how much h is allowed to increase in the tangential step, and, by being tangential, it is expected that h changes with the square of the variation of x. In an efficient tangential step, f_0 will vary linearly with the variation of x, and then (H4) will be true.

Now we show how (H4) can be replaced by simpler hypotheses made separately for the feasibility and optimality steps.

Feasibility step condition.

(H5) At all iterations $k \in \mathbb{N}$, the feasibility step must satisfy

$$h(x^k) - h(z^k) = \Omega(\|z^k - x^k\|). \tag{2.7}$$

This can also be stated as

$$\|z^k - x^k\| = O(h(x^k)), \tag{2.8}$$

because $h(z^k) \geq 0$. Note that since $\nabla f_0(\cdot)$ is bounded in X, by the mean-value theorem, for all $k \in \mathbb{N}$,

$$|f_0(z^k) - f_0(x^k)| = O(\|z^k - x^k\|).$$

Using this and (2.8) we have

$$|f_0(z^k) - f_0(x^k)| = O(h(x^k)). \tag{2.9}$$

Optimality step condition.

(H6) Given a feasible nonstationary point $\bar{x} \in X$, there exists a neighborhood V of $\bar{x}$ such that for any iterate $x^k \in V$,

$$f_0(z^k) - f_0(x^{k+1}) = \Omega(\sqrt{H_k}). \tag{2.10}$$

The assumption (H5) is used by Martínez [9] and is a global condition. It means that the feasibility step must be efficient, in the sense that the direction $z^k - x^k$ must be a good descent direction for h. Martínez discusses this hypothesis and shows that it

is satisfied under reasonable conditions. The assumption (H6) isolates the tangential step and is local (associated with each given nonstationary feasible point). It has the same interpretation as the one given for (H4), but now without the influence of the feasibility step.

Remark. Note, however, that condition (H6) is not completely independent of the feasibility phase, because it uses H_k, which is associated with x^k. Also, the condition is stated for $x^k \in V$, and not $z^k \in V$, but this is not important because $\|x^k - z^k\| = O(h(x^k))$: if x^k is near $\bar{x}$, then the same is true for z^k.

Before proving that (H5) and (H6) imply (H4), let us state one more hypothesis which is not needed here but which is very reasonable and will be useful ahead. It is similar to (H5) but applied to the objective function.

(H7) Given a feasible nonstationary point $\bar{x} \in X$, there exists a neighborhood V of $\bar{x}$ such that for any iterate $x^k \in V$,

$$f_0(z^k) - f_0(x^{k+1}) = \Omega(\|x^{k+1} - z^k\|).$$

With this hypothesis, (H6) can be stated as $\|z^k - x^{k+1}\| = \Omega(\sqrt{H_k})$ and has a simple interpretation: if the filter restricts the step (H_k is small), then this means that the variation of h is of the order of $\|x^k - x^{k+1}\|^2$, which is quite reasonable in a tangential step; otherwise (H_k is large), the condition means that $\|z^k - x^{k+1}\| = \Omega(1)$; i.e., near a fixed nonstationary point, an unconstrained tangential step is always large. Figure 2.1 illustrates the trajectory of the pair $(f_0(z^k + \lambda d), h(z^k + \lambda d))$ as λ grows and $d = x^{k+1} - z^k$.

Finally, we prove two lemmas, extending for the whole step the properties of the tangential step near a feasible nonstationary point.

LEMMA 2.8. (H5) *and* (H6) *imply* (H4).

Proof. Let $\bar{x}$ be a nonstationary feasible point, and let V_1 be the neighborhood defined by (H6). Since $\|x^k - z^k\| = O(h(x^k))$, there exists a neighborhood $\tilde{V}_1 \subset V_1$ of $\bar{x}$ such that for $x^k \in \tilde{V}_1$, $z^k \in V_1$ and $h(x^k) < 1$. Consider an iterate x^k in $\tilde{V}_1$. By definition of H_k, we have $h(x^k) \le H_k$. By (H5) and (H6), there are positive constants M and N such that

$$\begin{aligned} f_0(x^k) - f_0(x^{k+1}) &= f_0(x^k) - f_0(z^k) + f_0(z^k) - f_0(x^{k+1}) \\ &\ge M\sqrt{H_k} - Nh(x^k) \\ &\ge \left(M - N\sqrt{h(x^k)}\right)\sqrt{H_k}. \end{aligned}$$

By continuity of h at $\bar{x}$, there exists a neighborhood $V \subset \tilde{V}_1$ such that for any $x \in V$, $\sqrt{h(x)} \le 0.5M/N$. For any iterate x^k in this neighborhood, $f_0(x^k) - f_0(x^{k+1}) \ge 0.5M\sqrt{H_k}$, completing the proof. ∎

LEMMA 2.9. *Assume that* (H5)–(H7) *hold. Then given a feasible nonstationary point $\bar{x} \in X$, there exists a neighborhood V of $\bar{x}$ such that for any $x^k \in V$,*

$$f_0(x^k) - f_0(x^{k+1}) = \Omega(\|x^{k+1} - x^k\|).$$

Proof. Let V_1 and V_2 be the neighborhoods of a feasible nonstationary point $\bar{x}$ provided, respectively, by (H6) and (H7). As in the proof of Lemma 2.8, in some neighborhood $\tilde{V}_1 \subset V_1$ of $\bar{x}$, we have

$$f_0(x^k) - f_0(x^{k+1}) = f_0(x^k) - f_0(z^k) + f_0(z^k) - f_0(x^{k+1}),$$

with $|f_0(x^k) - f_0(z^k)| = O(h(x^k))$ by (2.9) and $f_0(z^k) - f_0(x^{k+1}) = \Omega(\sqrt{H^k})$. We easily deduce from these two facts that for x^k sufficiently near $\bar{x}$, say $x^k \in V_3 \subset \tilde{V}_1$, $|f_0(x^k) - f_0(z^k)| \leq 0.5(f_0(z^k) - f_0(x^{k+1}))$. It follows that

$$f_0(x^k) - f_0(x^{k+1}) \geq 0.5(f_0(z^k) - f_0(x^{k+1})). \tag{2.11}$$

We can also write

$$\|x^k - x^{k+1}\| \leq \|x^k - z^k\| + \|z^k - x^{k+1}\|,$$

with $\|x^k - z^k\| = O(h(x^k))$ by (H5) and $\|z^k - x^{k+1}\| = \Omega(f_0(z^k) - f_0(x^{k+1}))$ by the Lipschitz continuity of f_0. Again by the same reasoning as in the proof of Lemma 2.9, for $x^k \in \tilde{V}_2 \subset V_2$, $z^k \in V_2$, and we obtain from (H6) $\|z^k - x^{k+1}\| = \Omega(\sqrt{H_k})$. As above, we deduce that for x^k sufficiently near $\bar{x}$, say $x^k \in V_4 \subset \tilde{V}_2$, $\|x^k - z^k\| \leq \|z^k - x^{k+1}\|$, and hence

$$\|x^k - x^{k+1}\| \leq 2\,\|z^k - x^{k+1}\|.$$

Using in sequence (2.11), hypothesis (H7), and this expression in the neighborhood $V = V_3 \cap V_4$, we obtain

$$\begin{aligned} f_0(x^k) - f_0(x^{k+1}) &\geq 0.5(f_0(z^k) - f_0(x^{k+1})) \\ &= \Omega(\|z^k - x^{k+1}\|) \\ &= \Omega(\|x^k - x^{k+1}\|), \end{aligned}$$

completing the proof. □

3. Internal algorithms. In this section we discuss the internal steps used in each iteration of the main algorithm. We assume that Algorithm 2.1 has generated infinite sequences (x^k) and (z^k) and that hypotheses (H1)–(H3) are satisfied.

Feasibility step algorithm. The purpose of the feasibility phase is to find a point z^k such that $h(z^k) < (1-\alpha)h(x^k)$ and $z^k \notin \bar{\mathcal{F}}_k$. The procedure used in this phase could in principle be any iterative algorithm for decreasing h, and finite termination should be achieved because as we have seen above all filter entries $(f_0^j, h^j) \in F_k$ have $h^j > 0$.

The feasibility step studied by Martínez [9] satisfies assumption (H5) and applies directly to our case. Thus we shall not describe the feasibility procedure in detail in this paper.

Note that the feasibility algorithm may fail if $h(\cdot)$ has an infeasible stationary point. In this case, the method stops without success.

Optimality step algorithm. The optimality step must find x^{k+1} in the linearized set $L(z^k)$ such that $f_0(x^{k+1}) \leq f_0(z^k)$, and such that $x^{k+1} \notin \bar{\mathcal{F}}_k$. We shall describe a very general trust region method for this and then show that the resulting step satisfies the assumptions (H6) and (H7).

The main tool for the analysis (not necessarily for the construction) of such algorithms is the projected Cauchy direction described in the introduction.

The projected gradient method. A very simple (but impractical) method for the tangential step is the following: from z^k, compute the projected Cauchy direction $d_c(z^k) = P_{L(z^k)}(z^k - \nabla f_0(z^k)) - z^k$ and perform an Armijo search along $z^k + \lambda d_c(z^k)$, $\lambda \geq 0$. The search must avoid forbidden points, which can be achieved by using in the

Armijo search the objective $\theta(\lambda) = f_0(z^k + \lambda d_c(z^k))$ if $z^k + \lambda d_c(z^k) \notin \bar{\mathcal{F}}_k$, $\theta(\lambda) = +\infty$ otherwise.

We shall not prove the efficiency of this tangential step, because it is a particular case of the trust region iteration to be described from now on. The main requirement on the trust region step will be that it produces a point at least as good as the so-called Cauchy point, which lies on this projected gradient direction.

The quadratic model. Given $z^k \in X$ generated by Algorithm 2.1 in the feasibility phase, the trust region algorithm associates to z^k a quadratic model of f_0,

$$x \in \mathbb{R}^n \mapsto m_k(x) = f_0(z^k) + \nabla f_0(z^k)^T(x - z^k) + \frac{1}{2}(x - z^k)^T B_k(x - z^k), \tag{3.1}$$

where B_k is an $n \times n$ symmetric matrix. This matrix may be an approximation of $\nabla^2 f_0(z^k)$, or any other matrix, provided that the hypothesis (H8) below is verified. Usually, B_k will be an approximation of the Hessian of some Lagrangian function, and then m_k deviates from a straightforward model of f_0 by incorporating the curvature along the manifold of the constraints. Although this may be essential in the design of efficient algorithms, this discussion is out of the scope of this paper.

(H8) There exists $\beta > 0$ such that the quadratic model (3.1) satisfies $\|B_k\| \le \beta$ for all $k \in \mathbb{N}$.

The trust region step uses a radius $\Delta > 0$ and computes a step $d(z^k, \Delta) \in \mathbb{R}^n$ such that $\|d(z^k, \Delta)\| \le \Delta$. We define the *predicted reduction* produced by the step $d(z^k, \Delta)$ as

$$pred(z^k, \Delta) = m_k(z^k) - m_k(z^k + d(z^k, \Delta)) \tag{3.2}$$

and the *actual reduction* as

$$ared(z^k, \Delta) = f_0(z^k) - f_0(z^k + d(z^k, \Delta)). \tag{3.3}$$

LEMMA 3.1. *Consider $z^k \in X$ and $d(z^k, \Delta) \in \mathbb{R}^n$ generated by the trust region algorithm. Then*

$$ared(z^k, \Delta) = pred(z^k, \Delta) + o(z^k, \Delta), \tag{3.4}$$

where

$$\lim_{\Delta \to 0^+} \frac{o(z^k, \Delta)}{\Delta} = 0$$

uniformly in $z^k \in X$.

Proof. From (3.2)

$$\begin{aligned} -pred(z^k, \Delta) &= \nabla f_0(z^k)^T d(z^k, \Delta) + \tfrac{1}{2} d(z^k, \Delta)^T B_k d(z^k, \Delta) \\ &= \nabla f_0(z^k)^T d(z^k, \Delta) + O(\Delta^2) \end{aligned}$$

because $\|d(z^k, \Delta)\| \le \Delta$ and $\|B_k\| \le \beta$. From (3.3) and (1.3),

$$-ared(z^k, \Delta) = \nabla f_0(z^k)^T d(z^k, \Delta) + o(z^k, \Delta),$$

where

$$\lim_{\Delta \to 0^+} \frac{o(z^k, \Delta)}{\Delta} = 0.$$

This limit is uniform in $z^k \in X$ because $\nabla f_0(\cdot)$ is Lipschitz continuous in X. Hence

$$ared(z^k, \Delta) = pred(z^k, \Delta) + O(\Delta^2) - o(z^k, \Delta),$$

completing the proof. □

In the optimality step algorithm, which we will discuss below, we made the following choices which simplify the treatment:

(1) Each trust region computation starts with a radius $\Delta \geq \Delta_{\min}$, where $\Delta_{\min} > 0$ is fixed. The choice of Δ is irrelevant for the theory, and it usually comes from the former iteration. The use of this minimum radius $\Delta_{\min}$ simplifies the treatment substantially. In well-designed trust region algorithms for unconstrained problems this is not needed, but the convergence proofs become quite involved (see [13, Theorem 4.7]).

(2) A step $d(z^k, \Delta)$ is accepted only if the sufficient decrease condition is satisfied:

$$ared(z^k, \Delta) > \eta\, pred(z^k, \Delta) \tag{3.5}$$

for a given $\eta \in (0, 1)$.

(3) The trust region computation solves approximately the problem

$$\begin{array}{ll} \text{minimize} & m_k(x) \\ \text{subject to} & x \in L(z^k), \\ & \|x - z^k\| \leq \Delta, \end{array} \tag{3.6}$$

where $\|\cdot\|$ is any norm in $\mathbb{R}^n$.

Now we explain what we mean by "solving approximately." Given $z \in X$ and the set $L(z)$, the projected gradient direction is defined by

$$d_c(z) = P_{L(z)}(z - \nabla f_0(z)) - z. \tag{3.7}$$

Define

$$\varphi(z) = -\nabla f_0(z)^T \frac{d_c(z)}{\|d_c(z)\|}.$$

Then φ is the descent rate of f_0 along d_c. As usual, we denote $d_c^k = d_c(z^k)$, $\varphi^k = \varphi(z^k)$. As we saw in the introduction, $\varphi(z) > 0$ whenever z is a feasible nonstationary point.

Now we use known results about the minimization of $m_k(\cdot)$ along a direction; see the discussion on the Cauchy point in [13]. Defining the generalized Cauchy point as the minimizer of $m_k(\cdot)$ along d_c in the trust region $\{x \in \mathbb{R}^n \mid \|x - z^k\| \leq \Delta\}$,

$$x_c = \operatorname{argmin} \left\{ m_k(x) \mid \|x - z^k\| \leq \Delta, x = z^k + \lambda d_c^k, \lambda \geq 0 \right\},$$

we know that

$$m_k(z^k) - m_k(x_c) \geq \frac{\xi \varphi^k}{2} \min \left\{ \frac{\varphi^k}{\|B_k\|},\ \|d_c^k\|,\ \Delta \right\},$$

where ξ depends on the norms used. Using hypothesis (H8), this can be rewritten as

$$m_k(z^k) - m_k(x_c) \geq \frac{\xi \varphi^k}{2} \min \left\{ \frac{\varphi^k}{\beta},\ \|d_c^k\|,\ \Delta \right\}. \tag{3.8}$$

We accept as an approximate solution of (3.6) any feasible solution for this problem satisfying (3.8).

After stating the trust region step we shall study its properties.

ALGORITHM 3.2. *Optimality step.*
Data: $\eta \in (0,1)$, $\Delta_{\min} > 0$, $z^k \notin \bar{\mathcal{F}}_k$, $\Delta = \Delta^0 \geq \Delta_{\min}$.
REPEAT
Compute $d = d(z^k, \Delta)$ such that $\|d\| \leq \Delta$, $z^k + d \in L(z^k)$ and

$$pred(z^k, \Delta) \geq \frac{\xi \varphi^k}{2} \min\left\{ \frac{\varphi^k}{\|B_k\|}, \|d_c^k\|, \Delta \right\}.$$

Set $ared(z^k, \Delta) = f_0(z^k) - f_0(z^k + d)$.
if $z^k + d \notin \bar{\mathcal{F}}_k$ and $ared(z^k, \Delta) > \eta\, pred(z^k, \Delta)$
set $x^{k+1} = z^k + d$, $\Delta_k = \Delta$, and exit with success
else $\Delta = \Delta/2$.

Our task now is proving that this algorithm satisfies the assumptions (H6) and (H7) made for the optimality step.

LEMMA 3.3. *For any $z \in X$, $d \in \mathbb{R}^n$ such that $(z + d) \in L(z)$,*

$$|h(z + d) - h(z)| = O(\|d\|^2).$$

Proof. From (1.3), for any $z \in X$, $d \in \mathbb{R}^n$, $i = 0, 1, \ldots, m$,

$$f_i(z + d) - f_i(z) \leq \nabla f_i(z)^T d + O(\|d\|^2).$$

Since $(z + d) \in L(z)$, by the definition of $L(z)$ given in (1.4), we have for $i = 1, \ldots, m$

$$f_i(z) + \nabla f_i(z)^T d \leq f_i^+(z),$$

and hence

$$f_i(z + d) \leq f_i^+(z) + O(\|d\|^2).$$

We must prove that

$$f_i^+(z + d) \leq f_i^+(z) + O(\|d\|^2). \tag{3.9}$$

If $f_i(z+d) < 0$ and $i \in \mathcal{I}$, this is true because the right-hand side is positive. Otherwise $f_i^+(z + d) = f_i(z + d)$. Using (3.9) in the norm definition we set

$$\|f^+(z + d)\| = \|f^+(z)\| + O(\|d\|^2),$$

completing the proof. ∎

Now we study the optimality step near a nonstationary feasible point $\bar{x} \in X$. The first lemma says that if we ignore the filter, then the trust region step is large near $\bar{x}$.

LEMMA 3.4. *Let $\bar{x} \in X$ be a feasible nonstationary point satisfying an M-F condition. Then there exist a neighborhood $\tilde{V}$ of $\bar{x}$, $\tilde{\Delta} \in (0, \Delta_{\min})$, and a constant $\tilde{c} > 0$ such that for any $z^k \in \tilde{V}$,*

(i) *for any $\Delta > 0$, $pred(z^k, \Delta) \geq \tilde{c}\, \min\{\Delta, \tilde{\Delta}\}$;*
(ii) *for any $\Delta \in (0, \tilde{\Delta})$, $ared(z^k, \Delta) > \eta\, pred(z^k, \Delta) \geq \eta \tilde{c}\, \Delta$.*

Proof. From the generalized Cauchy decrease condition (3.8), which is satisfied by construction at each iteration,

$$pred(z^k, \Delta) \geq \frac{\xi \varphi(z^k)}{2}\ \min\left\{ \frac{\varphi(z^k)}{\beta},\ \|d_c^k\|,\ \Delta \right\}.$$

From Lemma 1.1 we deduce that $z \mapsto \|d_c(z)\|$ and $z \mapsto \varphi(z) = -\nabla f_0(z)^T \frac{d_c(z)}{\|d_c(z)\|}$ are continuous at $\bar{x}$. Hence there exists a neighborhood $\tilde{V}$ of $\bar{x}$ such that for $z^k \in \tilde{V}$, $\varphi(z^k) \geq \varphi(\bar{x})/2$ and $\|d_c(z^k)\| \geq \|d_c(\bar{x})\|/2$. Thus, for $z^k \in \tilde{V}$,

$$pred(z^k, \Delta) \geq \frac{\xi\varphi(\bar{x})}{4} \min\left\{\frac{\varphi(\bar{x})}{2\beta}, \frac{\|d_c(\bar{x})\|}{2}, \Delta\right\}.$$

This can be written as $pred(z^k, \Delta) \geq \tilde{c} \min\{\Delta_1, \Delta\}$, proving (i).

From Lemma 3.1, for any $k \in \mathbb{N}$ and $\Delta > 0$,

$$\begin{aligned} ared(z^k, \Delta) &= pred(z^k, \Delta) + o(z^k, \Delta) \\ &= \eta\, pred(z^k, \Delta) + (1-\eta)\, pred(z^k, \Delta) + o(z^k, \Delta), \end{aligned}$$

where $\lim_{\Delta \to 0^+} \frac{o(z^k,\Delta)}{\Delta} = 0$ uniformly in z^k. For $\Delta \leq \Delta_1$, $pred(z^k, \Delta) \geq \tilde{c}\,\Delta$ and then

$$ared(z^k, \Delta) \geq \eta\, pred(z^k, \Delta) + (1-\eta)\tilde{c}\,\Delta + o(z^k, \Delta).$$

For Δ sufficiently small, say, $\Delta \leq \tilde{\Delta} \leq \Delta_1$, $(1-\eta)\tilde{c}\,\Delta + o(z^k, \Delta) \geq 0$, completing the proof. ∎

Algorithm 3.2 in iteration k starts with $\Delta^0 \geq \Delta_{\min}$ and iterates by setting $\Delta^j = 2^{-j}\Delta^0$, $j = 0, 1, \ldots$, and computing for each Δ^j the step $d(z^k, \Delta^j)$. Whenever $\Delta^j < \tilde{\Delta}$, the condition $ared(z^k, \Delta^j) > \eta\, pred(z^k, \Delta^j)$ is satisfied; the radius Δ^j can only be rejected if $z^k + d(z^k, \Delta^j) \in \bar{\mathcal{F}}_k$.

By construction, $z^k \notin \bar{\mathcal{F}}_k$. Since $\bar{\mathcal{F}}_k$ is a closed set, for Δ^j sufficiently small, $z^k + d(z^k, \Delta^j) \notin \bar{\mathcal{F}}_k$. This shows that the algorithm always terminates.

LEMMA 3.5. *Let $\bar{x} \in X$ be a feasible nonstationary point satisfying an M-F condition, and assume that* (2.7) *holds. Then there exists a neighborhood V of $\bar{x}$ such that for $x^k \in V$,*

$$f_0(z^k) - f_0(x^{k+1}) = \Omega(\sqrt{H_k}), \tag{3.10}$$

$$f_0(z^k) - f_0(x^{k+1}) = \Omega(\|x^{k+1} - z^k\|), \tag{3.11}$$

where $x^{k+1} = z^k + d(z^k, \Delta)$ is computed by Algorithm 3.2.

Proof. By a usual argument, it is enough to prove that for any subsequence $(x^k)_{k\in\mathcal{K}}$ converging to $\bar{x}$, (3.10) and (3.11) are true for large $k \in \mathcal{K}$.

Assume that $x^k \overset{\mathcal{K}}{\to} \bar{x}$, where $\mathcal{K} \subset \mathbb{N}$. It follows that $z^k \overset{\mathcal{K}}{\to} \bar{x}$, because by (2.8) $\|x^k - z^k\| = O(h(x^k)) \overset{\mathcal{K}}{\to} 0$.

Let $\tilde{V} \subset X$ and $\tilde{\Delta} > 0$ be the neighborhood of $\bar{x}$ and radius given by Lemma 3.4. For large $k \in \mathcal{K}$, say $k \in \mathcal{K}_1 \subset \mathcal{K}$, $z^k \in \tilde{V}$. Let us now consider an iteration $k \in \mathcal{K}_1$, and denote $(\tilde{f}_0, \tilde{h}) = (f_0(x^k) - \alpha h(x^k), (1-\alpha)h(x^k))$ the temporary entry in the filter.

Algorithm 3.2 starts with a radius $\Delta^0 \geq \Delta_{\min}$ and computes $d(z^k, \Delta^j)$, $\Delta^j = 2^{-j}\Delta^0$, $j = 0, 1, \ldots$, until $z^k + d(z^k, \Delta^j) \notin \bar{\mathcal{F}}_k$ and $ared(z^k, \Delta^j) > \eta\, pred(z^k, \Delta^j)$. Then $\Delta_k = \Delta^j$. Let us define $\hat{\Delta}$ as the first Δ^j such that

$$ared(z^k, \Delta^j) > \eta\, pred(z^k, \Delta^j) \quad \text{and} \tag{3.12}$$

$$z^k + d(z^k, \Delta^j) \notin \bar{\mathcal{F}}_k \quad \text{or} \quad f_0(z^k + d(z^k, \Delta^j)) \geq \tilde{f}_0. \tag{3.13}$$

Let us denote $\hat{d} = d(z^k, \hat{\Delta})$ and $\hat{x} = z^k + \hat{d}$. Note that $\hat{\Delta} \geq \Delta_k$, and $\hat{\Delta} > \Delta_k$ happens only when $f_0(\hat{x}) \geq \tilde{f}_0$. We shall derive properties of this step $\hat{d}$ and then prove that this situation cannot occur when x^k is sufficiently near $\bar{x}$.

We shall first prove that $\hat{x}$ satisfies the bounds in the lemma. Choose $\bar{\Delta} \le \tilde{\Delta}/2$.

(i) First, the easy case: assume that $\hat{\Delta} \ge \bar{\Delta}$. Then by Lemma 3.4,

$$pred(z^k, \hat{\Delta}) \ge \tilde{c} \min\{\hat{\Delta}, \tilde{\Delta}\} \ge \tilde{c}\bar{\Delta}.$$

By definition of $\hat{\Delta}$, (3.12) holds, and hence

$$f_0(z^k) - f_0(\hat{x}) \ge \eta\tilde{c}\bar{\Delta} = \Omega(1).$$

It follows trivially that $f_0(z^k) - f_0(\hat{x}) = \Omega(\sqrt{H_k})$ and $f_0(z^k) - f_0(\hat{x}) = \Omega(\|x^k - \hat{x}\|)$, because in both cases the right-hand side is bounded in X.

(ii) Now, assume that $\hat{\Delta} < \bar{\Delta}$. Then the radius $2\hat{\Delta} < 2\bar{\Delta} \le \tilde{\Delta} < \Delta_{\min}$ does not satisfy (3.13) (and was rejected by Algorithm 3.2). By Lemma 3.4,

$$ared(z^k, d(z^k, 2\hat{\Delta})) > \eta\, pred(z^k, d(z^k, 2\hat{\Delta})),$$

and it follows from (3.13) that $z^k + d(z^k, 2\hat{\Delta}) \in \bar{\mathcal{F}}_k$ and $f_0(z^k + d(z^k, 2\hat{\Delta})) < \tilde{f}_0$. By definition of H_k, we must have $h(z^k + d(z^k, 2\hat{\Delta})) \ge H_k$.

By construction, $h(z^k) < (1-\alpha)h(x^k) \le (1-\alpha)H_k$. Thus,

$$h(z^k + d(z^k, 2\hat{\Delta})) - h(z^k) \ge \alpha H_k.$$

By Lemma 3.3,

$$h(z^k + d(z^k, 2\hat{\Delta})) - h(z^k) = O(\|d(z^k, 2\hat{\Delta})\|^2) = O(\hat{\Delta}^2), \tag{3.14}$$

because $\|d(z^k, 2\hat{\Delta})\| \le 2\hat{\Delta}$. Merging these two results, we obtain $\alpha H_k \le O(\hat{\Delta}^2)$, or

$$\hat{\Delta} = \Omega(\sqrt{H_k}). \tag{3.15}$$

Using Lemma 3.4 again with $\hat{\Delta} < \bar{\Delta} < \tilde{\Delta}$,

$$f_0(z^k) - f_0(\hat{x}) \ge \eta\tilde{c}\,\Omega(\sqrt{H_k}) = \Omega(\sqrt{H_k}), \tag{3.16}$$

$$f_0(z^k) - f_0(\hat{x}) \ge \eta\tilde{c}\hat{\Delta} = \Omega(\hat{\Delta}). \tag{3.17}$$

So, the step $\hat{d}$ satisfies the conditions in the lemma.

To finish the proof, we must show that for large $k \in \mathcal{K}_2$, $f_0(\hat{x}) < \tilde{f}_0$, which implies $\hat{x} \notin \bar{\mathcal{F}}_k$, and thus $\hat{x} = x^{k+1}$. From (3.16) and (2.9), there are positive constants M and N such that

$$f_0(\hat{x}) \le f_0(z^k) - M\sqrt{H_k},$$
$$f_0(z^k) \le f_0(x^k) + Nh(x^k).$$

Adding these expressions, we get $f_0(\hat{x}) \le f_0(x^k) - M\sqrt{H_k} + Nh(x^k)$. It is immediate to check that for $k \in \mathcal{K}_2$ such that $\sqrt{h(x^k)} < M/(N+\alpha)$ (say, $k \in \mathcal{K}_3$), $f_0(\hat{x}) < f_0(x^k) - \alpha h(x^k) = \tilde{f}_0$, completing the proof. ∎

4. Improvements. In this section we present three improvements to our treatment. First, we improve the convergence analysis by showing that under hypotheses (H5) and (H6) the objective function values always converge, thus very much limiting the possibility of reaching nonstationary accumulation points, especially when (H7) is added. Second, we show how a small change in the master algorithm totally precludes the possibility of generating nonstationary accumulation points. Third, we discuss a simplified optimality step using the Jacobian matrices already calculated in the feasibility phase instead of the ones for z^k.

4.1. Convergence of the objective function values. We shall continue the analysis of sequences (x^k) generated by the algorithm in section 2. We start by showing that f_0 cannot grow much in a single iteration.

LEMMA 4.1. *Assume that hypothesis* (H5) *holds. Then there exists a constant* $M > 0$ *such that in any iteration* k,

$$f_0(x^{k+1}) \leq f_0(x^k) + Mh(x^k).$$

Proof. Note that $f_0(\cdot)$ can only grow in an h-iteration. From (2.9), there exists a constant $M > 0$ such that in any iteration k, $f_0(z^k) \leq f_0(x^k) + Mh(x^k)$. By construction, $f(x^{k+1}) \leq f(z^k)$, completing the proof. ∎

Now we show that f_0 cannot grow much in a sequence of iterations.

LEMMA 4.2. *Assume that hypotheses* (H5) *and* (H6) *hold. Consider a finite sequence of iterations* $I = \{\bar{k}, \bar{k}+1, \ldots, K\}$ *such that for* $k \in I$, $f^k \equiv f_0(x^k) \geq f_0(x^{\bar{k}})$,[1] *and let* $M > 0$ *be given by Lemma* 4.1. *Then*

$$f^K \leq f^{\bar{k}} + \frac{M}{\alpha} h(x^{\bar{k}}).$$

Proof. Let us denote $f^k = f_0(x^k)$, $h^k = h(x^k)$ for $k \in I$ and $\bar{h} = h^{\bar{k}}$. Let us also define the following values:

$$\begin{aligned}
\phi_0 &= f^{\bar{k}}, \\
\phi_1 &= \phi_0 + M\bar{h}, \\
\phi_2 &= \phi_1 + M(1-\alpha)\bar{h} = \phi_0 + [1 + (1-\alpha)]M\bar{h}, \\
\phi_j &= \phi_0 + \left(\sum_{i=0}^{j-1} (1-\alpha)^i \right) M\bar{h} \leq \phi_0 + \frac{M}{\alpha}\bar{h}.
\end{aligned}$$

We show the following: there exists an integer $J \leq K - \bar{k}$ such that the sequence has at least one element in each interval $[\phi_j, \phi_{j+1}]$, $j = 0, 1, \ldots, J$, and $f^K \in [\phi_J, \phi_{J+1}]$. Consequently f^K will be smaller than $\phi_0 + M\bar{h}/\alpha$.

First interval. The iteration $\bar{k}$ is an h-iteration. The pair $(\phi_0 - \alpha\bar{h}, (1-\alpha)\bar{h})$ enters the permanent filter, and hence $h^k \leq (1-\alpha)\bar{h}$ for $k = \bar{k}+1, \ldots, K$ and $f^{\bar{k}+1} \leq \phi_0 + M\bar{h} = \phi_1$ by Lemma 4.1.

Let k_0 be the largest $k \in I$ such that $f^k \leq \phi_1$ (several h-iterations and f_0-iterations may have occurred between $\bar{k}$ and k_0). If $k_0 = K$, then the proof is complete. Otherwise f^{k_0+1} will be in the second interval.

Second interval. The iteration k_0 is an h-iteration, and as for the first interval,

$$(f^{k_0} - \alpha h^{k_0}, (1-\alpha)h^{k_0}) \leq (\phi_1, (1-\alpha)^2\bar{h})$$

enters the filter. Hence $h^k \leq (1-\alpha)^2\bar{h}$ for $k = k_0+1, \ldots, K$ and $\phi_1 \leq f^{k_0+1} \leq \phi_1 + M(1-\alpha)\bar{h} = \phi_2$ by Lemma 4.1.

Following the same process, we detect an h-iteration k_1, the last in the second interval. If $k_1 = K$, the proof is complete. Otherwise f^{k_1+1} will be in the third interval, and so on until $f^{k_J} = f^K$ is obtained. Then $f^K \leq \phi_0 + M\bar{h}/\alpha$, completing the proof. ∎

We can now prove the main result in this analysis.

[1] Note that (f^k) is not necessarily increasing, but $\bar{k}$ is an h-iteration.

THEOREM 4.3. *Assume that hypotheses* (H5) *and* (H6) *hold. Then the sequence* $(f_0(x^k))$ *converges.*

Proof. Let us denote $f^k \equiv f_0(x^k)$ for $k \in \mathbb{N}$. The sequence (f^k) is bounded by hypothesis. We shall use the following fact, which is a simple exercise in sequences: Given a sequence (f^k) such that $\limsup(f^k) > \liminf(f^k) + \delta$, $\delta > 0$, it is possible to extract two subsequences $(f^k)_{k\in\mathcal{K}}$ and $(f^{k+j_k})_{k\in\mathcal{K}}$, $\mathcal{K} \subset \mathbb{N}$, such that for any $k \in \mathcal{K}$,

$$f^{k+j_k} \geq f^k + \delta,$$
$$f^{k+r} \geq f^k \text{ for } r = 1, \ldots, j_k.$$

In fact, to prove this fact it is enough to take a subsequence convergent to $\limsup(f^k)$ and associate with each index (say, l) the last index $l - j_l$ such that $f^{l-j_l} \leq f^l - \delta$, if it exists. For large l, the construction will always be well defined.

Assume by contradiction that $\limsup(f^k) > \liminf(f^k) + \delta$ for some $\delta > 0$, and let the subsequence $(f^k)_{k\in\mathcal{K}}$ be given by the construction above. Then from Lemma 4.2, we conclude that for all $k \in \mathcal{K}$, the iteration k is an h-iteration and

$$f^k + \delta \leq f^{k+j_k} \leq f^k + \frac{M}{\alpha} h(x^k). \tag{4.1}$$

Taking subsequences if necessary, assume that $(x^k)_{k\in\mathcal{K}}$ converges to a point $\bar{x}$. Then $\bar{x}$ must be stationary by Lemma 2.5, and consequently $h(x^k) \xrightarrow{\mathcal{K}} 0$. This contradicts (4.1), completing the proof. ∎

Now we incorporate hypothesis (H7) and show that near a feasible nonstationary point the objective function always changes by a large amount, precluding the possibility of feasible nonstationary accumulation points.

LEMMA 4.4. *Assume that hypotheses* (H5)–(H7) *hold. Let* $\bar{x} \in X$ *be a feasible nonstationary point. Then there exist a neighborhood* V *of* $\bar{x}$ *and* $\delta > 0$ *such that whenever* $x^k \in V$*, there exists* $l_k \in \mathbb{N}$ *such that*

$$f_0(x^{k+l_k}) \leq f_0(x^k) - \delta. \tag{4.2}$$

Proof. Lemma 2.8 implies (H4). From (H4) and Lemma 2.9, there exist a neighborhood V_1 of $\bar{x}$ and constants $\beta_1, \beta_2 > 0$ such that for all $x^k \in V_1$,

$$f_0(x^k) - f_0(x^{k+1}) \geq \beta_1 \, \|x^{k+1} - x^k\|, \tag{4.3}$$
$$f_0(x^k) - f_0(x^{k+1}) \geq \beta_2 \, \sqrt{H_k}, \tag{4.4}$$

and the iteration k is an f_0-iteration.

Consider $\epsilon > 0$ such that $\mathcal{B}_\epsilon(\bar{x}) = \{x \in \mathbb{R}^n \mid \|x - \bar{x}\| < \epsilon\} \subset V_1$, and define $V = \mathcal{B}_{\epsilon/2}(\bar{x})$.

Let $k \in \mathbb{N}$ be such that $x^k \in V$. While x^{k+i}, $i = 1, 2, \ldots$, remain in $\mathcal{B}_\epsilon(\bar{x})$, the iterations $(k + i)$ are f_0-iterations, and the filter does not change, i.e.,

$$F_{k+i} = F_k \qquad \text{and} \qquad \mathcal{F}_{k+i} = \mathcal{F}_k \qquad \text{for } i = 1, 2, \ldots.$$

Consequently, from (4.4) f_0 decreases by at least the constant amount $\beta_2\sqrt{H_k}$. Hence, there exists a finite $l_k \in \mathbb{N}$ such that $x^{k+l_k} \notin \mathcal{B}_\epsilon(\bar{x})$, $x^{k+i} \in \mathcal{B}_\epsilon(\bar{x})$ for $i = 0, 1, \ldots, l_k - 1$. We have

$$\|x^{k+l_k} - x^k\| \geq \frac{\epsilon}{2} \tag{4.5}$$

because $x^k \in \mathcal{B}_{\epsilon/2}(\bar{x})$. Using (4.3), (4.5), and the triangle inequality,

$$\begin{aligned} f_0(x^k) - f_0(x^{k+l_k}) &= \sum_{i=0}^{l_k-1} f_0(x^{k+i}) - f_0(x^{k+i+1}) \\ &\geq \beta_1 \sum_{i=0}^{l_k-1} \|x^{k+i+1} - x^{k+i}\| \\ &\geq \beta_1 \|x^k - x^{k+l_k}\| \\ &\geq \beta_1 \epsilon/2, \end{aligned}$$

completing the proof. ∎

It is now trivial to prove (and this will be done in a moment) that feasible nonstationary points cannot be accumulation points of the sequence. The presence of nonstationary accumulation points is then reduced to a single seemingly unreasonable possibility: there must exist an infeasible accumulation point, reached by large jumps from points arbitrarily near a stationary solution, and the objective values must converge. We now show how a simple change in the algorithm precludes this possibility.

4.2. The modified algorithm. The only change is in the criterion used to introduce a point in the filter, which now becomes the following:

Filter update: Given $\varepsilon > 0$

if $f_0(x^{k+1}) < f_0(x^k) - \min\{(h(x^k))^2, \varepsilon\}$, then
  $F_{k+1} = F_k, \quad \mathcal{F}_{k+1} = \mathcal{F}_k$ (f_0-iteration)
else
  $F_{k+1} = \bar{F}_k, \quad \mathcal{F}_{k+1} = \bar{\mathcal{F}}_k$ (h-iteration)

This implies that potentially more points will be introduced in the filter.

Let us now study the sequence (x^k) generated by an application of the modified algorithm. Near feasible nonstationary points, the criterion for entering the filter becomes

$$f_0(x^k) - f_0(x^{k+1}) \leq (h(x^k))^2 = o(h(x^k)) = o(H_k).$$

The term $o(H_k)$ vanishes when added to $f_0(x^k) - f_0(x^{k+1}) = \Omega(\sqrt{H_k})$, and hence Lemma 2.5 remains true. Lemma 2.6 and Theorem 2.7 also remain true; it is immediate to check that the same proofs apply to the modified algorithm. Hence all the results of section 2 remain valid, and the sequence has a stationary accumulation point.

THEOREM 4.5. *Assume that hypotheses* (H1)–(H3) *and* (H5)–(H7) *hold. Then any accumulation point of* (x^k) *is stationary.*

Proof. By contradiction, assume that $x^k \xrightarrow{\mathcal{K}} \bar{x}$, $\bar{x}$ nonstationary, $\mathcal{K} \subset \mathbb{N}$. From Lemma 2.5, we know that for large $k \in \mathcal{K}$ all iterations are f_0-iterations.

If $h(\bar{x}) > 0$, then for large $k \in \mathcal{K}$, $h(x^k) > h(\bar{x})/2$ and hence $f_0(x^{k+1}) \leq f_0(x^k) - \delta_1$, with $\delta_1 = \min\{\varepsilon, (h(\bar{x}))^2/4\} > 0$.

If $h(\bar{x}) = 0$, then Lemma 4.4 ensures that for large $k \in \mathcal{K}$, there exist $\delta_2 > 0$ and $j_k \in \mathbb{N}$ such that

$$f_0(x^{k+j_k}) \leq f_0(x^k) - \delta_2.$$

In any case, we construct a subsequence $(x^{k+j_k})_{k\in\mathcal{K}}$ such that for $k \in \mathcal{K}$

$$f_0(x^{k+j_k}) \leq f_0(x^k) - \delta,$$

where $\delta = \min\{\delta_1, \delta_2\} > 0$.

It follows that the sequence $f_0(x^k)$ is not a Cauchy sequence, contradicting Theorem 4.3 and completing the proof. ∎

4.3. The simplified tangential step. In our algorithm the feasibility and tangential steps are independent. This means that the Jacobians $A_{\mathcal{E}}$ and $A_{\mathcal{I}}$ must be calculated both at x^k and z^k. In most algorithms based on feasibility and optimality steps, the tangential step uses at z^k the linear model computed at x^k, reducing the computations.

This makes sense if x^k is near z^k and if the feasibility algorithm has taken only one step to reach z^k from x^k. If multiple steps were used, then the tangential step can be simplified by approximating the Jacobians by the last ones computed in the feasibility procedure.

We shall now change the tangential step and use the following maps, which associate with each $(z, x) \in \mathbb{R}^{2n}$ the set

$$L(z,x) = \{y \in \mathbb{R}^n \mid A_{\mathcal{E}}(x)\,(y-z) = 0,\ f_{\mathcal{I}}(z) + A_{\mathcal{I}}(x)\,(y-z) \le f_{\mathcal{I}}^+(z)\} \tag{4.6}$$

and the point

$$d_c(z,x) = P_{L(z,x)}(z - \nabla f_0(z)) - z. \tag{4.7}$$

So, $L(z^k, x^k)$ is the same as $L(z^k)$ given by (1.4), with $A_{\mathcal{E}}(z^k)$, $A_{\mathcal{I}}(z^k)$ replaced by $A_{\mathcal{E}}(x^k)$, $A_{\mathcal{I}}(x^k)$. Similarly, the projected gradient direction is now projected into $L(z^k, x^k)$.

When $x^k \xrightarrow{\mathcal{K}} \bar{x}$, $\bar{x} \in X$ feasible, $\mathcal{K} \subset \mathbb{N}$, it is also true that $z^k \xrightarrow{\mathcal{K}} \bar{x}$, because by (2.8) $\|x^k - z^k\| = O(h(x^k))$. So, we need the continuity of (4.6) and (4.7) at a pair $(\bar{x}, \bar{x})$: this is guaranteed by straightforward changes in the proof of Lemma A.1.

The main change in the treatment is in Lemma 3.3, which now becomes the following.

LEMMA 4.6. *For any $z, x \in X$ such that $\|z - x\| = O(h(x))$ and $d \in \mathbb{R}^n$ such that $(z + d) \in L(z, x)$,*

$$|h(z+d) - h(z)| = h(x)\,O(\|d\|) + O(\|d\|^2).$$

Proof. From (1.3), for any $z \in X$, $d \in \mathbb{R}^n$, $i = 0, 1, \ldots, m$,

$$f_i(z+d) - f_i(z) \le \nabla f_i(z)^T d + O(\|d\|^2).$$

By the Lipschitz condition on ∇f_i, $i = 0, 1, \ldots, m$,

$$\|\nabla f_i(z) - \nabla f_i(x)\| = O(\|z - x\|) = O(h(x))$$

and

$$\begin{aligned} \nabla f_i(z)^T d &= \nabla f_i(x)^T d + O(h(x))\,\|d\| \\ &= \nabla f_i(x)^T d + h(x)\,O(\|d\|). \end{aligned}$$

Now, proceeding as in the proof of Lemma 3.3, we get

$$\|f^+(z+d)\| = \|f^+(z)\| + h(x)\,O(\|d\|) + O(\|d\|^2),$$

completing the proof. ∎

Lemma 3.4 is not affected by the changes, and we only need to change the proof of Lemma 3.5.

The only place where Lemma 3.3 is used in the proof is in the expressions (3.14) and (3.15). We will show now that with a good choice of the value $\bar{\Delta}$ (defined in the beginning of the proof), (3.15) is still valid.

Let us modify (3.14), using Lemma 4.4:

$$\alpha H_k \leq h(z^k + d(z^k, 2\hat{\Delta})) - h(z^k) = h(x^k)\, O(\|d(z^k, 2\hat{\Delta})\|) + O(\|d(z^k, 2\hat{\Delta})\|^2).$$

Since $d(z^k, 2\hat{\Delta}) \leq 2\hat{\Delta}$ and $h(x^k) < H_k$,

$$H_k \leq cH_k\hat{\Delta} + O(\hat{\Delta}^2),$$

where $c > 0$ is a constant dependent only on $\bar{x}$. For a choice of $\bar{\Delta}$ so that $c\bar{\Delta} < 1/2$, we obtain

$$\frac{1}{2}H_k = O(\hat{\Delta}^2),$$

which implies (3.15).

From this point on, the proof is identical, proving that the simplified algorithm has the same convergence properties as the original one.

5. A graphical example. In this section we present a graphical example of the mechanics of the algorithms. Consider the bidimensional problem

$$\begin{array}{ll} \text{minimize} & x_2 \\ \text{subject to} & f(x) = x_2 + (2 + x_1)\cos(x_1) = 0. \end{array}$$

Figure 5.1 (as well as all figures to follow) shows the level curves of $h(x) = |f(x)|$ and a local minimizer. The figure on the right shows the pairs $(f(x), h(x))$.

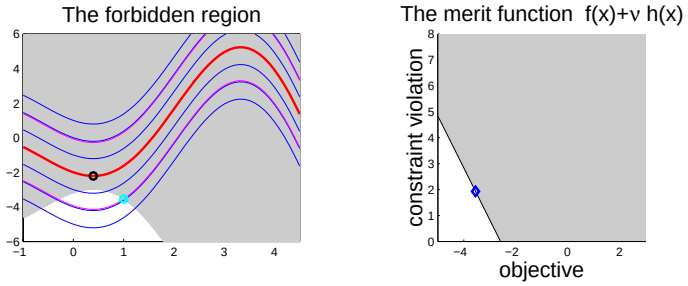


FIG. 5.1. *The merit function forbids the local optimizer.*

Using a merit function $\psi(x) = f(x) + \nu h(x)$ with $\nu = 0.5$, the figure shows the forbidden points associated with the point $(1, -4)$, i.e., the points x such that $\psi(x) \geq \psi((1, -4))$. Notice that the local optimizer is forbidden for this value of ν. This happens because ν is too small, smaller than the KKT multiplier at the optimum.

Figure 5.2 shows the same situation for $\nu = 1.5$, and now the local optimizer is never forbidden. This is actually true for any value of $\nu \geq 1$, the value of the optimal multiplier.

The following figures show some iterations of the filter method, programmed in Matlab and using internal algorithms which are intentionally imprecise but satisfy all

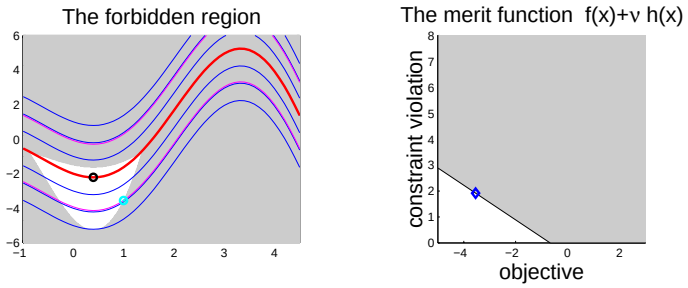


FIG. 5.2. *Now the optimizer is not forbidden.*

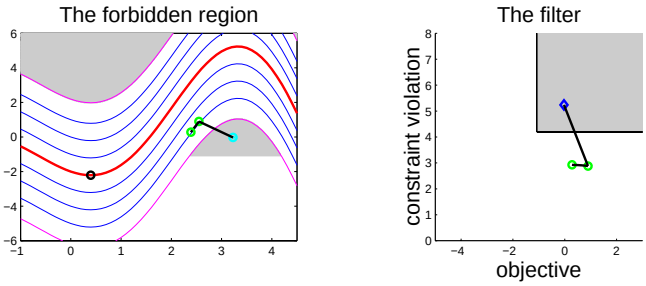


FIG. 5.3. *First iteration of a filter method.*

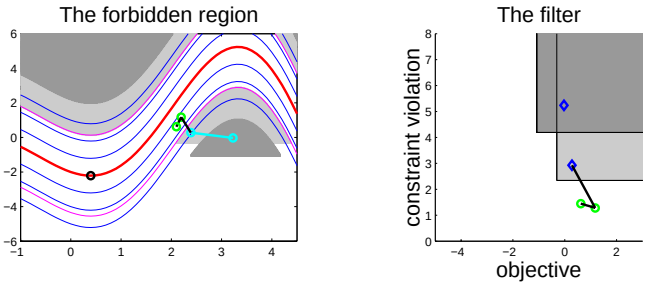


FIG. 5.4. *Temporarily and permanently forbidden points after the first iteration.*

the hypotheses. Figure 5.3 shows the first iteration. On the left are the temporarily forbidden region associated with the first iterate and a feasibility step followed by a tangential step. The figure on the right shows the filter: now $F_0 = \emptyset$, and $\bar{F}_0$ contains only the point $(f_0(x^0) - \alpha h(x^0), (1-\alpha)h(x^0))$. The pairs resulting from the feasibility and tangential steps are also shown. For the tangential step, we show the pairs corresponding to $z^k + \lambda(x^{k+1} - z^k)$, $\lambda \geq 0$.

The first iteration was an h-iteration, because $f_0(x^1) > f_0(x^0)$. So, $(f_0(x^0), h(x^0))$ becomes a permanent entry in the filter. Figure 5.4 shows the second iteration, where the permanently forbidden points and pairs are in the darker region.

The second iteration was also an h-iteration, and the permanent filter has two points. The third iteration is an f_0-iteration (see Figure 5.5).

After one more h-iteration (iteration 4), filter entries dominated by the new one can be eliminated from the filter. Figure 5.6 shows the fifth iteration.

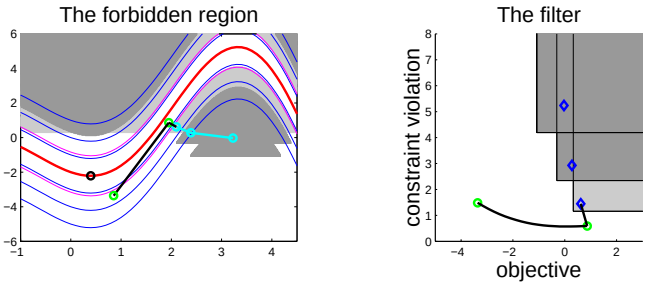


FIG. 5.5. *Third iteration: an f_0-iteration.*

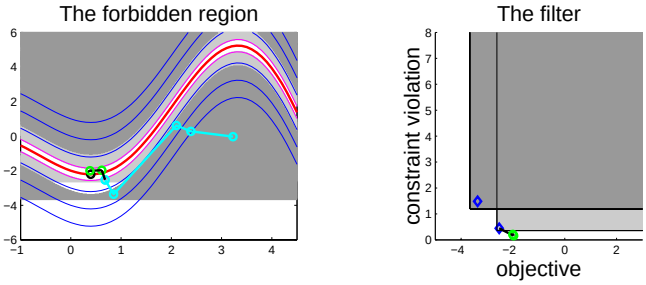


FIG. 5.6. *Fifth iteration: two filter elements were eliminated.*

Appendix. Continuity properties of maps. Let $L : \mathbb{R}^n \to \mathcal{P}(\mathbb{R}^n)$ be the map defined in (1.4),

$$z \in \mathbb{R}^n \mapsto L(z) = \{x \in \mathbb{R}^n \mid A_{\mathcal{E}}(z)(x - z) = 0, f_{\mathcal{I}}(z) + A_{\mathcal{I}}(z)(x - z) \leq f^+_{\mathcal{I}}(z)\},$$

where $z \mapsto A_{\mathcal{E}}(z)$ and $z \mapsto A_{\mathcal{I}}(z)$ are continuous.

We say that $x \in \mathbb{R}^n$ is an interior point of $L(z)$ if $x \in L(z)$ and $f_{\mathcal{I}}(z) + A_{\mathcal{I}}(z)(x - z) < f^+_{\mathcal{I}}(z)$.

LEMMA A.1. *Consider $\bar{z} \in \mathbb{R}^n$ such that $A_{\mathcal{E}}(\bar{z})$ has linearly independent rows and $L(\bar{z})$ has an interior point (i.e., the M-F qualification condition is satisfied at $\bar{z}$). Then the point to set map $L(\cdot)$ is continuous at $\bar{z}$.*

Proof. Consider a sequence (z^k) such that $z^k \to \bar{z}$ and the sets $L(z^k)$.

(1) Upper semicontinuity: Let $x^k \in L(z^k)$, $k \in \mathbb{N}$, be such that $x^k \to \bar{x}$. Using the continuity of all functions involved in the definition of $L(\cdot)$, the fact that $\bar{x} \in L(\bar{z})$ is straightforward.

(2) Lower semicontinuity: Consider an arbitrary point $\bar{x} \in L(\bar{z})$. We must exhibit a sequence $x^k \in L(z^k)$, $k \in \mathbb{N}$, such that $x^k \to \bar{x}$.

Define $x^k = P_{L(z^k)}(\bar{x})$, where $P_\Gamma(w)$ denotes the orthogonal projection of $w \in \mathbb{R}^n$ onto the closed set $\Gamma \subset \mathbb{R}^n$.

By contradiction assume that there exist an infinite set $\mathcal{K} \subset \mathbb{N}$ and $\varepsilon > 0$ such that for all $k \in \mathcal{K}$, $\|x^k - \bar{x}\| > \varepsilon$. We shall establish a contradiction by obtaining $k \in \mathcal{K}$ and a point $w^k \in L(z^k)$ such that $\|w^k - \bar{x}\| < \varepsilon$.

Let $y \in \mathbb{R}^n$ be an interior point of $L(\bar{z})$. Then for any $\lambda \in (0, 1)$,

$$y_\lambda = \lambda\, y + (1 - \lambda)\bar{x}$$

is an interior point of $L(\bar{z})$. Choose λ such that $\|y_\lambda - \bar{x}\| < \varepsilon/2$, and define w^k as the projection of y_λ onto $\{x \in \mathbb{R}^n \mid A_{\mathcal{E}}(z^k)(x - z^k) = 0\}$. For z^k sufficiently near $\bar{z}$,

$A_{\mathcal{E}}(z^k)$ has linearly independent rows, and the projection is given by

$$(y_\lambda - w^k) = A_{\mathcal{E}}(z^k)^T \left(A_{\mathcal{E}}(z^k)A_{\mathcal{E}}(z^k)^T\right)^{-1} A_{\mathcal{E}}(z^k)(y_\lambda - z^k).$$

The projection is continuous at $\bar{z}$, and hence $y_\lambda - w^k \to 0$, because $A_{\mathcal{E}}(\bar{z})(y_\lambda - \bar{z}) = 0$. From the continuity of $A_{\mathcal{I}}$ and $f_{\mathcal{I}}(\bar{z}) + A_{\mathcal{I}}(\bar{z})(y_\lambda - \bar{z}) < f^+_{\mathcal{I}}(\bar{z})$ and the facts that $z^k \overset{\mathcal{K}}{\to} \bar{z}$ and $w^k \overset{\mathcal{K}}{\to} y_\lambda$, for large $k \in \mathcal{K}$

$$f_{\mathcal{I}}(z^k) + A_{\mathcal{I}}(z^k)(w^k - z^k) < f^+_{\mathcal{I}}(z^k).$$

Thus, for large $k \in \mathcal{K}$, we have $w^k \in L(z^k)$ and $\|w^k - y_\lambda\| < \varepsilon/2$. For such w^k,

$$\|\bar{x} - w^k\| \le \|\bar{x} - y_\lambda\| + \|y_\lambda - w^k\| < \varepsilon,$$

completing the proof. ∎

LEMMA A.2. *Let the point to set map $z \in \mathbb{R}^n \mapsto L(z) \in \mathcal{P}(\mathbb{R}^n)$ and the function $z \in \mathbb{R}^n \mapsto p(z) \in \mathbb{R}^n$ be continuous at $\bar{z} \in \mathbb{R}^n$. Then $z \in \mathbb{R}^n \mapsto P_{L(z)}\,(p(z))$ is continuous at $\bar{z}$.*

Proof. Consider a sequence $z^k \to \bar{z} \in \mathbb{R}^n$, $x^k = P_{L(z^k)}\,(p(z^k))$. We must prove that $x^k \to \bar{x} = P_{L(\bar{z})}\,(p(\bar{z}))$.

From the lower semicontinuity of $L(\cdot)$, there exists a sequence $y^k \in L(z^k)$ such that $y^k \to \bar{x}$. By definition of projection,

$$\|p(z^k) - x^k\| \le \|p(z^k) - y^k\|. \tag{A.1}$$

Hence $(p(z^k) - x^k)$ is bounded, and consequently (x^k) is bounded. Consider $\tilde{x} \in \mathbb{R}^n$ and $\mathcal{K} \subset \mathbb{N}$ such that $(x^k) \overset{\mathcal{K}}{\to} \tilde{x}$. Using the upper semicontinuity of $L(\cdot)$, $\tilde{x} \in L(\bar{z})$, and hence by definition of projection,

$$\|\tilde{x} - p(\bar{z})\| \ge \|\bar{x} - p(\bar{z})\|.$$

Taking limits in (A.1) for $k \in \mathcal{K}$, $k \to \infty$,

$$\|p(\bar{z}) - \tilde{x}\| \le \|p(\bar{z}) - \bar{x}\|.$$

It follows that $\|p(\bar{z}) - \tilde{x}\| = \|p(\bar{z}) - \bar{x}\|$, and thus $\tilde{x} = \bar{x}$ by uniqueness of the projection onto a convex set. This proves that $\bar{x}$ is the unique accumulation point of (x^k), completing the proof. ∎

Acknowledgments. We thank Andreas Wächter for keenly finding some mistakes in our proofs and giving suggestions which helped us very much. We also thank José Mario Martínez for many suggestions. Finally, we thank an extremely dedicated and knowledgeable anonymous referee, whose suggestions led to great improvements in the paper.

REFERENCES

[1] J. ABADIE AND J. CARPENTIER, *Generalization of the Wolfe reduced-gradient method to the case of nonlinear constraints*, in Optimization, R. Fletcher, ed., Academic Press, New York, 1968, pp. 37–47.

[2] D. P. BERTSEKAS, *Nonlinear Programming*, Athena Scientific, Belmont, MA, 1995.

[3] R. H. BYRD, *Robust trust region methods for constrained optimization*, presented at the 3rd SIAM Conference on Optimization, Boston, MA, 1987.

[4] R. H. Byrd, J. C. Gilbert, and J. Nocedal, *A trust region method based on interior point techniques for nonlinear programming*, Math. Program., 89 (2000), pp. 149–185.
[5] R. H. Byrd, M. E. Hribar, and J. Nocedal, *An interior point algorithm for large-scale nonlinear programming*, SIAM J. Optim., 9 (1999), pp. 877–900.
[6] 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., SIAM, Philadelphia, 1985, pp. 71–82.
[7] R. Fletcher, N. I. M. Gould, S. Leyffer, P. L. Toint, and A. Wächter, *Global convergence of trust-region SQP-filter algorithm for general nonlinear programming*, SIAM J. Optim., 13 (2002), pp. 635–659.
[8] R. Fletcher and S. Leyffer, *Nonlinear programming without a penalty function*, Math. Program., 91 (2002), pp. 239–269.
[9] J. M. Martínez, *Inexact-restoration method with Lagrangian tangent decrease and new merit function for nonlinear programming*, J. Optim. Theory Appl., 111 (2001), pp. 39–58.
[10] J. M. Martínez and E. A. Pilotta, *Inexact restoration algorithms for constrained optimization*, J. Optim. Theory Appl., 104 (2000), pp. 135–163.
[11] J. M. Martínez and E. A. Pilotta, *Inexact restoration methods for nonlinear programming: Advances and perspectives*, in Optimization and Control with Applications, L. Q. Qi, K. L. Teo, and X. Q. Yang, eds., Kluwer Academic, Dordrecht, The Netherlands, 2001.
[12] J. M. Martínez and B. F. Svaiter, *A Practical Optimality Condition without Constraint Qualifications for Nonlinear Programming*, Tech. rep., Institute of Mathematics, University of Campinas, Brazil, 2001.
[13] J. Nocedal and S. J. Wright, *Numerical Optimization*, Springer Series in Operations Research, Springer-Verlag, New York, 1999.
[14] E. Omojokun, *Trust Region Algorithms for Optimization with Nonlinear Equality and Inequality Constraints*, Ph.D. thesis, Department of Computer Science, University of Colorado, Boulder, CO, 1991.
[15] M. J. D. Powell, *Convergence properties of a class of minimization algorithms*, in Nonlinear Programming 2, O. L. Mangasarian, R. R. Meyer, and S. M. Robinson, eds., Academic Press, New York, 1975, pp. 1–27.
[16] J. B. Rosen, *The gradient projection method for nonlinear programming.* I. *Linear constraints*, J. Soc. Indust. Appl. Math., 8 (1960), pp. 181–217.
[17] M. Ulbrich, S. Ulbrich, and L. N. Vicente, *A Globally Convergent Primal-Dual Interior-Point Filter Method for Nonconvex Nonlinear Programming*, Tech. rep. 00-11, Department of Mathematics, University of Coimbra, Portugal, 2000.
[18] A. Wächter and L. T. Biegler, *Global and Local Convergence of Line Search Filter Methods for Nonlinear Programming*, Tech. rep. B-01-09, CAPD, Department of Chemical Engineering, Carnegie Mellon University, Pittsburgh, PA, 2001.

SIAM J. OPTIM.
Vol. 14, No. 3, pp. 670–698

SOLVING LARGE SCALE SEMIDEFINITE PROGRAMS VIA AN ITERATIVE SOLVER ON THE AUGMENTED SYSTEMS*

KIM-CHUAN TOH†

Abstract. The search directions in an interior-point method for large scale semidefinite programming (SDP) can be computed by applying a Krylov iterative method to either the Schur complement equation (SCE) or the augmented equation. Both methods suffer from slow convergence as interior-point iterates approach optimality. Numerical experiments have shown that a diagonally preconditioned conjugate residual method on the SCE typically takes a huge number of steps to converge. However, it is difficult to incorporate cheap and effective preconditioners into the SCE. This paper proposes to apply the preconditioned symmetric quasi-minimal residual (PSQMR) method to a reduced augmented equation that is derived from the augmented equation by utilizing the eigenvalue structure of the interior-point iterates. Numerical experiments on SDP problems arising from maximum clique and selected SDPLIB (SDP Library) problems show that moderately accurate solutions can be obtained with a modest number of PSQMR steps using the proposed preconditioned reduced augmented equation. An SDP problem with 127600 constraints is solved in about 6.5 hours to an accuracy of 10^{-6} in relative duality gap.

Key words. large scale semidefinite programming, interior-point methods, augmented systems, conjugate residual method, symmetric quasi-minimal residual method, preconditioners, maximum-clique problem

AMS subject classification. 90C05

DOI. 10.1137/S1052623402419819

1. Introduction.

Let $\mathcal{S}^n$ be the vector space of $n \times n$ real symmetric matrices endowed with the inner product $A \bullet B = \text{Trace}(AB)$. Given a positive integer n, we let $\bar{n} = n(n+1)/2$. We use the notation $X \succeq 0$ ($X \succ 0$) to denote that X is symmetric positive semidefinite (symmetric positive definite). Given $k \times l$ matrices G, H, we define the linear map $G \circledast H : \mathcal{S}^l \to \mathcal{S}^k$ by $G \circledast H(M) = (HMG^T + GMH^T)/2$ for $M \in \mathcal{S}^l$.

Consider the standard primal semidefinite program (SDP)

$$\begin{array}{rrcl} \min_X, & C \bullet X, & & \\ & \mathcal{A}(X) & = & b, \\ & X & \succeq & 0, \end{array} \tag{1}$$

where $\mathcal{A} : \mathcal{S}^n \to \mathbb{R}^m$ is the linear map defined by

$$\mathcal{A}(X) \;=\; [A_1 \bullet X \;\; \cdots \;\; A_m \bullet X]^T .$$

Here $b \in \mathbb{R}^m$ and $A_1, \ldots, A_m$,$C \in \mathcal{S}^n$ are given data. The dual of (1) is

$$\begin{array}{rl} \max_{y,Z}, & b^T y, \\ & \mathcal{A}^T y + \;\; Z \;=\; C, \\ & Z \succeq 0, \end{array} \tag{2}$$

*Received by the editors December 14, 2002; accepted for publication (in revised form) July 15, 2003; published electronically December 19, 2003. This research was supported in part by National University of Singapore Research grant R-146-000-032-112 and the Hitachi Scholarship Foundation while the author was visiting the Tokyo Institute of Technology.

http://www.siam.org/journals/siopt/14-3/41981.html

†Department of Mathematics, National University of Singapore, 2 Science Drive 2, Singapore 117543 (mattohkc@math.nus.edu.sg) and Singapore-MIT Alliance, 4 Engineering Drive 3, Singapore 117576.

where $\mathcal{A}^T : \mathbb{R}^m \to \mathcal{S}^n$ is the adjoint of $\mathcal{A}$ defined by

$$\mathcal{A}^T y \;=\; \sum_{k=1}^{m} y_k A_k.$$

In this paper, we assume that (1) and (2) are strictly feasible, and the set $\{A_1, \ldots, A_m\}$ is linearly independent in $\mathcal{S}^n$.

We consider primal–dual path-following methods [31, 34] for SDP using the Nesterov–Todd direction in which the general framework in each iteration is as follows. Given a current iterate (X, y, Z) and a centering parameter $\sigma \in [0, 1)$, where $X, Z \succ 0$, the methods find a search direction $(\Delta X, \Delta y, \Delta Z) \in \mathcal{S}^n \times \mathbb{R}^m \times \mathcal{S}^n$ so as to generate the next iterate by solving the following linear system of equations:

$$\mathcal{A}\Delta X = R_p \;:=\; b - \mathcal{A}X, \tag{3a}$$

$$\mathcal{A}^T\Delta y \,+\, \Delta Z = R_d \;:=\; C - Z - \mathcal{A}^T y, \tag{3b}$$

$$\mathcal{E}\Delta X \,+\, \mathcal{F}\Delta Z = R_c \;:=\; \sigma\mu I - \Sigma^2, \tag{3c}$$

where $\mu = X \bullet Z/n$, $\mathcal{E} = G^{-T} \circledast GZ$, and $\mathcal{F} = G^{-T}X \circledast G$. Here G is the unique matrix such that $\Sigma := GZG^T = G^{-T}XG^{-1}$ is a positive definite diagonal matrix. Note that $W := G^TG$ is the NT scaling matrix such that $WZW = X$; see [31]. Instead of solving (3a)–(3c) directly, one can substitute $\Delta Z = R_d - \mathcal{A}^T\Delta y$ from (3b) into (3c) and solve the following augmented system:

$$-\mathcal{U}\Delta X + \mathcal{A}^T\Delta y = \mathcal{R} \;:=\; R_d - \mathcal{F}^{-1}R_c \;=\; R_d - \sigma\mu X^{-1} + Z, \tag{4a}$$

$$\mathcal{A}\Delta X = R_p, \tag{4b}$$

where $\mathcal{U} := \mathcal{F}^{-1}\mathcal{E} = W^{-1} \circledast W^{-1}$.

One can further eliminate ΔX from the augmented system above by substituting $\Delta X = \mathcal{U}^{-1}(\mathcal{A}^T\Delta y - R_d + \mathcal{F}^{-1}R_c)$ from (4a) into (4b) to obtain the following Schur complement equation (SCE) involving only Δy:

$$\underbrace{\mathcal{A}\mathcal{U}^{-1}\mathcal{A}^T}_{M}\Delta y = h \;:=\; R_p + \mathcal{A}\mathcal{U}^{-1}R_d - \mathcal{A}\mathcal{E}^{-1}R_c. \tag{5}$$

The $m \times m$ matrix M is known as the Schur complement matrix, and its (i, j) element is given by $M_{ij} = A_i \bullet WA_jW$. Most implementations of interior-point methods for SDP use (5) to compute the search direction. Generally, (5) is solved by a direct method by first computing and storing the matrix M and then computing its Cholesky factorization to find Δy. Substantial reduction in the cost of computing M is possible when the SDP data is sparse; see [13] for the details. However, M is generally fully dense even when the data is sparse. Thus when m is larger than a few thousands, it is impossible to store M in the memory of most current workstations. Furthermore, the $m^3/3$ flops required to compute the Cholesky factor of M also becomes prohibitively expensive. Consequently, when m is large, it is extremely difficult to solve (5) by a direct method, and a Krylov subspace iterative method such as the preconditioned conjugate gradient (PCG) or preconditioned conjugate residual (PCR) method becomes necessary, as these methods do not require M to be stored explicitly.

Earlier research works on using the PCG or PCR method to solve the SCE arising from large scale SDPs include [8, 20, 22, 23, 35]. As the coefficient matrix M is

dense, traditional preconditioning techniques that are designed for sparse matrices, such as incomplete Cholesky factorizations, cannot be readily applied to M without incurring a significant computational cost and memory usage. Thus in all the above-mentioned papers, except [20], only simple preconditioners such as diagonal or block-diagonal preconditioners were used. In [20], attempts had been made to use incomplete Cholesky factors as preconditioners, but no substantial improvement over diagonal preconditioners was observed. The preconditioners just mentioned are ineffective when the Schur complement matrix becomes increasingly ill-conditioned as the interior-point iterates approach an optimal solution. As a result, in all these works, only low accuracies in the duality gap can be achieved at reasonable costs.

The difficulties in constructing cheap and effective preconditioners for the SCE lead one to believe that second-order methods like those presented in [1, 18, 21, 31, 34] are too expensive for large scale SDPs. Thus, despite the success of second-order methods in solving small and medium size SDPs, attention has been diverted to first-order methods for large scale SDPs. Currently, there are three main classes of first-order methods. In [16], the dual SDP (2) was first formulated as a nonsmooth convex optimization problem and was solved by a spectral bundle (SB) method based on standard nonsmooth optimization techniques. On the other hand, Burer, Monterio, and Zhang [4] converted the dual SDP into a nonconvex nonlinear program in $\mathbb{R}^n_{++} \times \mathbb{R}^m$ and used a log-barrier method to solve the resulting nonlinear program. The third class of first-order methods [3] is based on the primal SDP (1). In this class of methods, the primal positive semidefinite constraint $X \succeq 0$ is eliminated by employing the factorization $X = VV^T$ for some matrix $V \in \mathbb{R}^{n\times p}$, where p is an estimate on the rank of an optimal primal solution. Such a technique transforms (1) into a nonlinear nonconvex program. In [3], an infeasible first-order augmented Lagrangian method (called BMPR method) is used to solve the resulting nonlinear program.

However, there are recent advances in using second-order methods to solve large SDPs. In [30], Toh and Kojima constructed preconditioners for the SCE based on orthogonal projectors derived from the eigenvalue structure of W. It was shown that these preconditioners can improve the convergence rate of the PCR method substantially in solving the SCE. However, each preconditioning step is rather expensive. Furthermore, the construction of these preconditioners requires the computation of a dense $\bar{p} \times \bar{p}$ matrix and its Cholesky factorization. Though $\bar{p}$ is generally a few times smaller than m, it does grow proportionately with m, and when m is very large, computing these preconditioners will require excessive memory space and time. Such a drawback poses a limit on the size of SDPs one can solve using these preconditioners. In [14], Fukuda, Kojima, and Shida used a predictor-corrector approach to numerically trace the central path in the space of Lagrange multipliers. The method uses the BFGS quasi-Newton method in the corrector procedure to locate points on the central path and the PCG method with BFGS preconditioners to solve a Schur complement-type equation in the predictor procedure. Preliminary numerical results on small SDPs show that this approach is promising for solving large scale SDPs, but careful numerical implementations have yet to be done to actualize this goal.

We should mention that in a primal–dual interior-point method, memory problems can also occur when n is large, since the primal variable X is typically dense even if the SDP data and the resulting dual variable Z are sparse. However, the root cause of this problem lies in the primal-dual framework used to solve the SDP, and it cannot be easily overcome by simply using an iterative method to compute the search direction. For such a problem, it is more appropriate to use methods, such as the

dual scaling method in [7], that avoid the need to form X explicitly. Another method that can alleviate the memory demand of the primal variable is the matrix completion method proposed in [15]. However, the implementation of the latter method is more complex than the dual scaling method.

In this paper, we will mainly focus on SDPs where m is large but n is moderate, say, less than 1000. We propose an efficient preconditioned iterative method to solve the augmented system (4a)–(4b). Like the SCE, the augmented system also suffers from ill-conditioning as the interior-point iterates approach optimality. We overcome the ill-conditioning problem by transforming the original augmented system into a better-conditioned reduced augmented system based on a newly developed block preconditioning technique in [29]. The basic idea is to analyze the eigenvalue structure of the (1,1) block $\mathcal{U}$ of the augmented system and eliminate the small eigenvalues by applying the technique in [29]. For SDP problems that are primal and dual nondegenerate and strict complementarity holds at optimality, the coefficient matrix of the reduced augmented system is shown to have a bounded condition number even as the interior-point iterates approach optimality. Like the Schur complement matrix, the reduced augmented matrix is dense even if the SDP data is sparse. Thus, to further improve the conditioning of the reduced augmented matrix without incurring significant computational and storage costs, we are restricted to consider only diagonal preconditioners. Fortunately, the class of diagonal preconditioners that are proposed in [26] for augmented systems arising from soil consolidation problems in civil engineering is also quite effective for our reduced augmented systems. To solve the reduced augmented system, we use the preconditioned symmetric quasi-minimal residual (PSQMR) method [12].

Because the cost of applying the PCR method to the SCE is typically two to three times cheaper than that of applying the PSQMR method to the reduced augmented system, it is desirable to use the SCE unless the Schur complement matrix is highly ill-conditioned. By using the hybrid approach of applying the PCR method to the SCE when interior-point iterates are not close to optimality and switching to the PSQMR method applied to the reduced augmented system when they are, we are able to solve some large SDPs arising from maximum clique problems of graphs, and selected SDPLIB problems [6] to moderately high accuracies, but at reasonable costs. Numerical experiments indicate that our method is promising in solving large SDPs. But there is a slight limitation in that our method cannot be adapted for the HRVW/KSH/M direction [17, 18, 21] for reasons that we will explain later.

The paper is organized as follows. In section 2, the derivation of the reduced augmented system is presented. The implementation of the PCR method for solving the SCE is given in section 3. The implementation of the PSQMR method for solving the reduced augmented system and the class of diagonal preconditioners used are presented in section 4. This is followed by numerical results in section 5 showing the effectiveness of the preconditioned reduced augmented systems on two collections of SDPs arising from maximum clique problems of graphs. Section 6 presents further numerical results for SDPs (selected SDPLIB problems and those arising from frequency assignment problems) whose degeneracies make them potentially ill-suited for computation via an iterative solver. In section 7, we conclude our paper.

We end this section by introducing some notations. We let $\mathbf{svec} : \mathcal{S}^n \to \mathbb{R}^{\bar{n}}$ be the isometry defined by

$$\mathbf{svec}(U) \;=\; \Big[U_{11},\ \sqrt{2}U_{12},\ U_{22},\ \sqrt{2}U_{13},\ \sqrt{2}U_{23},\ U_{33},\ \ldots\ ,\sqrt{2}U_{1n},\ \ldots\ ,U_{nn}\Big]^T. \tag{6}$$

Note that for $K, L \in \mathcal{S}^n$, $K \bullet L = \mathbf{svec}(K)^T \mathbf{svec}(L)$. We let $\mathbf{smat} : \mathbb{R}^{\bar{n}} \to \mathcal{S}^n$ be the inverse of $\mathbf{svec}$. We define the linear map $\mathbf{vec} : \mathbb{R}^{m\times n} \to \mathbb{R}^{mn}$ by

$$\mathbf{vec}(U) \;=\; [U_{11},\, \ldots\, , U_{m1},\, U_{12},\, \ldots\, , U_{m2},\;\; \ldots\;\; , U_{1n},\, \ldots\, , U_{mn}]^T \tag{7}$$

and $\mathbf{tvec} : \mathbb{R}^{n\times n} \to \mathbb{R}^{\bar{n}}$ by

$$\mathbf{tvec}(U) \;=\; [U_{11},\, U_{12},\, U_{22},\, U_{13},\, U_{23},\, U_{33},\, \ldots\, , U_{1n},\, \ldots\, , U_{nn}]^T\,. \tag{8}$$

For a vector d, $\mathrm{diag}(d)$ denotes the diagonal matrix with d as its diagonal. The MATLAB notation $[x\,;\,y]$ is used to denote the column vector formed by appending a column vector y to x. We use $\|\cdot\|$ to denote the vector and matrix 2-norms, and we use $\|\cdot\|_F$ to denote the Frobenius norm. The notation $\mathrm{diag}(P, Q)$ is used to denote the block-diagonal matrix with P and Q as its diagonal blocks. The condition number $\kappa(P)$ of a matrix P is defined to be the ratio between the largest and smallest singular values of P. For a vector $x \in \mathbb{R}^n_{++}$, the notation $x = \Theta(\epsilon)$ means that there exist positive constants γ_1, γ_2 such that $\gamma_1\epsilon \le x_i \le \gamma_2\epsilon$ for all $i = 1, \ldots, n$.

2. Reduced augmented system. Given an interior-point iterate (X, y, Z), let $\mu := X \bullet Z/n$ and W be the associated NT scaling matrix. Let $W^{-1} = QDQ^T$ be the eigenvalue decomposition of W^{-1}. Then the eigenvalue decomposition of $\mathcal{U}$ is given by

$$\mathcal{U} \;=\; (Q \circledast Q)(D \circledast D)(Q^T \circledast Q^T). \tag{9}$$

With the above decomposition, the augmented system (4a)–(4b) can be rewritten as follows:

$$-(D \circledast D)\,(Q^T \Delta X Q) \,+\, (Q^T \circledast Q^T)\mathcal{A}^T\,\Delta y = Q^T \mathcal{R} Q, \tag{10a}$$

$$\mathcal{A}(Q \circledast Q)\,(Q^T \Delta X Q) = R_p. \tag{10b}$$

Suppose (X, y, Z) is close to some optimal solution (X^*, y^*, Z^*) of the primal and dual SDP. If (X^*, Z^*) satisfies the strict complementarity condition defined in [2] (that is, $\mathrm{rank}(X^*) + \mathrm{rank}(Z^*) = n$), then as (X, Z) approaches this optimal solution (i.e., when μ is sufficiently small), the eigenvalues of W^{-1} will separate into two groups, one with small magnitude of the order $\Theta(\sqrt{\mu})$ and the other with large magnitude of the order $\Theta(1/\sqrt{\mu})$. Now suppose that W^{-1} has a group of p small eigenvalues and a group of $q := n - p$ large eigenvalues. Let the vector of small and large eigenvalues be d_1 and d_2, respectively. We can rewrite W^{-1} as

$$W^{-1} \;=\; Q_1 D_1 Q_1^T \,+\, Q_2 D_2 Q_2^T, \tag{11}$$

according to the partition $D = \mathrm{diag}(D_1, D_2)$ and $Q = [Q_1 \;\; Q_2]$, with $D_1 = \mathrm{diag}(d_1) \in \mathbb{R}^{p\times p}$, $Q_1 \in \mathbb{R}^{n\times p}$ corresponding to the small eigenvalues, and $D_2 = \mathrm{diag}(d_2) \in \mathbb{R}^{q\times q}$, $Q_2 \in \mathbb{R}^{n\times q}$ corresponding to the large eigenvalues. When μ is sufficiently small, the number of eigenvalues of W^{-1} with magnitudes $\Theta(\sqrt{\mu})$ is equal to the rank of X^*. Thus p is usually equal to the rank of X^*. In actual computation, however, we can set p to be any integer such that $\bar{p} \le m$, and it is not necessary to know the exact rank of X^*. Recall that by a theorem of Pataki [25], there exists an optimal solution X^* whose rank p satisfies the inequality $\bar{p} \le m$. Thus it is legitimate to choose p such that $\bar{p} \le m$ in actual computation.

Based on the eigenstructure of W^{-1}, we will now propose a method to overcome the ill-conditioning problem in the SCE and augmented equation when μ is small. We start from the augmented system (10a)–(10b) by diagonalizing $\mathcal{U}$ based on the eigenvalue decomposition of W^{-1}.

As a reminder, we have $d_1 = \mathrm{diag}(D_1)$ and $d_2 = \mathrm{diag}(D_2)$.

THEOREM 2.1. *With the partition in* (11)*, the augmented system* (10a)–(10b) *can be rewritten as*

(12)
$$\begin{bmatrix} -\mathcal{D}_{11} & & & \mathcal{B}_{11}^T \\ & -\mathcal{D}_{12} & & \mathcal{B}_{12}^T \\ & & -\mathcal{D}_{22} & \mathcal{B}_{22}^T \\ \mathcal{B}_{11} & \mathcal{B}_{12} & \mathcal{B}_{22} & \end{bmatrix} \begin{bmatrix} \mathbf{svec}(Q_1^T \Delta X Q_1) \\ \sqrt{2}\mathbf{vec}(Q_1^T \Delta X Q_2) \\ \mathbf{svec}(Q_2^T \Delta X Q_2) \\ \Delta y \end{bmatrix} = \begin{bmatrix} \mathbf{svec}(Q_1^T \mathcal{R} Q_1) \\ \sqrt{2}\mathbf{vec}(Q_1^T \mathcal{R} Q_2) \\ \mathbf{svec}(Q_2^T \mathcal{R} Q_2) \\ R_p \end{bmatrix},$$

where

$$\mathcal{B}_{11}^T = \left[\begin{array}{ccc} \mathbf{svec}(Q_1^T A_1 Q_1) & \cdots & \mathbf{svec}(Q_1^T A_m Q_1) \end{array} \right] \in \mathbb{R}^{\bar{p} \times m},$$
$$\mathcal{B}_{12}^T = \left[\begin{array}{ccc} \sqrt{2}\mathbf{vec}(Q_1^T A_1 Q_2) & \cdots & \sqrt{2}\mathbf{vec}(Q_1^T A_m Q_2) \end{array} \right] \in \mathbb{R}^{pq \times m},$$
$$\mathcal{B}_{22}^T = \left[\begin{array}{ccc} \mathbf{svec}(Q_2^T A_1 Q_2) & \cdots & \mathbf{svec}(Q_2^T A_m Q_2) \end{array} \right] \in \mathbb{R}^{\bar{q} \times m},$$

and

$$\mathcal{D}_{11} = \mathrm{diag}(\mathbf{tvec}(d_1 d_1^T)), \ \mathcal{D}_{12} = \mathrm{diag}(\mathbf{vec}(d_1 d_2^T)), \ \mathcal{D}_{22} = \mathrm{diag}(\mathbf{tvec}(d_2 d_2^T)).$$

Proof. Using the fact that for any $U \in \mathcal{S}^n$,

$$Q^T U Q = \left[\begin{array}{cc} Q_1^T U Q_1 & Q_1^T U Q_2 \\ Q_2^T U Q_1 & Q_2^T U Q_2 \end{array} \right],$$

we get from (10a) the following equations:

$$-D_1(Q_1^T \Delta X Q_1) D_1 + \sum_{k=1}^m (Q_1^T A_k Q_1) \Delta y_k = Q_1^T \mathcal{R} Q_1,$$

$$-D_1(Q_1^T \Delta X Q_2) D_2 + \sum_{k=1}^m (Q_1^T A_k Q_2) \Delta y_k = Q_1^T \mathcal{R} Q_2,$$

$$-D_2(Q_2^T \Delta X Q_2) D_2 + \sum_{k=1}^m (Q_2^T A_k Q_2) \Delta y_k = Q_2^T \mathcal{R} Q_2.$$

It is readily shown that these three equations correspond to the first three block equations in (12). Now, from (10b), we have

$$\mathcal{A}(Q \circledast Q)(Q^T \Delta X Q) = \begin{bmatrix} (Q^T A_1 Q) \bullet (Q^T \Delta X Q) \\ \vdots \\ (Q^T A_m Q) \bullet (Q^T \Delta X Q) \end{bmatrix}$$

$$= \begin{bmatrix} (Q_1^T A_1 Q_1) \bullet (Q_1^T \Delta X Q_1) + 2(Q_1^T A_1 Q_2) \bullet (Q_1^T \Delta X Q_2) + (Q_2^T A_1 Q_2) \bullet (Q_2^T \Delta X Q_2) \\ \vdots \\ (Q_1^T A_m Q_1) \bullet (Q_1^T \Delta X Q_1) + 2(Q_1^T A_m Q_2) \bullet (Q_1^T \Delta X Q_2) + (Q_2^T A_m Q_2) \bullet (Q_2^T \Delta X Q_2) \end{bmatrix}$$

$$= \mathcal{B}_{11}\mathbf{svec}(Q_1^T \Delta X Q_1) + \sqrt{2}\,\mathcal{B}_{12}\mathbf{vec}(Q_1^T \Delta X Q_2) + \mathcal{B}_{22}\mathbf{svec}(Q_2^T \Delta X Q_2).$$

This corresponds to the last block equation in (12). ∎

Through (10a)–(10b), the system (12) is orthogonally equivalent to the augmented system (4a)–(4b), and thus the condition numbers of the coefficient matrices are the same. To improve the conditioning of (12), we apply the block splitting introduced in [29] to (12) to get a smaller reduced augmented system as shown in the next theorem.

THEOREM 2.2. *Let $\beta \in \mathbb{R}^p$ be a given positive vector. Suppose*

$$E_{11} = \mathrm{diag}(\mathbf{tvec}(\beta\beta^T + \beta d_1^T + d_1\beta^T)), \tag{13}$$

$$S_{11} := \mathcal{D}_{11} + E_{11} = \mathrm{diag}(\mathbf{tvec}((d_1+\beta)(d_1+\beta)^T)). \tag{14}$$

The augmented system (12) *can be solved via the following reduced augmented system:*

$$\underbrace{\begin{bmatrix} \mathcal{H} & \mathcal{B}_{11}S_{11}^{-1/2} \\ S_{11}^{-1/2}\mathcal{B}_{11}^T & -\Psi \end{bmatrix}}_{\mathcal{K}} \begin{bmatrix} \Delta y \\ S_{11}^{-1/2}E_{11}\mathbf{svec}(Q_1^T \Delta X Q_1) \end{bmatrix}$$

$$= \begin{bmatrix} R_p + \mathcal{B}\mathrm{diag}(S_{11}^{-1},\, \mathcal{D}_{12}^{-1},\, \mathcal{D}_{22}^{-1})\mathbf{svec}(Q^T \mathcal{R} Q) \\ S_{11}^{-1/2}\mathbf{svec}(Q_1^T \mathcal{R} Q_1) \end{bmatrix}, \tag{15}$$

where $\mathcal{B} = [\mathcal{B}_{11}\ \mathcal{B}_{12}\ \mathcal{B}_{22}]$, *and*

$$\mathcal{H} = \mathcal{B}\,\mathrm{diag}(S_{11}^{-1},\, \mathcal{D}_{12}^{-1},\, \mathcal{D}_{22}^{-1})\,\mathcal{B}^T, \quad \Psi = \mathcal{D}_{11}E_{11}^{-1}.$$

Note that

$$\mathcal{H} = \mathcal{A}(P_1 \circledast P_1)\mathcal{A}^T + \mathcal{A}((2P_2 + P_3) \circledast P_3)\mathcal{A}^T, \tag{16}$$

where

$$P_1 = Q_1\,\mathrm{diag}(\beta + d_1)^{-1}Q_1^T, \quad P_2 = Q_1\,D_1^{-1}Q_1^T, \quad P_3 = Q_2\,D_2^{-1}Q_2^T.$$

Note that once Δy *and* $Q_1^T \Delta X Q_1$ *are computed,* $Q^T \Delta X Q$ *can be computed as follows:*

$$Q_1^T \Delta X Q_2 = D_1^{-1}\left(Q_1^T(\mathcal{A}^T\Delta y - \mathcal{R})Q_2\right) D_2^{-1}, \tag{17}$$

$$Q_2^T \Delta X Q_2 = D_2^{-1}\left(Q_2^T(\mathcal{A}^T\Delta y - \mathcal{R})Q_2\right) D_2^{-1}. \tag{18}$$

Proof. The derivation of (15) follows readily by applying Theorem 2.1 in [29] to the system in (12). Next we will derive (16). Note that

$$\mathcal{H} = \mathcal{B}_{11}\mathrm{diag}(S_{11}^{-1})\mathcal{B}_{11}^T + \mathcal{B}_{12}\mathrm{diag}(\mathcal{D}_{12}^{-1})\mathcal{B}_{12}^T + \mathcal{B}_{22}\mathrm{diag}(\mathcal{D}_{22}^{-1})\mathcal{B}_{22}^T.$$

We shall just show that $\mathcal{B}_{11}\mathrm{diag}(S_{11}^{-1})\mathcal{B}_{11}^T = \mathcal{A}(P_1 \circledast P_1)\mathcal{A}^T$, and it is easy to simplify the other two terms similarly. For any $v \in \mathbb{R}^m$, we have

$$\mathcal{B}_{11}^T v = \mathbf{svec}(G),$$

$$\mathrm{diag}(S_{11}^{-1})\mathcal{B}_{11}^T v = \mathbf{svec}(\Lambda G\,\Lambda),$$

where $G = Q_1^T(\mathcal{A}^T v)Q_1$ and $\Lambda = \mathrm{diag}(\beta + d_1)^{-1}$. Hence

$$\mathcal{B}_{11}\mathrm{diag}(S_{11}^{-1})\mathcal{B}_{11}^T v \;=\; \mathcal{B}_{11}\mathbf{svec}(\Lambda G\,\Lambda) \;=\; \begin{bmatrix} \mathbf{svec}(Q_1^T A_1 Q_1)^T \mathbf{svec}(\Lambda G\,\Lambda) \\ \vdots \\ \mathbf{svec}(Q_1^T A_m Q_1)^T \mathbf{svec}(\Lambda G\,\Lambda) \end{bmatrix}$$

$$= \begin{bmatrix} A_1 \bullet Q_1\Lambda Q_1^T(\mathcal{A}^T v)Q_1\Lambda Q_1^T \\ \vdots \\ A_m \bullet Q_1\Lambda Q_1^T(\mathcal{A}^T v)Q_1\Lambda Q_1^T \end{bmatrix} = \begin{bmatrix} A_1 \bullet P_1(\mathcal{A}^T v)P_1 \\ \vdots \\ A_m \bullet P_1(\mathcal{A}^T v)P_1 \end{bmatrix}$$

$$= \mathcal{A}(P_1 \circledast P_1)\mathcal{A}^T v.$$

Thus, we have derived the first term in (16). □

Notice that the reduced augmented matrix $\mathcal{K} \in S^{m+\bar{p}}$ in (15) is smaller in size compared to the augmented matrix in (12), whose dimension is $m + \bar{n}$. It is also potentially better conditioned, as shown in Theorem 2.4 below. Before we present that theorem, it is beneficial for us to recall the concept of primal and dual nondegeneracy introduced in [2].

THEOREM 2.3 (see [2]). *Suppose (X^*, Z^*) satisfies the strict complementarity condition. Then X^* is primal nondegenerate if and only if the matrix $[\mathcal{B}_{11}\ \mathcal{B}_{12}]$ has full row rank, and a necessary condition for primal nondegeneracy is $\bar{n} - \bar{q} \geq m$. The solution Z^* is dual nondegenerate if and only if the matrix $\mathcal{B}_{11}$ has full column rank, and a necessary condition for dual nondegeneracy is $\bar{p} \leq m$.*

Proof. The proof follows readily from Theorems 6 and 9 in [2]. □

THEOREM 2.4. *Under the assumption that (X^*, Z^*) satisfies the strict complementarity condition, and the primal and dual nondegeneracy conditions defined in* [2]*, the coefficient matrix $\mathcal{K}$ in* (15) *has a condition number that is bounded independent of μ (when μ is small).*

Proof. When μ is sufficiently small and (X, Z) is close to a strictly complementary optimal solution (X^*, Z^*) with $p = \mathrm{rank}(X^*)$, by Theorem 6 in [2], primal nondegeneracy implies that $[\mathcal{B}_{11}\ \ \mathcal{B}_{12}]$ has full row rank; and by Theorem 9 in [2], dual nondegeneracy implies that $\mathcal{B}_{11}$ has full column rank. By Theorem 3.2 in [29], the theorem follows. □

For an SDP that is primal and dual nondegenerate, and where the strict complementarity condition holds, Theorem 2.4 implies that one can expect a Krylov subspace method applied to (15) to have a better rate of convergence than one that is applied to (5). There is another advantage in using the reduced augmented system. Because the (1,1) and (1,2) blocks of $\mathcal{K}$ are not ill-conditioned, the task of constructing effective preconditioners for $\mathcal{K}$ is likely to be easier than that for the highly ill-conditioned matrix M.

Remark. (a) Notice that the derivation of the reduced augmented system (15) depends on our ability to find the eigenvalue decomposition of $W^{-1} \circledast W^{-1}$. For the HRVW/KSH/M direction described in [17, 18, 21], $W^{-1} \circledast W^{-1}$ is replaced by $(X \circledast Z^{-1})^{-1}$. Unfortunately, unlike the former, the eigenvalue decomposition of the latter is not readily available even if those of X and Z are known. For this reason, the augmented system (12) cannot be reduced to the form in (15) for the HRVW/KSH/M direction. However, for the dual scaling direction in [7], $W^{-1} \circledast W^{-1}$ is replaced by $Z \circledast Z$, and the corresponding reduced augmented system can be found readily once

the eigenvalue decomposition of Z is known.

(b) For the numerical experiments in this paper, the vector β in Theorem 2.2 is chosen to be

$$\beta \;=\; \max(1, \max(d_1))\,(1, 1, \ldots, 1)^T.$$

(c) Our reduced augmented system (15) can also be applied to interior-point methods for linear programs (LPs). Such a system can potentially produce a search direction that is numerically more accurate than that computed straightforwardly from the SCE (5). This topic is currently being investigated. Another method that had been proposed to overcome the stability/accuracy problems encountered when solving the SCE arising from LP is the stabilization method of Kovacevic-Vujcic and Asic [33]. The stabilization method in [33] is based on a novel pivoting strategy to avoid excessive loss of numerical accuracies due to the mixing of elements corresponding to large and small scaling factors when the Schur complement matrix M is factorized. As far as we know, the reduced augmented system approach and the stabilization method in [33] are not directly related, although both can be used to avoid excessive numerical errors for computing the search directions in interior-point methods for LP.

2.1. Nondegeneracy and condition number of the reduced augmented matrix. Now we present some examples to illustrate the validity of Theorem 2.4, as well as examples to demonstrate what may happen to the condition number $\kappa(\mathcal{K})$ when the nondegeneracy conditions in Theorem 2.4 do not hold. In order to know the ranks of X^* and Z^* unambiguously, we need to compute very accurate approximate optimal solutions. But it is well known that the standard approach of computing the search direction from (5) in each interior-point iteration usually does not deliver very accurate approximate optimal solutions because of highly ill-conditioned Schur complement matrices. Thus we have to rely on an alternative approach to compute the search directions.

It turns out that the approach of computing the directions from (15) via the LDL^T factorization of $\mathcal{K}$ can usually deliver more accurate approximate solutions than the standard approach. On a limited set of examples that we have tested, the accuracy gained is usually more than two digits in the infeasibilities and duality gap. Better accuracy is plausible because $\mathcal{K}$ is potentially better conditioned than M, and so the search direction computed via (15) is potentially more accurate than that computed from (5). When the assumption in Theorem 2.4 holds, the condition number of the coefficient matrix in (15) is bounded independent of μ. This implies that the unknowns Δy and $Q_1^T \Delta X Q_1$ can be computed accurately even when μ is small. From (18), it is easy to see that $Q_2^T \Delta X Q_2$ can be computed accurately since $d_2 = \Theta(1/\sqrt{\mu})$. From (17), we have

$$(Q_1^T \Delta X Q_2)_{ij} \;=\; \frac{\big(Q_1^T(\mathcal{A}^T \Delta y - \mathcal{R})Q_2\big)_{ij}}{d_1^{(i)} d_2^{(j)}};$$

thus $Q_1^T \Delta X Q_2$ can also be computed accurately since $d_1^{(i)} d_2^{(j)} = \Theta(\sqrt{\mu})\Theta(1/\sqrt{\mu}) = \Theta(1)$. Therefore ΔX can be computed accurately from $Q_1^T \Delta X Q_1$, $Q_1^T \Delta X Q_2$, and $Q_2^T \Delta X Q_2$. Finally, ΔZ can also be computed accurately from $\Delta Z = R_d - \mathcal{A}^T \Delta y$.

As our purpose in this paper is the application of iterative methods for solving large SDPs, we shall not discuss further the issue of solving an SDP via (15) by using

the LDL^T factorization. We leave this issue for a more detailed investigation in the future.

To illustrate the validity of Theorem 2.4, in Table 1 we give the condition number of $\mathcal{K}$ in (15) and M in (5) for some of the interior-point iterates generated by the SDP software, SDPT3 version 3.0 [32]. The SDP problem is the problem `theta2` (with $m = 498$ and $n = 100$) taken from the SDPLIB [6]. The default parameters in SDPT3 are used. But when μ is small, the search direction in each interior-point iteration is computed via (15) instead of via the system (5) that is implemented in SDPT3.

The table shows that $\kappa(\mathcal{K})$ is bounded at the level 2.5×10^6 when $\mu = X \bullet Z/n$ is approaching 0, while $\kappa(M)$ grows like $3 \times 10^4/\mu$. In the table,

$$\phi \;=\; \max\left(\frac{\|R_p\|}{1+\|b\|}, \frac{\|R_d\|_F}{1+\|C\|_F}\right). \tag{19}$$

The approximate optimal solution (X, y, Z) of `theta2` is strictly complementary, and it satisfies the necessary conditions in Theorem 2.3 for primal and dual nondegeneracy. Suppose the eigenvalues of X and Z are ordered in decreasing and increasing order, respectively. We have $\min_i\{\lambda_i(X) + \lambda_i(Z)\} = 3.2 \times 10^{-3}$, and $p = 16$, $q = 84$. For this problem, the matrix $[\mathcal{B}_{11}\ \mathcal{B}_{12}] \in \mathbb{R}^{m\times 1480}$ has singular values in the range $[0.1, 4.1]$, while those of $\mathcal{B}_{11} \in \mathbb{R}^{m\times 136}$ are contained in $[5 \times 10^{-2}, 4.1]$.

TABLE 1

Condition number of reduced augmented and Schur complement matrices corresponding to interior-point iterates generated by SDPT3 for the SDP problem `theta2`. *The approximate optimal solution has a relative duality gap of* 3.6e-14. *This SDP is primal and dual nondegenerate.*

Iteration	$X \bullet Z/n$	ϕ	$\kappa(\mathcal{K})$	$\kappa(\mathcal{H})$	$\kappa(\mathcal{B}_{11}S_{11}^{-1/2})$	$\kappa(M)$
13	2.8e-08	4.8e-15	2.0e+06	1.9e+06	6.8e+01	1.4e+12
14	3.0e-09	4.0e-16	2.4e+06	2.3e+06	6.8e+01	1.0e+13
15	2.5e-10	4.0e-16	2.5e+06	2.3e+06	6.9e+01	9.9e+13
16	5.9e-12	4.2e-16	2.5e+06	2.4e+06	6.9e+01	4.2e+14
17	1.7e-13	2.1e-16	2.4e+06	2.3e+06	6.9e+01	4.5e+14

Next we give an example to illustrate what may happen to $\kappa(\mathcal{K})$ when the conditions in Theorem 2.4 are not satisfied. For this purpose, we use the SDPLIB problem `qap6` (with $m = 229$ and $n = 37$). The approximate solution delivered by SDPT3 is strictly complementary with $\min_i\{\lambda_i(X) + \lambda_i(Z)\} = 5.0 \times 10^{-4}$, and we have $p = 12$ and $q = 25$. Although $\bar{n} - \bar{q} = 378 \geq m$ satisfies the necessary condition in Theorem 2.3 for primal nondegeneracy, the problem `qap6` is in fact nearly primal degenerate, because the matrix $[\mathcal{B}_{11}\ \mathcal{B}_{12}] \in \mathbb{R}^{m\times 378}$ has 13 small singular values that are in the range $[1 \times 10^{-5}, 7 \times 10^{-5}]$, while the rest are in the range $[1.2 \times 10^{-1}, 4.2 \times 10^{1}]$. Note that `qap6` is dual nondegenerate, which can be seen from the fact that the singular values of $\mathcal{B}_{11}$ are contained in the interval $[7.0 \times 10^{-2}, 2.3 \times 10^{1}]$.

Because of near primal degeneracy, we see from Table 2 that for `qap6`, $\kappa(\mathcal{K})$ is no longer bounded independent of μ due to the fact that the (1,1) block $\mathcal{H}$ of $\mathcal{K}$ is nearly singular. In fact, both $\kappa(\mathcal{K})$ and $\kappa(\mathcal{H})$ have order equal to the reciprocal of the machine precision ($\approx 2 \times 10^{-16}$). This example illustrates that for a nearly primal or dual degenerate problem, the matrix $\mathcal{K}$ can also be very ill-conditioned, just like the matrix M.

TABLE 2
Same as Table 1 *but for the SDP problem* `qap6`. *The approximate optimal solution has a relative duality gap of* 6.5e-9. *This SDP appears to be primal degenerate but dual nondegenerate.*

Iteration	$X \bullet Z/n$	ϕ	$\kappa(\mathcal{K})$	$\kappa(\mathcal{H})$	$\kappa(\mathcal{B}_{11}S_{11}^{-1/2})$	$\kappa(M)$
24	1.8e-06	5.2e-12	8.6e+16	7.3e+16	3.2e+02	1.7e+20
25	4.6e-07	4.8e-12	4.6e+17	9.7e+17	3.2e+02	5.0e+20
26	1.9e-07	1.1e-11	4.8e+18	2.9e+18	3.2e+02	1.2e+20
27	1.0e-07	7.5e-12	7.5e+18	1.6e+18	3.2e+02	1.3e+20
28	7.6e-08	4.3e-12	3.0e+18	1.0e+19	3.2e+02	2.0e+21
29	6.7e-08	7.3e-12	2.2e+18	1.1e+18	3.2e+02	3.3e+20

Our third example is the problem `mcp250-1` (with $m = 250$, $n = 250$) from SDPLIB. This problem has a strictly complementary approximate optimal solution, with $p = 25$ and $q = 225$. This problem is primal nondegenerate since the singular values of $[\mathcal{B}_{11}\ \mathcal{B}_{12}]$ are contained in the interval $[1.3 \times 10^{-1}, 1]$. But it is clearly dual degenerate since $\bar{p} > m$ violates the necessary condition for dual nondegeneracy in Theorem 2.3. This is also reflected in Table 3 with $\kappa(\mathcal{B}_{11}S_{11}^{-1/2})$ numerically equal to infinity.

TABLE 3
Same as Table 1 *but for the SDP problem* `mcp250-1`. *The approximate optimal solution has a relative duality gap of* 1.5e-13. *This SDP is primal nondegenerate, but it is dual degenerate.*

Iteration	$X \bullet Z/n$	ϕ	$\kappa(\mathcal{K})$	$\kappa(\mathcal{H})$	$\kappa(\mathcal{B}_{11}S_{11}^{-1/2})$	$\kappa(M)$
14	2.4e-07	1.1e-15	1.1e+11	9.8e+03	Inf	5.9e+08
15	3.8e-08	9.7e-16	2.9e+11	6.1e+03	Inf	2.3e+09
16	3.3e-09	6.4e-16	5.3e+12	7.2e+03	Inf	2.8e+10
17	1.0e-10	6.3e-16	1.9e+14	7.7e+03	Inf	9.0e+11
18	3.3e-12	6.7e-16	5.5e+15	7.3e+03	Inf	2.7e+13

A closer inspection of the problem data of `mcp250-1` reveals that it has 20 constraints that fix for a given i, $X_{ii} = 1$, and $X_{ij} = 0$ for $j \neq i$. That is, X is actually a block-diagonal matrix where one of the blocks is the 20×20 identity matrix. The presence of such a fixed block makes the problem dual degenerate. By removing the fixed block, the resulting problem has $m = 230$ and $n = 230$. The new problem becomes dual nondegenerate, and now the condition number of $\mathcal{K}$ is bounded independent of μ, as shown in Table 4. The singular values of $\mathcal{B}_{11}S_{11}^{-1/2}$ are now in the interval $[3.5 \times 10^{-2}, 2.2 \times 10^{-1}]$. This example shows that preprocessing SDP data is an important step to avoid degeneracies, and hence also potential numerical difficulties. Preprocessing to avoid degeneracies is especially important when one chooses to use an iterative solver to compute the search direction since degeneracies can seriously increase the condition number of the coefficient matrix, and hence worsen the convergence rate.

Our last example is on an SDP that is both primal and dual degenerate. This problem, `fap01`, is an SDP relaxation of a frequency assignment problem considered in [5]. This SDP has a semidefinite variable in $\mathcal{S}_+^{52}$ and a linear variable in $\mathbb{R}_+^{1160}$. The number of constraints is $m = 1378$. The approximate optimal solution is strictly

TABLE 4

Same as Table 3 *for the SDP problem* `mcp250-1` *but with fixed diagonal block removed. The approximate optimal solution has a relative duality gap of* 5.7e-14. *This SDP is primal and dual nondegenerate.*

Iteration	$X \bullet Z/n$	ϕ	$\kappa(\mathcal{K})$	$\kappa(\mathcal{H})$	$\kappa(\mathcal{B}_{11}S_{11}^{-1/2})$	$\kappa(M)$
14	2.5e-07	6.4e-16	2.7e+06	9.2e+03	6.2e+00	4.5e+08
15	3.5e-08	6.8e-16	1.2e+06	6.1e+03	6.2e+00	2.2e+09
16	2.7e-09	6.1e-16	1.7e+06	7.5e+03	6.2e+00	3.7e+10
17	9.0e-11	7.0e-16	1.8e+06	7.8e+03	6.2e+00	1.0e+12
18	3.0e-12	7.7e-16	1.7e+06	7.3e+03	6.2e+00	3.0e+13
19	2.3e-13	6.7e-16	1.7e+06	7.5e+03	6.2e+00	3.9e+14

complementary with $\min_i\{\lambda_i(X) + \lambda_i(Z)\} = 3.9 \times 10^{-4}$. We have $p = 48$, $q = 4$ for the semidefinite block, and $p = 30$, $q = 1130$ for the linear block. The matrix $[\mathcal{B}_{11}\ \mathcal{B}_{12}] \in \mathbb{R}^{m\times 1368}$ has 6 singular values that are smaller than 5×10^{-16}, with the rest contained in the interval $[8.2 \times 10^{-2}, 2]$. It is clear that the problem is primal degenerate since $[\mathcal{B}_{11}\ \mathcal{B}_{12}]$ does not have full row rank. The matrix $\mathcal{B}_{11}$ has 4 singular values that are smaller than 10^{-11}, with the rest lying in the interval $[1.7\times 10^{-2}, 2]$. Since $\mathcal{B}_{11}$ has very small singular values, the problem can be considered dual degenerate. The condition numbers of $\mathcal{K}$, $\mathcal{H}$, and $\mathcal{B}_{11}S_{11}^{-1/2}$ in Table 5 clearly reflect the fact that the problem is primal and dual degenerate.

TABLE 5

Same as Table 1 *but for the SDP problem* `fap01`. *The approximate optimal solution has a relative duality gap of* 1.6e-14. *This SDP appears to be both primal and dual degenerate.*

Iteration	$X \bullet Z/n$	ϕ	$\kappa(\mathcal{K})$	$\kappa(\mathcal{H})$	$\kappa(\mathcal{B}_{11}S_{11}^{-1/2})$	$\kappa(M)$
24	2.8e-08	8.7e-11	4.8e+06	5.7e+06	9.1e+13	5.6e+12
25	1.3e-09	8.1e-16	5.0e+11	2.2e+08	9.7e+14	1.4e+15
26	3.0e-11	1.9e-15	1.5e+13	1.0e+10	1.4e+14	1.9e+18
27	6.6e-13	2.1e-15	3.9e+15	4.5e+11	2.2e+13	4.0e+16
28	1.6e-14	2.1e-15	3.7e+16	1.4e+13	5.2e+11	4.5e+17
29	3.9e-16	1.7e-15	2.1e+18	3.5e+14	6.2e+10	1.5e+20

The reader would have noticed that for all the examples, except `qap6`, we are able to compute very accurate approximate solutions (with ϕ and relative duality gap both smaller than 2×10^{-13}). It is rather surprising that this is possible for the last example since $\mathcal{K}$ is highly ill-conditioned.

3. Solving the SCE via the conjugate residual method. The use of an iterative method to solve the SCE (5) requires less computer memory compared to using a direct method. It also has the added advantage that one can terminate the iterative solver whenever an approximate solution of (5) is deemed sufficiently accurate. This can lead to a significant saving in the CPU time required in each interior-point iteration, especially during the initial phase, where accurate computation of the search direction is not necessary. In [19], Kojima, Shida, and Shindoh (KSS) proposed inexact search directions, where (3a) and (3b) are satisfied exactly but (3c) is relaxed. If Δy is an approximate solution of (5), the KSS inexact search direction requires

the computation of the matrix $U := \mathcal{A}^T(\mathcal{A}\mathcal{A}^T)^{-1}(h - M\Delta y)$ for computing ΔX and determining whether $\|(\mathcal{E}\Delta X + \mathcal{F}\Delta Z) - R_c\|_F = \|\mathcal{E}(U)\|_F$ is sufficiently small. However, such a computation can be expensive when either $\mathcal{A}\mathcal{A}^T$ is not easily invertible or when computing $\mathcal{E}(U)$ is expensive. Due to these drawbacks, we decide to use the heuristic rule described below to compute an inexact search direction.

Suppose Δy satisfies (5) only approximately. Let $r = h - M\Delta y$. Given such an Δy, we compute ΔZ and ΔX via the following equations:

$$\Delta Z \;=\; R_d - \mathcal{A}^T\Delta y, \qquad \Delta X \;=\; \mathcal{E}^{-1}R_c - \mathcal{E}^{-1}\mathcal{F}\Delta Z. \tag{20}$$

Then $(\Delta X, \Delta y, \Delta Z)$ satisfies (3a)–(3c) approximately, where the residual vector is $[r\ ;\ 0\ ;\ 0]^T$. As our interest is to solve (3a)–(3c), it is reasonable to insist that the relative residual norm of the approximate solution $(\Delta X, \Delta y, \Delta Z)$ must be smaller than some prescribed threshold, say θ. Let

$$\|(R_p, R_d, R_c)\| \;:=\; \max(\|R_p\|,\, \|R_d\|_F,\, \|R_c\|_F). \tag{21}$$

That is, we want

$$\|r\| \;\le\; \theta\,\|(R_p, R_d, R_c)\|. \tag{22}$$

Note that for the dual variables, once dual feasibility is achieved, it is maintained because (3b) is satisfied exactly. However, for the primal variable, primal infeasibility may deteriorate since (3a) is satisfied only approximately. But we can ensure that the primal infeasibility is reduced proportionately to $\|(R_p, R_d, R_c)\|$ in each iteration. Suppose the new primal iterate is $X^+ = X + \alpha\Delta X$, where $\alpha \in (0, 1]$ is the step-length; then we have

$$\|b - AX^+\| \;\le\; (1-\alpha)\|R_p\| + \alpha\|r\| \;\le\; \Big(1 - \alpha(1-\theta)\Big)\|(R_p, R_d, R_c)\|.$$

The behavior of the preconditioned conjugate residual (PCR) method on the SCE was discussed in detail in [30]. Because the matrix M is dense, it is difficult to adapt existing preconditioning techniques that are mainly designed for sparse matrices to M, and the only obvious and easily implementable choices are diagonal preconditioners. In [30], the PCR method was applied to following preconditioned version of (5):

$$\underbrace{L^{-1}ML^{-T}}_{\widehat{M}}(L^T\Delta y) \;=\; L^{-1}h, \tag{23}$$

where $L = \mathrm{diag}(\sqrt{M_{11}}, \ldots, \sqrt{M_{mm}})$. It was observed that the PCR method on (23) is highly efficient in computing an approximate solution when the iterate (X, y, Z) is not close to optimality, i.e, when the duality gap $X \bullet Z$ is not too small. However, when the iterate is close to optimality, the PCR method becomes exceedingly slow because the matrix $\widehat{M}$ becomes very ill-conditioned (with a condition number of the order $1/\mu$), and also a more accurate solution of the system (23) is needed when the duality gap is small.

As we shall compare with the reduced augmented equation approach in section 4, the strength of solving (23) by an iterative method such as the PCR method lies in its simplicity and inexpensive matrix-vector products (where each cost about $3\rho_s n^3 + 2\rho_t mn^2$ flops; ρ_s and ρ_t are defined in section 4.1). Thus it is desirable to use the PCR method whenever its convergence rate is not too slow.

4. Computing the search direction via the reduced augmented system. Assume that Δy and ΔX are computed inexactly from Theorem 2.2 and the residual vector from (15) is denoted by

$$\begin{bmatrix} \xi \\ \eta \end{bmatrix}. \tag{24}$$

Then simple algebraic manipulations show that we have

$$-\mathcal{U}\Delta X + \mathcal{A}^T\Delta y = \mathcal{R} - Q_1\,\mathbf{smat}(S_{11}^{1/2}\eta)\,Q_1^T,$$

$$\mathcal{A}\Delta X = R_p - \xi + \mathcal{A}\,\mathbf{svec}(Q_1\mathbf{smat}(S_{11}^{-1/2}\eta)Q_1^T).$$

Now if we compute ΔZ via the equation

$$\Delta Z = R_d - \mathcal{A}^T\Delta y - Q_1\,\mathbf{smat}(S_{11}^{1/2}\eta)Q_1^T, \tag{25}$$

then we have

$$\begin{aligned}\mathcal{E}\Delta X + \mathcal{F}\Delta Z &= \mathcal{F}\left(\mathcal{U}\Delta X + R_d - \mathcal{A}^T\Delta y - Q_1\,\mathbf{smat}(S_{11}^{1/2}\eta)Q_1^T\right)\\ &= \mathcal{F}\,(R_d - \mathcal{R}) \;=\; R_c,\end{aligned}$$

where $\mathcal{R}$ is defined as in (4a). Thus, for the inexact search direction $(\Delta X, \Delta y, \Delta Z)$ computed from (15) and (25), it satisfies (3a)–(3c) approximately, and the residual vector is

$$\begin{bmatrix} \xi - \mathcal{A}\,\mathbf{svec}(Q_1\mathbf{smat}(S_{11}^{-1/2}\eta)Q_1^T) \\ Q_1\,\mathbf{smat}(S_{11}^{1/2}\eta)\,Q_1^T \\ 0 \end{bmatrix}. \tag{26}$$

Again, we want the relative residual norm of our inexact search direction $(\Delta X, \Delta y, \Delta Z)$ to be sufficiently small. That is, we want

$$\max(\|\xi - \mathcal{A}\,\mathbf{svec}(Q_1\mathbf{smat}(S_{11}^{-1/2}\eta)Q_1^T)\|,\ \|S_{11}^{1/2}\eta\|) \;\le\; \theta\,\|(R_p, R_d, R_c)\|. \tag{27}$$

Remark. Notice that we computed ΔZ as in (25) so as to satisfy the linearized complementarity equation (3c) exactly. However, if it is desirable to maintain dual feasibility, then we can compute ΔZ via $\Delta Z = R_d - \mathcal{A}^T\Delta y$ to make (3b) exact but (3c) approximately satisfied. In the latter case, if we let $V = \mathbf{smat}(S_{11}^{1/2}\eta)$, then the residual associated with (3c) is given by

$$\mathcal{F}(Q_1VQ_1^T) \;=\; [\Sigma(GQ_1)V(GQ_1)^T + (GQ_1)V(GQ_1)^T\Sigma]/2,$$

which can be computed in $2p^2n + pn^2$ flops (with symmetry taken into account) if GQ_1 is precomputed. Because of the extra cost incurred in the present case, this explains why we prefer to compute ΔZ via (25).

Observe that (3a) and (3b) are not satisfied exactly; primal and dual feasibilities are not maintained even if the iterate happens to be feasible. However, in each

iteration, the infeasibilities are reduced proportionately with $\|(R_p, R_d, R_c)\|$. From (26) and (27), the primal infeasibility for the new iterate satisfies

$$\begin{aligned}\|b - \mathcal{A}X^+\| &\le (1-\alpha)\|R_p\| + \alpha\|\xi - \mathcal{A}\,\mathbf{svec}(Q_1\mathbf{smat}(S_{11}^{-1/2}\eta)Q_1^T)\| \\ &\le \Big(1 - \alpha(1-\theta)\Big)\|(R_p, R_d, R_c)\|.\end{aligned}$$

It is easy to see that a similar inequality holds for the new dual iterate.

4.1. Preconditioned symmetric quasi-minimal residual method. Recall that the reduced augmented equation (15) is symmetric but indefinite. In this subsection, we will discuss an appropriate Krylov subspace method to solve such a linear system.

The standard Krylov subspace methods for solving a symmetric indefinite system are SYMMLQ and MINRES due to Paige and Saunders [24]. When preconditioning is used, both methods above require the preconditioner to be symmetric positive definite, and this excludes the use of indefinite preconditioners that are perhaps more appropriate since the coefficient matrix itself is indefinite. Here, we choose the preconditioned symmetric quasi-minimal residual (PSQMR) method proposed in [12] that allows the use of symmetric indefinite preconditioners. Note that if no preconditioning is used, the SQMR method and MINRES are mathematically equivalent.

Let $\mathcal{I}$ be the set of indices of nonzero elements of the matrix $\sum_{k=1}^m |A_k|$ (where $|A_k|$ is the matrix whose (i,j) element is the magnitude of the corresponding element of A_k), and

$$\rho_s = \left(\text{number of nonzero elements of the matrix } \sum_{k=1}^m |A_k|\right)/n^2,$$

$$\rho_t = (\text{total number of nonzero elements of } A_1, A_2, \ldots, A_m)/(mn^2).$$

Note that ρ_s and ρ_t are the ratios of the actual number of nonzero elements over the maximum possible number of nonzero elements.

In each PSQMR iteration, we compute the matrix-vector product $\mathcal{K}[u\,;\,v]$ for the reduced augmented system (15) via the procedure described in Table 6, where the cost is also estimated.

The cost of a matrix-vector product for the reduced augmented system is $3p^2n + 3\rho_s\, pn^2 + 7\rho_s\, n^3 + 2\rho_t\, mn^2$, as estimated in Table 6. In contrast, the corresponding cost for the SCE (5) is $3\rho_s n^3 + 2\rho_t mn^2$, as estimated in [30]. In our numerical experiments in section 5, we have found that the cost of the former range from 2 to 4 times more expensive than the latter. For the projected SCE approach proposed in [30], a matrix-vector product would cost about $6p^2n + 6\rho_s pn^2 + 4\rho_s n^3 + 6\rho_t mn^2 + 2\bar{p}^2$, and this is usually more expensive than that for the reduced augmented system.

In the current literature, most preconditioning techniques are proposed for a sparse matrix that is stored explicitly, and preconditioners such as incomplete Cholesky factors are generally quite effective for matrices that are not too ill-conditioned [28]. However, as the reader may have recalled, our matrix $\mathcal{K}$ is dense and is not formed explicitly. Thus, most of the current preconditioning techniques [28, Chapter 10] are not applicable to our linear system. The only obvious and easily implementable choices for our system are diagonal preconditioners.

In [26], some effective diagonal preconditioners were proposed for a symmetric indefinite matrix of the form $\mathcal{K}$ that arises from the finite element solution of Biot's

TABLE 6
Computational cost required in the matrix-vector product for (15).

Computing	Number of flops required
$T := \mathcal{A}^T u$	$\rho_t\, mn^2$
$\{U_{ij}^{(1)} \mid (i,j) \in \mathcal{I}\}$, where $U_1 := P_1 \circledast P_1(T)$	$3\rho_s\, n^3$
$\{U_{ij}^{(2)} \mid (i,j) \in \mathcal{I}\}$, where $U_2 := (2P_2 + P_3) \circledast P_3(T)$	$4\rho_s\, n^3$
$\{U_{ij}^{(3)} \mid (i,j) \in \mathcal{I}\}$, where $U_3 := Q_1 \circledast Q_1 \mathbf{smat}(S_{11}^{-1/2} v)$	$2p^2 n + \rho_s\, pn^2$
$\mathcal{A}(U_1 + U_2 + U_3)$	$\rho_t\, mn^2$
$S_{11}^{-1/2} \mathbf{svec}(Q_1^T \circledast Q_1^T(T)) - \Psi v$	$p^2 n + 2\rho_s\, pn^2$
$\mathcal{K}[u\,;\, v]$	$3p^2 n + 3\rho_s\, pn^2 + 7\rho_s\, n^3 + 2\rho_t\, mn^2$

soil consolidation equations. Those diagonal preconditioners were derived from some theoretical forms that are proven to have tight eigenvalue clustering properties. By adapting those preconditioners for our matrix $\mathcal{K}$, we get

$$\begin{bmatrix} \mathrm{diag}(\mathcal{H}) & 0 \\ 0 & \alpha\, \mathrm{diag}\Big(S_{11}^{-1/2} \mathcal{B}_{11}^T \mathrm{diag}(\mathcal{H})^{-1} \mathcal{B}_{11} S_{11}^{-1/2} + \Psi \Big) \end{bmatrix}, \tag{28}$$

where α is a given scalar. In our numerical experiments in section 5, we take $\alpha = -20$. Notice that the diagonal preconditioner (28) is indefinite.

5. Numerical experiments on SDPs arising from maximum clique problems. We will now present numerical experiments to show the convergence behavior of the PCR method on (23) versus the PSQMR method on (15).

All the numerical results presented in this paper are computed using MATLAB on a 700MHz HP workstation c3700 with 1G of RAM. Note that computational intensive parts such as the PCR and PSQMR methods are implemented in C but with interface to MATLAB. To give an idea of the speed of this machine, we run the MATLAB benchmark command, `bench`. Compared to a 300MHz SGI R1200 IRIX 64 machine, our machine is about 2 times faster on LU factorization and has about the same speed on sparse matrix operations.

The interior-point method we used is the primal–dual path-following method (without corrector) described in [32], except that the direct solver used to solve (5) is replaced by an iterative solver. The following starting iterates (slightly modified from the default in [32]) are used throughout:

$$y^0 \,=\, 0, \quad X^0 \,=\, \xi\, I, \quad Z^0 \,=\, \eta\, I,$$

where

$$\xi = n \max\left(\sqrt{n},\ n\ \max_k \left\{ \frac{1 + |b_k|}{1 + \|A_k\|_F} \right\} \right),$$

$$\eta = \sqrt{n}\ \max\left(n,\ \|C\|_F,\ \max_k \{\|A_k\|_F\} \right).$$

For easy reference, we will call the interior-point method in [32] that uses the PCR method to solve the preconditioned SCE (23) Algorithm PFsch ("PF" for "path-following"). The parameter θ in (22) is set to 0.01. In view of the efficiency of the

PCR method in computing an inexact search direction via (23) when the duality gap $X \bullet Z$ is not too small, for the experiments we use a hybrid method that combines the advantage of applying the PCR method to (23) and the PSQMR method to (15) for computing the search direction in each interior-point iteration. The details of the hybrid method are given in Algorithm PFaug in Table 7. The parameter θ in (27) is set to $\theta = 0.05$.

Our test problems consist of the following two collections of SDPs:

1. the first consists of SDPs arising from maximum clique problems on randomly generated graphs;
2. the second consists of SDPs associated with maximum clique problems for graphs from the Second DIMACS Implementation Challenge [9].

We choose these SDP collections because they are likely to be primal and dual nondegenerate. These are problems with m large and n moderate. Thus they are also well suited for solution via a primal–dual interior point method with an iterative solver. There are two commonly used equivalent SDP relaxations [27, equations (2.6) and (2.9)] for the maximum clique problems. The relaxation we used for a given simple undirected graph (G, V) follows equation (2.6) in [27]. That is,

$$\min\{-(ee^T) \bullet X \,:\, \mathrm{Trace}(X) = 1,\ X_{ij} = 0\ \forall\, (i,j) \in E,\ X \succeq 0\},$$

where e is the vector of all ones. We also tested on the second formulation given in [27, equation (2.9)] but found that the SDPs are typically either primal or dual degenerate.

Let

$$N_k \ = \ \begin{cases} \text{the number of PCR/PSQMR steps required at the } k\text{th interior-} \\ \text{point iteration to solve (23)/(15) so that the admissible condition} \\ \text{(22)/(27) is satisfied.} \end{cases}$$

The maximum numbers of PCR and PSQMR steps allowed in each interior-point iteration are set to $5m$ and $3m$, respectively.

Table 8 shows the primal and dual objective values obtained by Algorithm PFaug. Table 9–12 compare the cumulative CPU time taken by Algorithms PFsch and PFaug at various interior-point iterations so as to achieve the following accuracy:

$$\max(\texttt{relgap}, \phi) \ \le \ 10^{-4},\ 10^{-5},\ 10^{-6}.$$

Here `relgap` is the relative duality gap defined by

$$\texttt{relgap} \ = \ \frac{X \bullet Z}{1 + (|C \bullet X| + |b^T y|)/2}, \tag{29}$$

and ϕ is the infeasibility measure defined in (19). For each problem, three rows of data are reported, and they correspond to the CPU time needed to solve the problem to an accuracy of 10^{-4}, 10^{-5}, and 10^{-6}, respectively.

Tables 9–11 show that Algorithm PFaug is much faster than Algorithm PFsch on the majority of the problems tested. For example, consider the problem `theta82` with $m = 23872$; Algorithm PFaug is about 7 times faster than Algorithm PFsch to achieve an accuracy of 10^{-6} in $\max(\texttt{relgap}, \phi)$. On the set of maximum clique problems on randomly generated graphs considered in Table 9, Algorithm PFaug is 3–14 times faster than Algorithm PFsch. For those SDPs arising from [9] in Table 11, Algorithm PFaug is 2–9 times faster than Algorithm PFsch. The reader may have

observed that the speedup in these problems is mainly gained on the last few interior-point iterations. Comparing Tables 9 and 11, we see that the number of PSQMR steps needed to solve (15) is far less than that required by (23) when the iterates are close to optimality. This also confirms the usefulness of Theorem 2.4. But because computing a matrix-vector product for (15) is more expensive, the saving in CPU time is not as impressive as the reduction in the number iterative steps.

It is worth noting that in Table 9, we are able to solve an SDP (`theta162`) with 127600 constraints in 6.5 hours to an accuracy of 10^{-6}. If the required accuracy is 10^{-4}, then only 3.5 hours is needed.

Observe that in Table 10 the reduced augmented system (15) in Algorithm PFaug is never invoked, indicating that the condition number of the NT scaling matrix W never exceeds 5×10^3 in Algorithm PFaug described in Table 7. It is surprising that the collection of `hamming` problems can be solved so efficiently via the SCE alone. For example, the problem `hamming-9-5-6` is solved to an accuracy of 10^{-6} in 3 minutes, whereas the CPU time reported in [5] by using the first-order nonlinear programming method in [4] (we will called it the BMZ method for convenience) is more than 10 hours. Although the comparison here between Algorithm PFaug and the BMZ method is not entirely fair because the latter solves a different, though equivalent, SDP relaxation [27, equation (2.9)] of the maximum clique problem, the fact that such a disparity is possible indicates that one should not totally abandon second-order methods in favor of first-order methods when solving large scale SDPs.

Despite the success of Algorithm PFaug reported in Tables 9–11, for the problems in Table 12 the performance of Algorithm PFaug is worse than Algorithm PFsch. For example, it performs badly compared to Algorithm PFsch on the problem `p-hat300-1`. To understand why the reduced augmented equation approach does not perform well, we need to know whether the problem is degenerate. This can be done by estimating the condition number of $\mathcal{H}$ and $\mathcal{B}_{11}S_{11}^{-1/2}$. Since the matrices are large, we used the Lanczos method to estimate the largest and smallest eigenvalues of the matrices $\mathcal{H}$ and $S_{11}^{-1/2}\mathcal{B}_{11}^T\mathcal{B}_{11}S_{11}^{-1/2}$. The ratio between these eigenvalues would then give a lower bound on $\kappa(\mathcal{H})$ and $\kappa(\mathcal{B}_{11}S_{11}^{-1/2})^2$. A lower bound we get for $\kappa(\mathcal{H})$ is 1.5×10^8. As for $\kappa(\mathcal{B}_{11}S_{11}^{-1/2})$, we are able to get an accurate estimate of 9.0×10^1. From these numbers, we may conclude that the problem is dual nondegenerate, but it is possibly primal degenerate due to the large condition number estimate we have for $\mathcal{H}$. The poorer performance of Algorithm PFaug on `G51--G54` compared to Algorithm PFsch is also due to degeneracies.

Methods for solving large SDPs are still in the infancy state. Currently, the most successful methods are the spectral bundle (SB) method in [16], the BMZ method in [4], and the BMPR method in [3]. Detailed comparison between the SB (version 1.1.1) and BMZ methods are given in [5], where the BMZ method appeared to perform generically better than the SB method on the tested set of SDPs arising from maximum clique problems from the Second DIMACS Implementation Challenge. Comparison between the BMPR and SB (version 1.1.1) methods on the same set of SDPs are reported in [3]. Based on the results reported in [3] and [5], it is known that the BMPR is superior to the BMZ method on this set of SDPs. The latter is in turn superior to the SB method. Thus in this section we shall compare our reduced augmented equation approach mainly with the BMPR method.

For the set of SDPs listed in Tables 10 to 12, except `G43--G47`, `G51--G54`, Algorithm PFaug is comparable to the BMPR method (Tables 4 and 7 in [3]) in terms of computational time (although we must take into account that different machines are

used, and our machine is 1–2 times faster based on MATLAB's `bench` command). For example, the problems `brock400-1` and `c-fat-200-1` are solved by Algorithm PFaug in 1016 and 67 seconds, respectively. The corresponding numbers for the BMPR method reported in [3] are 2028 and 742 seconds.

For `G43--G47`, our method is able to solve them in about 1.5 hours to an accuracy of 10^{-6}. The BMPR method, however, is much faster, taking an average of 15 minutes to solve these problems (see Table 7 in [3]), though less accurately than ours. The fact that the latter is far superior to Algorithm PFaug on these problems can be explained. First, because the matrix variable has a relatively large dimension of $n = 1000$, computing the NT scaling matrix W and eigenvalue decomposition of W^{-1} in Algorithm PFaug takes more than 50% of the total computation time. Second, the rank of the optimal primal variable (about 60) is small compared to n; the BMPR method can fully exploit such an advantage, whereas Algorithm PFaug is not designed to do the same. The reasons above apply also to the problems `G51--G54`. For the problems `G52` and `G53`, our interior-point based algorithms perform much worse than the BMPR method. For example, Algorithm PFsch takes 6.5 hours to achieve an accuracy of 10^{-4}, whereas the BMPR method takes only 2 hours to achieve a comparable accuracy. If the required accuracy is 10^{-6}, then Algorithm PFsch would take about 33.5 hours to solve the problem. Comparing the results in Table 11 and 12, obviously `G52` and `G53` are much harder to solve compared to `G43--47`. We suspect that this is because the former are highly degenerate problems. For example, for `G52` we have $\bar{p} = 48828 \gg m = 5917$, which violates the necessary condition for dual nondegeneracy in Theorem 2.3.

We note that the objective values reported in Table 8 are generally better than those reported in [3]. For example, the primal objective value we obtained for `brock400-1` is -39.7018863, with a primal infeasibility of 5.1×10^{-10}, whereas the corresponding number obtained by the BMPR method is -39.652, with a primal infeasibility of 1.4×10^{-4}.

Other than computational time, we should mention a comparison criterion between interior-point methods (such as Algorithm PFaug) and first-order methods (such as the BMPR method) that is perhaps underappreciated. An advantage of the former is that it can produce a duality gap that measures how close the approximate optimal solution is to optimality. The latter, however, can only obtain either an approximate primal or dual optimal solution, and there is no optimality guarantee on the approximate solution delivered.

6. Numerical experiments on other SDPs. In this section, we further investigate the performance of Algorithms PFaug and PFsch but on some SDPs that are not necessarily well suited for the reduced augmented equation or the primal–dual interior-point framework. The problems we considered are as follows.

`mcp`: this collection consists of the preprocessed version of the SDPLIB problems `mcp500-1`–`mcp500-4`. These are SDPs arising from relaxation of maximum cut problems. The original SDPs are dual degenerate, but a simple preprocessing step to remove fixed diagonal blocks render them dual nondegenerate. For these problems, $m \approx 500$ and $n = 500$.

`arch`: this consists of the SDPLIB problems `arch0, arch2, arch4`, and `arch8`. Each of these problems has a semidefinite variable of dimension 161 and a linear variable of dimension 174, and $m = 174$.

`fap`: these are SDPs arising from semidefinite relaxation of frequency assignment problems [11]. The explicit form of the primal SDP is given in [5, equation

TABLE 7
Algorithm PFaug.

Algorithm PFaug. Suppose we are given an initial iterate (X^0, y^0, Z^0) with X^0, Z^0 positive definite. Set $\gamma^0 = 0.9$ and $\sigma^0 = 0.5$.

For $k = 0, 1, \ldots$

Let the current and the next iterate be (X, y, Z) and (X^+, y^+, Z^+) respectively. Also, let the current and the next step-length (centering) parameter be denoted by γ and γ^+ (σ and σ^+), respectively.

1. Set $\mu = X \bullet Z/n$. Stop the iteration if the infeasibility measure ϕ defined in (19) and `relgap` defined in (29) are sufficiently small.
2. Compute the NT scaling matrix W and the eigenvalue decomposition $W^{-1} = QDQ^T$. Let $d = \mathrm{diag}(D)$, where d is sorted in ascending order.

   If $\max(d)/\min(d) > 5 \times 10^3$
   choose p to be the integer such that d_{p+1}/d_p is the maximum,
   else
   set $p = 0$,
   end
3. (a) If $p = 0$;
   Compute an inexact direction $(\Delta X, \Delta y, \Delta Z)$ via the PCR method on (23) with diagonal preconditioner $\mathrm{diag}(M)$.

   (b) If $p > 0$;
   Compute an inexact search direction $(\Delta X, \Delta y, \Delta Z)$ via the PSQMR method on (15) with diagonal preconditioner described in (28).
4. Update (X, y, Z) to (X^+, y^+, Z^+) by

$$X^+ = X + \alpha\,\Delta X, \quad y^+ = y + \beta\,\Delta y, \quad Z^+ = Z + \beta\,\Delta Z,$$

   where $\alpha = \min\big(1, -\gamma/\lambda_{\min}(X^{-1}\Delta X)\big)$, $\beta = \min\big(1, -\gamma/\lambda_{\min}(Z^{-1}\Delta Z)\big)$. (Here $\lambda_{\min}(U)$ denotes the minimum eigenvalue of U; if the minimum eigenvalue in either expression is positive, we ignore the corresponding term.)
5. Update the step-length and centering parameters by

$$\gamma^+ = 0.9 + 0.08\min(\alpha, \beta), \quad \sigma^+ = 1 - 0.9\min(\alpha, \beta).$$

(5)] (note the difference between maximizing and minimizing the objective function in [5, equation (5)] and (1)). Note that this collection of SDPs are likely to be both primal and dual degenerate (evident from Table 5 for `fap01`). Each of these problems has a semidefinite variable with moderate dimension n and a linear variable with dimension slightly less than m, where m is the number of constraints and $m \gg n$.

Before we discuss the numerical results for the above SDPs, we would like to mention that many of the SDPs in SDPLIB [6] appear to be either primal or dual degenerate or ill-posed in the sense that the primal and dual problems are not both strictly feasible. The `mcp` problems are dual degenerate if fixed diagonal blocks are not removed. The `qap` problems are nearly primal degenerate; the `control` problems either do not appear to have strictly complementary approximate optimal solutions or they are primal degenerate. The `gpp` problems do not have strictly primal feasible points.

TABLE 8
Primal and dual objective values obtained by Algorithm PFaug.

Problem	n	m	Primal obj	Dual obj
theta6	300	4375	-63.4770649	-63.4770915
theta62	300	13390	-29.6412339	-29.6412589
theta8	400	7905	-73.9535154	-73.9535717
theta82	400	23872	-34.3668848	-34.3668981
theta83	400	39862	-20.3018839	-20.3018980
theta10	500	12470	-83.8059524	-83.8059706
theta102	500	37467	-38.3905171	-38.3905620
theta103	500	62516	-22.5285606	-22.5285800
theta104	500	87245	-13.3361385	-13.3361438
theta12	600	17979	-92.8016040	-92.8016958
theta123	600	90020	-24.6686484	-24.6686554
theta162	800	127600	-37.0097262	-37.0097436
MANN-a27	378	703	-132.7628635	-132.7628930
johnson8-4-4	70	561	-13.9999840	-14.0000044
johnson16-2-4	120	1681	-7.9999998	-8.0000017
san200-0.7-1	200	5971	-29.9999629	-30.0000002
c-fat200-1	200	18367	-11.9999970	-12.0000002
hamming-6-4	64	1313	-5.3333301	-5.3333351
hamming-8-4	256	11777	-15.9999977	-16.0000010
hamming-9-8	512	2305	-223.9996367	-224.0000138
hamming-10-2	1025	23040	-102.3999498	-102.4000165
hamming-7-5-6	128	1793	-42.6666515	-42.6666678
hamming-8-3-4	256	16129	-25.5999744	-25.6000043
hamming-9-5-6	512	53761	-85.3331694	-85.3333369
brock200-1	200	5067	-27.4566346	-27.4566445
brock200-4	200	6812	-21.2934670	-21.2934817
brock400-1	400	20078	-39.7018863	-39.7019055
keller4	171	5101	-14.0122384	-14.0122440
sanr200-0.7	200	6033	-23.8361531	-23.8361601
G43	1000	9991	-280.6245145	-280.6245830
G44	1000	9991	-280.5831314	-280.5832102
G45	1000	9991	-280.1848899	-280.1851375
G46	1000	9991	-279.8365727	-279.8369756
G47	1000	9991	-281.8938988	-281.8939612
p-hat300-1	300	33918	-10.0679626	-10.0679686
G51	1000	5910	-348.9996545	-349.0001073
G52	1000	5917	-348.3860739	-348.3864065
G53	1000	5915	-348.3469748	-348.3473550
G54	1000	5917	-340.9998601	-341.0000203

The `mcp` and `arch` problems are problems with $m \approx n$, and they are not large scale. Thus they are not ideal examples to evaluate the viability of using iterative methods to solve large SDPs in a primal–dual interior-point method. However, they are included here to evaluate the merit of the reduced augmented equation (15) over the SCE (5) when solved via an iterative method. The CPU time given in Table 14 for these problems is not indicative of the time spent in solving these linear systems because a substantial part is spent on computing W and its eigenvalue decomposition.

TABLE 9

Comparison of Algorithms PFsch and PFaug on a number of SDP problems arising from maximum clique problems on randomly generated graphs.

		Algorithm PFsch				Algorithm PFaug				
n m	Iter. no.	relgap	ϕ	Cum. time	N_k	relgap	ϕ	Cum. time	N_k	p
theta6	22 (22)	7.8 -5	2.2 -5	1:11	715	9.3 -5	4.1 -6	1:00	170	47
300	24 (24)	4.3 -6	7.9 -7	6:12	6782	1.8 -6	1.6 -7	1:49	365	47
4375	25 (25)	6.2 -7	1.6 -7	12:11	11641	5.0 -7	3.1 -8	2:28	540	48
theta62	21 (21)	7.9 -5	1.2 -5	3:35	1327	7.9 -5	1.2 -5	3:17	1327	0
300	23 (23)	3.2 -6	4.9 -7	26:31	15796	3.2 -6	9.2 -8	7:43	1015	106
13390	24 (24)	8.7 -7	1.0 -7	1:03:13	33831	8.4 -7	6.9 -9	10:38	1105	106
theta8	23 (23)	2.4 -5	5.0 -6	7:57	3247	2.3 -5	2.8 -8	4:00	400	66
400	24 (24)	4.4 -6	1.1 -6	18:21	7896	3.0 -6	9.5 -9	5:18	415	66
7905	25 (25)	7.2 -7	1.8 -7	43:24	18436	7.6 -7	3.9 -8	6:37	430	66
theta82	23 (23)	2.2 -5	3.6 -6	15:27	3036	2.2 -5	2.3 -7	10:35	510	138
400	24 (24)	2.7 -6	5.1 -7	45:44	11751	2.7 -6	2.4 -9	14:37	620	138
23872	25 (25)	3.9 -7	7.6 -8	2:19:49	36504	3.9 -7	6.6-10	18:16	555	138
theta83	22 (22)	2.8 -5	2.9 -6	16:04	2959	2.8 -5	2.9 -6	15:49	2959	0
400	23 (23)	4.0 -6	4.5 -7	48:23	12031	4.0 -6	1.4 -7	29:18	1785	204
39862	24 (24)	6.9 -7	8.0 -8	2:31:30	38507	6.9 -7	4.4 -9	38:21	1160	201
theta10	22 (22)	9.0 -5	2.2 -5	10:01	1327	9.7 -5	9.1 -6	7:13	265	81
500	24 (24)	1.9 -6	6.5 -7	55:50	9120	1.8 -6	9.5 -8	13:36	460	81
12470	25 (25)	2.3 -7	9.6 -8	2:32:49	28412	2.2 -7	2.6 -8	17:52	560	82
theta102	23 (23)	6.3 -5	7.9 -6	22:25	1800	6.3 -5	7.9 -6	22:04	1800	0
500	24 (24)	8.8 -6	1.3 -6	54:53	6437	8.8 -6	2.6 -7	32:33	820	170
37467	26 (26)	1.7 -7	3.3 -8	7:51:53	62416	1.7 -7	8.9-10	1:02:15	1120	174
theta103	22 (22)	3.3 -5	2.9 -6	35:16	3719	3.3 -5	2.9 -6	34:50	3719	0
500	23 (23)	5.3 -6	4.8 -7	1:38:14	12045	5.3 -6	3.6 -8	54:21	1190	252
62516	24 (24)	8.6 -7	9.0 -8	4:26:17	32131	8.6 -7	1.8 -9	1:11:38	1005	252
theta104	22 (22)	7.3 -5	4.0 -6	24:28	2070	7.3 -5	4.0 -6	24:12	2070	0
500	24 (24)	2.2 -6	1.0 -7	3:31:02	26022	2.2 -6	1.0 -8	2:25:53	4285	332
87245	25 (25)	4.0 -7	2.1 -8	9:24:10	65693	4.0 -7	3.4-10	3:08:09	2160	328
theta12	24 (24)	5.3 -5	1.5 -5	25:21	1430	5.8 -5	3.3 -6	18:39	255	98
600	25 (25)	9.2 -6	2.3 -6	1:07:37	5626	9.2 -6	3.7 -8	25:13	395	98
17979	27 (26)	1.4 -7	5.9 -8	8:03:22	40323	9.8 -7	3.4 -8	33:13	485	98
theta123	23 (23)	5.2 -5	4.7 -6	47:35	2707	5.2 -5	4.7 -6	47:25	2707	0
600	24 (24)	7.9 -6	7.4 -7	2:10:12	9587	7.9 -6	1.2 -7	1:20:28	1135	301
90020	26 (26)	2.8 -7	2.7 -8	16:29:46	72630	2.8 -7	5.0-10	2:20:16	885	301
theta162	25 (25)	1.4 -5	1.8 -6	3:20:59	6126	1.4 -5	1.8 -6	3:21:34	6126	0
800	26 (26)	2.3 -6	3.0 -7	10:00:53	20400	2.3 -6	5.9 -8	4:58:07	1670	335
127600	27 (27)	4.7 -7	6.0 -8	27:47:52	54234	4.7 -7	4.2 -9	6:34:37	1650	340

For the mcp problems, the iterative solvers use only less than 30% of the total CPU time. Thus the number of iterative steps used to solve the linear systems would be a better indicator of the relative merit between (5) and (15). From Table 14, we observe that the PSQMR method on (15) takes significantly fewer steps to converge than the PCR method on (5) when μ is small. This confirms again the merit of the reduced augmented equation over the SCE when μ is small.

The fap problems are SDPs that are both primal and dual degenerate. Because these problems are expected to be hard to solve via an interior-point method

TABLE 10
Comparison of Algorithms PFsch and PFaug on SDPs from the Second DIMACS Challenge on Maximum Clique Problems.

		Algorithm PFsch				Algorithm PFaug				
n m	Iter. no.	`relgap`	ϕ	Cum. time	N_k	`relgap`	ϕ	Cum. time	N_k	p
MANN-a27	39 (39)	7.5 -5	1.8 -5	1:48	28	7.5 -5	1.8 -5	1:43	28	0
378	41 (41)	2.0 -6	9.5 -7	1:56	50	2.0 -6	9.5 -7	1:51	50	0
703	42 (42)	2.0 -7	9.8 -8	2:00	61	2.0 -7	9.8 -8	1:55	61	0
johnson8-4-4	16 (16)	1.5 -5	1.1-10	0:18	2	1.5 -5	1.1-10	0:16	2	0
70	17 (17)	1.5 -6	6.0-10	0:19	2	1.5 -6	6.0-10	0:17	2	0
561	18 (18)	1.5 -7	3.6 -9	0:20	2	1.5 -7	3.6 -9	0:18	2	0
johnson16-2-4	17 (17)	2.4 -5	7.8-13	0:20	2	2.4 -5	7.8-13	0:18	2	0
120	18 (18)	2.4 -6	6.4-12	0:22	2	2.4 -6	6.4-12	0:19	2	0
1681	19 (19)	2.4 -7	5.3-11	0:23	2	2.4 -7	5.3-11	0:20	2	0
san200-0.7-1	23 (23)	1.2 -5	8.4 -7	0:35	22	1.2 -5	8.4 -7	0:32	22	0
200	24 (24)	1.2 -6	9.6 -8	0:37	30	1.2 -6	9.6 -8	0:34	30	0
5971	25 (25)	1.2 -7	1.0 -8	0:39	49	1.2 -7	1.0 -8	0:36	49	0
c-fat200-1	21 (21)	2.8 -5	1.6 -6	0:54	175	2.8 -5	1.6 -6	0:51	175	0
200	22 (22)	2.8 -6	2.1 -7	1:01	255	2.8 -6	2.1 -7	0:58	255	0
18367	23 (23)	2.8 -7	1.9 -8	1:10	313	2.8 -7	1.9 -8	1:07	313	0
hamming-6-4	15 (15)	9.4 -5	5.0 -7	0:17	3	9.4 -5	5.0 -7	0:15	3	0
64	16 (16)	9.4 -6	4.9 -8	0:18	3	9.4 -6	4.9 -8	0:16	3	0
1313	17 (17)	9.4 -7	4.9 -9	0:19	3	9.4 -7	4.9 -9	0:17	3	0
hamming-8-4	20 (20)	2.1 -5	1.9 -7	0:34	4	2.1 -5	1.9 -7	0:31	4	0
256	21 (21)	2.1 -6	1.3 -8	0:36	5	2.1 -6	1.3 -8	0:33	5	0
11777	22 (22)	2.1 -7	1.2 -8	0:38	4	2.1 -7	1.2 -8	0:35	4	0
hamming-9-8	19 (19)	1.7 -5	1.3 -6	2:10	4	1.7 -5	1.3 -6	2:08	4	0
512	20 (20)	1.7 -6	7.6 -7	2:17	4	1.7 -6	7.6 -7	2:15	4	0
2305	21 (21)	1.7 -7	2.3 -9	2:25	5	1.7 -7	2.3 -9	2:23	5	0
hamming-10-2	21 (21)	6.5 -5	3.4-10	17:31	3	6.5 -5	3.4-10	17:36	3	0
1025	22 (22)	6.5 -6	1.5 -9	18:26	3	6.5 -6	1.5 -9	18:31	3	0
23040	23 (23)	6.5 -7	5.1 -8	19:19	2	6.5 -7	5.1 -8	19:24	2	0
hamming-7-5-6	17 (17)	3.8 -5	4.0 -6	0:21	2	3.8 -5	4.0 -6	0:18	2	0
128	18 (18)	3.8 -6	5.2 -9	0:22	4	3.8 -6	5.2 -9	0:19	4	0
1793	19 (19)	3.8 -7	1.9 -8	0:23	4	3.8 -7	1.9 -8	0:21	4	0
hamming-8-3-4	19 (19)	1.2 -5	2.6 -8	0:32	3	1.2 -5	2.6 -8	0:30	3	0
256	20 (20)	1.2 -6	5.5 -8	0:34	4	1.2 -6	5.5 -8	0:31	4	0
16129	21 (21)	1.2 -7	1.5 -9	0:36	6	1.2 -7	1.5 -9	0:33	6	0
hamming-9-5-6	20 (20)	2.0 -5	2.7 -6	2:47	6	2.0 -5	2.7 -6	2:46	6	0
512	21 (21)	2.0 -6	8.7 -8	2:56	5	2.0 -6	8.7 -8	2:56	5	0
53761	22 (22)	2.0 -7	6.4 -9	3:06	5	2.0 -7	6.4 -9	3:06	5	0

using an iterative solver, now the accuracy tolerance is set to $\max(\texttt{relgap}, \phi) \leq 10^{-2}, 10^{-3}, 10^{-4}$ in Table 15. Also, because these problems have convergence difficulty in a purely primal–dual path-following method, we use a primal–dual path-following method with Mehrotra's predictor-corrector. Note that in each iteration of the predictor-corrector method, two linear systems with the same coefficient matrix have to be solved. But having the same coefficient matrix offers no savings in computation time for an iterative solver, unlike the case of a direct solver where the same factorization can be used for both linear systems. Thus, unless necessary, predictor-

TABLE 11
Comparison of Algorithms PFsch and PFaug on SDPs from the Second DIMACS Challenge on Maximum Clique Problems.

		Algorithm PFsch				Algorithm PFaug				
n m	Iter. no.	relgap	ϕ	Cum. time	N_k	relgap	ϕ	Cum. time	N_k	p
brock200-1	21 (21)	3.8 -5	6.6 -6	1:33	1646	3.6 -5	1.1 -7	1:14	335	63
200	22 (22)	6.2 -6	1.1 -6	3:35	6498	5.9 -6	8.0 -9	1:30	330	63
5067	24 (24)	3.2 -7	1.0 -7	16:50	25334	3.6 -7	3.3-10	1:57	240	63
brock200-4	21 (21)	1.8 -5	2.5 -6	2:15	3185	1.8 -5	8.7 -9	1:29	640	79
200	22 (22)	3.5 -6	4.8 -7	5:52	11188	3.5 -6	1.6 -9	2:14	900	79
6812	23 (23)	7.2 -7	8.2 -8	14:17	26054	6.9 -7	1.1 -9	2:41	545	77
brock400-1	23 (23)	3.0 -5	5.8 -6	11:55	2077	3.0 -5	6.2 -8	9:33	515	123
400	24 (24)	3.3 -6	7.4 -7	34:22	8883	3.3 -6	2.9 -9	13:04	560	123
20078	25 (25)	4.8 -7	1.1 -7	1:53:34	31285	4.8 -7	5.1-10	16:56	620	123
keller4	21 (21)	4.2 -5	4.2 -6	0:51	1008	4.2 -5	2.9 -7	0:25	110	67
171	22 (22)	4.2 -6	4.2 -7	1:33	3180	4.2 -6	1.3 -7	0:34	235	67
5101	23 (23)	4.2 -7	4.2 -8	2:30	5063	4.2 -7	3.4 -8	0:58	765	67
sanr200-0.7	21 (21)	2.6 -5	4.4 -6	1:45	2149	2.6 -5	7.0 -8	1:07	410	71
200	22 (22)	4.6 -6	7.9 -7	4:33	8718	4.6 -6	1.1 -8	1:32	510	71
6033	23 (24)	9.3 -7	1.4 -7	11:39	22494	3.0 -7	2.7-10	2:15	375	71
G43	27 (27)	4.0 -5	3.1 -5	48:39	754	5.4 -5	3.6 -7	43:18	245	56
1000	29 (28)	4.7 -6	8.3 -7	3:37:58	11230	9.7 -6	2.2 -7	51:22	300	58
9991	30 (32)	4.7 -7	3.8 -7	5:16:54	8943	2.6 -7	3.8 -8	1:25:23	265	58
G44	28 (28)	2.0 -5	1.2 -5	1:09:00	1497	1.4 -5	1.2 -7	54:55	220	60
1000	29 (29)	7.4 -6	3.7 -6	2:38:09	8070	2.8 -6	2.7 -9	1:04:46	380	60
9991	31 (30)	2.5 -7	2.0 -7	9:44:55	14144	2.8 -7	2.3-10	1:11:52	255	60
G45	28 (29)	1.8 -5	1.4 -5	1:48:32	854	8.9 -5	5.9 -7	1:09:36	310	57
1000	29 (30)	3.6 -6	2.4 -6	3:09:17	7315	8.9 -6	3.7 -7	1:15:55	220	58
9991	30 (31)	7.7 -7	2.7 -7	7:38:25	24263	8.9 -7	4.0 -8	1:24:22	320	58
G46	27 (27)	3.9 -5	3.1 -5	48:01	652	5.5 -5	1.6 -6	44:23	235	56
1000	29 (28)	7.6 -6	2.9 -6	3:35:40	10852	8.9 -6	1.0 -7	53:19	340	60
9991	31 (30)	2.6 -7	2.1 -7	7:20:06	10383	1.8 -7	9.8-11	1:09:28	335	60
G47	27 (27)	5.3 -5	4.0 -5	44:42	608	6.3 -5	6.7 -7	44:50	190	58
1000	30 (28)	2.0 -6	1.4 -6	2:36:55	2108	9.7 -6	1.4 -7	51:11	225	58
9991	31 (30)	4.1 -7	1.5 -7	4:54:55	12555	2.2 -7	8.3 -9	1:07:37	375	58

corrector approach is not the preferred option when an iterative solver is used.

Because of degeneracies, the reduced augmented equation would offer no advantage over the SCE for the fap problems. Since each matrix-vector product in (15) is more expensive, it is logical only to expect that Algorithm PFsch would be more efficient than Algorithm PFaug. This expectation is confirmed by the numerical results presented in Table 15. It is evident that Algorithm PFaug consistently takes a longer time than Algorithm PFsch to solve the problems. Furthermore, Algorithm PFaug fails to solve eight of the problems (entries with boldface fonts) to the required accuracy of 10^{-4}, whereas Algorithm PFsch successfully solved all. This set of SDPs illustrates that for problems that are degenerate, it is not advisable to use an iterative method to solve the reduced augmented equation (15). Unless modifications on (15) are done to handle the ill-conditioning of $\mathcal{K}$, it appears that the simplest approach of using the PCR method on (5) should be used.

Table 13 shows the primal and dual objective values for the mcp, arch, and fap

TABLE 12
Comparison of Algorithms PFsch and PFaug on SDPs from the Second DIMACS Challenge on Maximum Clique Problems.

		Algorithm PFsch				Algorithm PFaug				
n m	Iter. no.	relgap	ϕ	Cum. time	N_k	relgap	ϕ	Cum. time	N_k	p
p-hat300-1	24 (24)	9.3 -5	1.3 -6	13:19	4357	9.3 -5	1.1 -9	1:42:24	26790	209
300	26 (26)	6.8 -6	1.8 -7	42:52	17512	6.9 -6	3.8 -8	9:49:14	101750	200
33918	28 (28)	7.1 -7	1.2 -8	2:12:47	43234	5.9 -7	1.7-10	20:47:54	100775	200
G51	44 (43)	4.2 -5	1.5 -5	1:19:28	305	8.7 -5	5.8 -6	2:56:49	1505	5
1000	45 (45)	4.3 -6	3.5 -6	1:23:51	482	1.4 -6	1.4 -6	3:06:26	414	0
5910	46 (46)	4.8 -7	4.2 -7	1:28:16	490	1.4 -7	1.4 -7	3:11:11	514	0
G52	60 (58)	9.3 -5	5.2 -6	4:28:59	3291	8.7 -5	1.2 -6	13:02:23	5395	6
1000	68 (65)	6.8 -6	6.7 -7	9:04:03	4212	7.9 -6	2.7 -7	25:58:35	9985	8
5917	71 (69)	6.0 -7	2.2 -7	11:31:50	5907	9.5 -7	5.9 -8	33:19:59	14257	0
G53	56 (58)	8.4 -5	3.0 -6	6:25:55	6874	7.7 -5	6.4 -6	7:25:49	3905	6
1000	62 (63)	8.8 -6	1.2 -6	14:06:25	26162	8.7 -6	4.6 -7	17:32:15	12955	6
5915	68 (68)	5.2 -7	3.2 -7	33:24:45	29574	8.6 -7	4.9 -7	37:15:59	17740	6
G54	45 (45)	6.6 -5	3.9 -5	1:17:04	288	3.7 -5	6.3 -6	3:35:28	2290	7
1000	46 (46)	6.6 -6	5.8 -6	1:21:25	477	3.7 -6	3.5 -6	3:40:09	494	0
5917	48 (47)	3.4 -7	1.1 -7	1:45:27	2325	3.7 -7	3.3 -7	3:46:16	699	0

TABLE 13
Primal and dual objective values obtained by Algorithm PFsch.

Problem	n	m	Primal obj	Dual obj
mcp500-1	451	451	-598.1479228	-598.1485310
mcp500-2	493	493	-1070.0563326	-1070.0567704
mcp500-3	500	500	-1847.9696030	-1847.9700289
mcp500-4	500	500	-3566.7373504	-3566.7380666
arch0	161	174	-0.5665156	-0.5665177
arch2	161	174	-0.6715133	-0.6715158
arch4	161	174	-0.9726271	-0.9726275
arch8	161	174	-7.0569738	-7.0569811
fap01	52	1378	0.0329454	0.0328773
fap02	61	1866	0.0007310	0.0006973
fap03	65	2145	0.0493711	0.0493676
fap04	81	3321	0.1749789	0.1748222
fap05	84	3570	0.3083974	0.3082823
fap06	93	4371	0.4595326	0.4593247
fap07	98	4851	2.1180259	2.1176137
fap08	120	7260	2.4363666	2.4362657
fap09	174	15225	10.7982702	10.7976727
fap10	183	14479	0.0096992	0.0096708
fap11	252	24292	0.0298764	0.0297662
fap12	369	26462	0.2734163	0.2732371

TABLE 14
Comparison of Algorithms PFsch and PFaug on `mcp` *and* `arch` *problems from SDPLIB.*

		Algorithm PFsch				Algorithm PFaug				
n m	Iter. no.	`relgap`	ϕ	Cum. time	N_k	`relgap`	ϕ	Cum. time	N_k	p
mcp500-1	25 (25)	1.0 -5	1.1 -6	1:24	225	1.1 -5	3.1 -6	1:49	95	11
451	26 (26)	1.0 -6	1.1 -7	1:38	411	3.3 -6	2.9 -7	1:58	95	11
451	27 (27)	1.0 -7	9.9 -9	1:55	536	5.1 -7	5.4 -8	2:10	140	11
mcp500-2	22 (22)	8.2 -5	6.5 -6	1:43	336	8.2 -5	6.5 -6	1:47	336	0
493	24 (24)	2.5 -6	3.3 -7	2:21	550	3.1 -6	4.9 -7	2:19	170	8
493	25 (25)	4.1 -7	4.6 -8	3:04	1161	9.2 -7	1.8 -7	2:44	270	8
mcp500-3	20 (20)	7.4 -5	1.3 -5	1:40	195	7.5 -5	2.8 -5	1:48	85	8
500	21 (22)	9.9 -6	2.6 -6	1:55	322	6.2 -6	1.6 -6	2:17	110	9
500	23 (24)	2.3 -7	5.2 -8	2:55	848	6.0 -7	5.0 -8	2:47	115	9
mcp500-4	19 (19)	5.4 -5	2.5 -5	1:45	183	5.4 -5	2.5 -5	1:54	183	0
500	21 (21)	7.7 -6	1.7 -6	2:43	885	7.7 -6	1.8 -6	2:31	70	11
500	23 (24)	2.0 -7	1.3 -7	3:35	681	7.2 -7	9.6 -8	3:12	60	11
arch0	52 (52)	9.1 -5	8.3 -7	0:57	869	9.4 -5	5.8 -9	0:45	100	8
161	57 (57)	6.0 -6	4.1 -6	1:17	869	6.3 -6	2.4-11	0:54	110	8
174	58 (62)	**2.8 -6**	**3.7 -6**	1:21	869	5.8 -7	7.5-11	1:02	105	8
arch2	45 (45)	5.9 -5	1.0 -6	0:49	869	5.9 -5	3.0 -9	0:30	80	8
161	48 (48)	6.0 -6	1.0 -6	1:03	869	6.0 -6	4.7-10	0:35	90	8
174	50 (52)	**2.2 -6**	**7.2 -6**	1:11	869	7.6 -7	4.1-12	0:41	100	8
arch4	50 (51)	9.9 -5	3.3 -6	0:56	869	5.9 -5	8.9-10	0:36	75	5
161	52 (53)	4.4 -6	3.7 -7	1:05	869	2.8 -6	1.6-11	0:39	80	5
174	53 (54)	7.2 -7	**2.6 -6**	1:09	869	2.9 -7	2.1-11	0:40	75	5
arch8	52 (52)	2.1 -5	1.4 -8	0:50	587	3.5 -5	6.8 -9	0:33	35	7
161	53 (53)	7.0 -6	7.2 -7	0:54	869	9.7 -6	1.2 -9	0:34	55	7
174	58 (57)	**1.0 -6**	2.5 -7	1:16	869	1.1 -7	6.9-11	0:39	50	8

problems obtained by Algorithm PFsch.

It has been reported in [5] that the BMZ method is highly successful in solving the `fap` problems compared to the SB method. (The BMPR method is not tested on the `fap` problems in [3].) By comparing the performance of Algorithm PFsch in Table 15 with the results report in [5, Table 6], we observe that our interior-point method fared reasonably well compared to the first-order BMZ method. The CPU time taken to solve all the problems, except `fap12`, are comparable for both methods (again, we must take into account that different machines are used, and our machine is 1–2 times faster). For example, the problem `fap11` is solved in 9 hours by Algorithm PFsch, and the CPU time reported in [5] is 10.8 hours.

The objective values we obtained in Table 13 for the `fap` problems are better than those obtained in [5]. Take `fap11`, for example; the dual objective value we obtained is 0.0297662, with a dual infeasibility of 1.1×10^{-16}. This value is better (the larger the absolute value the better) than the absolute value of 0.0296136 reported in [5]. For `fap12`, the BMZ method is superior to Algorithm PFsch, where the former is about 3 times faster if an accuracy requirement of 10^{-4} is set for the latter. But if the accuracy requirement is set to 10^{-3}, then Algorithm PFsch can solve `fap12` in about 19 hours compared to 12.5 hours in the BMZ method.

Our comparison here between Algorithm PFsch and the BMZ method indicates that interior-point methods are not totally uncompetitive compared to first-order methods.

TABLE 15
Comparison of Algorithms PFsch and PFaug on `fap` *problems.*

		Algorithm PFsch				Algorithm PFaug				
n m	Iter. no.	`relgap`	ϕ	Cum. time	N_k	`relgap`	ϕ	Cum. time	N_k	p
fap01	26 (26)	2.5 -3	4.3 -7	0:10	4474	2.5 -3	4.3 -7	0:10	4474	0
52	27 (27)	3.7 -4	9.8 -8	0:15	10071	3.7 -4	9.8 -8	0:15	10071	0
1378	28 (28)	6.7 -5	3.4 -7	0:21	11022	**3.8 -4**	4.6 -5	0:25	6885	46
fap02	23 (23)	5.9 -3	4.1 -7	0:05	1116	5.9 -3	4.1 -7	0:05	1116	0
61	25 (25)	7.6 -4	4.5 -8	0:09	2660	7.6 -4	4.5 -8	0:09	2660	0
1866	27 (27)	3.4 -5	4.0 -9	0:15	6096	2.9 -5	**1.4 -4**	0:31	9325	56
fap03	30 (30)	3.3 -3	3.3 -7	0:17	4431	3.3 -3	3.3 -7	0:17	4431	0
65	31 (31)	8.2 -4	8.6 -8	0:25	9612	3.2 -3	1.5 -4	0:47	10715	60
2145	33 (31)	2.8 -5	1.1 -6	0:55	17158	**3.2 -3**	**1.5 -4**	0:47	10715	60
fap04	37 (37)	4.1 -3	3.3 -7	1:47	23146	5.8 -3	9.6 -7	2:28	15250	76
81	39 (39)	4.2 -4	8.3 -7	3:12	26566	2.5 -3	1.2 -5	5:25	16600	76
3321	40 (39)	8.7 -5	4.7 -6	3:55	26566	**2.5 -3**	1.2 -5	5:25	16600	76
fap05	41 (41)	6.4 -3	2.6 -7	2:26	28558	7.5 -3	8.6 -7	4:12	14620	79
84	43 (45)	8.4 -4	1.4 -5	4:04	28558	2.7 -3	2.6 -6	11:38	17840	79
3570	45 (45)	3.6 -5	2.9 -5	5:40	28558	**2.7 -3**	2.6 -6	11:38	17840	79
fap06	43 (43)	4.2 -3	1.7 -7	3:36	31778	5.7 -3	2.2 -6	10:16	21850	83
93	45 (45)	4.6 -4	6.9 -6	6:30	34966	9.8 -4	5.8 -7	16:18	20005	86
4371	47 (49)	3.4 -5	1.3 -5	9:16	34966	**3.9 -4**	2.7 -6	29:09	21850	86
fap07	43 (44)	7.7 -3	6.1 -7	4:49	34270	7.6 -3	9.4 -7	8:13	20145	93
98	46 (47)	4.0 -4	4.1 -6	10:15	38806	8.1 -4	1.8 -6	20:26	24250	90
4851	48 (50)	4.3 -5	1.1 -5	13:49	38806	**2.5 -4**	8.7 -7	33:14	24250	91
fap08	45 (45)	5.0 -3	3.0 -7	8:29	30394	6.2 -3	1.1 -6	15:51	11560	110
120	47 (48)	5.8 -4	2.1 -7	18:04	58078	5.6 -4	1.2 -7	37:13	26745	110
7260	49 (51)	4.3 -5	3.8 -6	27:59	58078	5.2 -5	2.5 -8	1:04:43	27890	110
fap09	70 (72)	5.0 -3	8.0 -7	28:53	23255	8.2 -3	2.2 -6	1:10:58	10885	157
174	73 (76)	4.3 -4	7.4 -8	1:25:14	92031	7.6 -4	2.6 -7	2:35:05	28440	156
15225	75 (77)	5.3 -5	2.2 -6	2:38:43	121798	**3.7 -4**	1.6 -7	3:14:07	42065	156
fap10	65 (65)	7.9 -3	1.1 -7	25:27	25835	7.9 -3	1.1 -7	25:36	25835	0
183	67 (67)	5.5 -4	1.4 -8	1:12:04	99556	5.5 -4	7.9 -6	1:47:55	72390	144
14479	69 (70)	1.3 -5	2.2 -6	2:28:29	115830	7.1 -5	5.7 -5	5:11:44	72390	140
fap11	71 (71)	9.3 -3	1.2 -7	1:22:14	30988	9.3 -3	1.2 -7	1:22:58	30988	0
252	73 (73)	7.9 -4	1.1 -8	3:54:12	130230	7.9 -4	1.2 -6	10:07:30	121455	175
24292	75 (75)	1.5 -5	9.2 -7	9:01:15	194334	4.5 -5	2.9 -5	18:49:02	121455	171
fap12	68	9.3 -3	1.3 -7	5:21:43	64851					
369	70	9.5 -4	4.0 -8	18:43:24	211694	excluded since it will take too long to run				
26462	72	4.0 -5	8.2 -7	33:34:41	211694					

7. Conclusion and future research. We introduced the reduced augmented equation for computing the search directions in primal–dual interior-point methods. For SDPs that are primal and dual nondegenerate and have strictly complementary optimal solutions, the coefficient matrices of the reduced augmented equations have condition numbers that are bounded independent of the barrier parameter μ, even when μ approaches 0.

We proposed Algorithm PFaug, which is based on a hybrid between the PCR method applied to the SCE and the PSQMR method applied to the reduced augmented equation. Numerical experiments on SDPs arising from maximum clique

problems show that Algorithm PFaug performs much better than Algorithm PFsch, which is based solely on applying the PCR method to the SCE.

Our interior-point based methods, Algorithms PFaug and PFsch, are competitive (timewise) compared to the first-order BMPR method on the majority of the maximum clique problems considered in [5]. Our interior-point based method, Algorithm PFsch, is also competitive compared to the first-order BMZ method on the `fap` problems. The numerical results presented in this paper indicate that interior-point methods like Algorithms PFaug and PFsch are not totally uncompetitive compared to first-order methods such as the SB, BMZ, and BMPR methods. On many of the problems tested in this paper, we are able to obtain objective values that are better or comparable to those obtained by the first-order methods.

Algorithm PFaug is well suited for primal and dual nondegenerate problems with optimal solutions that are strictly complementary. It appears that significant modifications to the reduced augmented equation are needed to effectively solve problems that are degenerate. Besides this important issue, there are a number of other issues that we hope to address in the future.

(a) We would like to investigate the performance of the reduced augmented equation approach in a dual scaling interior-point framework for solving SDPs with n large, especially large SDPs arising from maximum cut and graph partitioning problems.
(b) The construction of more sophisticated preconditioners for the reduced augmented matrix.
(c) The use of a direct method to solve the reduced augmented equation so as to generate accurate approximate optimal solutions. Our numerical results in section 2.1 indicate that the outcome would be promising.

Acknowledgments. Part of this paper was written while the author was visiting the Tokyo Institute of Technology under the Hitachi Fellowship. He thanks Professor Masakazu Kojima for hosting his visit and for many stimulating discussions.

The author thanks Sam Burer for providing him the data to generate the frequency assignment problems. Finally, he thanks the referees for reading the paper very carefully and for their helpful comments.

REFERENCES

[1] F. Alizadeh, J.-P. A. Haeberly, and M. L. Overton, *Primal-dual interior-point methods for semidefinite programming: Convergence results, stability and numerical results*, SIAM J. Optim., 8 (1998), pp. 746–768.
[2] F. Alizadeh, J. A. Haeberly, and M. Overton, *Complementarity and nondegeneracy in semidefinite programming*, Math. Programming, 77 (1997), pp. 111–128.
[3] S. Burer and R. Monterio, *A nonlinear programming algorithm for solving semidefinite programs via low-rank factorization*, Math. Program., 95 (2003), pp. 329–357.
[4] S. Burer, R. Monterio, and Y. Zhang, *Solving a class of semidefinite programs via nonlinear programming*, Math. Program., 93 (2002), pp. 97–122.
[5] S. Burer, R. Monterio, and Y. Zhang, *A computational study of a gradient-based log-barrier algorithm for a class of large scale SDPs*, Math. Program., 95 (2003), pp. 359–379.
[6] B. Borchers, *SDPLIB* 1.2, *a library of semidefinite programming test problems*, Optim. Methods Softw., 11/12 (1999), pp. 683–690.
[7] S. J. Benson, Y. Ye, and X. Zhang, *Solving large-scale sparse semidefinite programs for combinatorial optimization*, SIAM J. Optim., 10 (2000), pp. 443-461.
[8] C. Choi and Y. Ye, *Solving Sparse Semidefinite Programs Using the Dual Scaling Algorithm with an Iterative Solver*, working paper, Computational Optimization Laboratory, University of Iowa, Iowa City, IA, 2000.

[9] M. Trick, V. Chvatal, W. Cook, D. Johnson, C. McGeoch, and R. Tarjan, *The Second DIMACS Implementation Challenge: NP Hard Problems: Maximum Clique, Graph Coloring, and Satisfiability*, Rutgers University, http://dimacs.rutgers.edu/Challenges/ (1992).

[10] T. A. Driscoll, K.-C. Toh, and L. N. Trefethen, *From potential theory to matrix iterations in six steps*, SIAM Rev., 40 (1998), pp. 547–578.

[11] A. Eisenblätter, M. Grötschel, and A. M. C. A. Koster, *Frequency planning and ramification of coloring*, Discuss. Math. Graph Theory, 22 (2002), pp. 51–88.

[12] R. W. Freund and N. M. Nachtigal, *A new Krylov-subspace method for symmetric indefinite linear systems*, in Proceedings of the 14th IMACS World Congress on Computational and Applied Mathematics, W. F. Ames, ed., Atlanta, GA, 1994, pp. 1253–1256.

[13] K. Fujisawa, M. Kojima, and K. Nakata, *Exploiting sparsity in primal-dual interior-point methods for semidefinite programming*, Math. Programming, 79 (1997), pp. 235–253.

[14] M. Fukuda, M. Kojima, and M. Shida, *Lagrangian dual interior-point methods for semidefinite programs*, SIAM J. Optim., 12 (2002), pp. 1007–1031.

[15] M. Fukuda, M. Kojima, K. Murota, and K. Nakata, *Exploiting sparsity in semidefinite programming via matrix completion* I: *General framework*, SIAM J. Optim., 11 (2000), pp. 647–674.

[16] C. Helmberg and K. C. Kiwiel, *A spectral bundle method with bounds*, Math. Program., 93 (2002), pp. 173–194.

[17] C. Helmberg, F. Rendl, R. J. Vanderbei, and H. Wolkowicz, *An interior-point method for semidefinite programming*, SIAM J. Optim., 6 (1996), pp. 342–361.

[18] M. Kojima, S. Shindoh, and S. Hara, *Interior-point methods for the monotone semidefinite linear complementarity problem in symmetric matrices*, SIAM J. Optim., 7 (1997), pp. 86–125.

[19] M. Kojima, M. Shida, and S. Shindoh, *Search directions in the SDP and the monotone SDLCP: Generalization and inexact computation*, Math. Program., 85 (1999), pp. 51–80.

[20] C.-J. Lin and R. Saigal, *An incomplete Cholesky factorization for dense symmetric positive definite matrices*, BIT, 40 (2000), pp. 536–558.

[21] R. D. C. Monteiro, *Primal–dual path-following algorithms for semidefinite programming*, SIAM J. Optim., 7 (1997), pp. 663–678.

[22] K. Nakata, K. Fujisawa, and M. Kojima, *Using the conjugate gradient method in interior-points methods for semidefinite programs* (in Japanese), Proc. Inst. Statist. Math., 46 (1998), pp. 297–316 (in Japanese).

[23] K. Nakata, S.-L. Zhang, and M. Kojima, *Preconditioned Conjugate Gradient Methods for Large Scale and Dense Linear Systems in Semidefinite Programming*, abstract based on talks delivered at INFORMS Meeting, Philadelphia, 1999.

[24] C. C. Paige and M. A. Saunders, *Solution of sparse indefinite systems of linear equations*, SIAM J. Numer. Anal., 12 (1975), pp. 617–629.

[25] G. Pataki, *On the rank of extreme matrices in semidefinite programs and the multiplicity of optimal eigenvalues*, Math. Oper. Res., 23 (1998), pp. 339–358.

[26] K. K. Phoon, K. C. Toh, S. H. Chan, and F. H. Lee, *An efficient diagonal preconditioner for finite element solution of Biot's consolidation equations*, Internat. J. Numer. Methods Engrg., 55 (2002), pp. 377–400.

[27] G. Gruber and F. Rendl, *Computational experience with stable set relaxations*, SIAM J. Optim., 13 (2003), pp. 1014–1028.

[28] Y. Saad, *Iterative Methods for Sparse Linear Systems*, PWS, Boston, 1996.

[29] K. C. Toh, *An Analysis of Ill-Conditioned Equilibrium Systems*, Technical report, Department of Mathematics, National University of Singapore, Singapore, 2002.

[30] K.-C. Toh and M. Kojima, *Solving some large scale semidefinite programs via the conjugate residual method*, SIAM J. Optim., 12 (2002), pp. 669–691.

[31] M. J. Todd, K. C. Toh, and R. H. Tütüncü, *On the Nesterov–Todd direction in semidefinite programming*, SIAM J. Optim., 8 (1998), pp. 769–796.

[32] K. C. Toh, M. J. Todd, and R. H. Tütüncü, *SDPT*3 — *a* Matlab *software package for semidefinite programming, version* 1.3, Optim. Methods Softw., 11 (1999), pp. 545–581.

[33] V. V. Kovacevic-Vujcic and M. D. Asic, *Stabilization of interior-point methods for linear programming*, Comput. Optim. Appl., 14 (1999), pp. 331–346.

[34] Y. Zhang, *On extending some primal-dual interior-point algorithms from linear programming to semidefinite programming*, SIAM J. Optim., 8 (1998), pp. 365–386.

[35] Q. Zhao, S. E. Karisch, F. Rendl, and H. Wolkowicz, *Semidefinite programming relaxations for the quadratic assignment problem*, J. Comb. Optim., 2 (1998), pp. 71–109.

SIAM J. OPTIM.
Vol. 14, No. 3, pp. 699–707

ON STEEPEST DESCENT ALGORITHMS FOR DISCRETE CONVEX FUNCTIONS*

KAZUO MUROTA†

Abstract. This paper investigates the complexity of steepest descent algorithms for two classes of discrete convex functions: M-convex functions and L-convex functions. Simple tie-breaking rules yield complexity bounds that are polynomials in the dimension of the variables and the size of the effective domain. Combining the present results with a standard scaling approach leads to an efficient algorithm for L-convex function minimization.

Key words. discrete optimization, discrete convex function, steepest descent algorithm, M-convex function, L-convex function

AMS subject classifications. 90C10, 90C25, 90C35, 90C27

DOI. 10.1137/S1052623402419005

1. Introduction.

Discrete convex functions have long been attracting research interest in the area of discrete optimization. Miller [15] was a forerunner in the early 1970s. The relationship between submodularity and convexity was discussed in Edmonds [3], and deeper understanding of this relationship was gained in the 1980s by Frank [5], Fujishige [6], and Lovász [13] (see also [7]). Favati and Tardella [4] introduced integrally convex functions to show a local characterization for global minimality, and Dress and Wenzel [2] considered valuated matroids in terms of a greedy algorithm. Recently, Murota [17, 18, 20, 21] advocated "discrete convex analysis," where M-convex and L-convex functions play central roles. $M^\natural$-convex and $L^\natural$-convex functions,[1] introduced, respectively, by Murota and Shioura [22] and Fujishige and Murota [8], are variants of M-convex and L-convex functions. It was shown in [8] that $L^\natural$-convex functions are the same as the submodular integrally convex functions considered in [4].

Minimization of discrete convex functions is most fundamental in discrete optimization. In fact, we have recently witnessed dramatic progress of algorithms for submodular set-function minimization; see, e.g., Iwata [10], Iwata, Fleischer, and Fujishige [11], Schrijver [23], and a survey by McCormick [14].

M-convex function minimization contains the minimum-weight matroid-base problem (see, e.g., [1]) as a very special case. Minimization of an M-convex function on $\{0,1\}$-vectors is equivalent to maximization of a matroid valuation, for which the greedy algorithm of Dress and Wenzel [2] works. The first polynomial time algorithm for general M-convex functions was given by Shioura [24], and scaling algorithms were considered by Moriguchi, Murota, and Shioura [16], Tamura [26], and Shioura [25].

For L-convex function minimization the algorithm of Favati and Tardella [4], originally meant for submodular integrally convex functions, works with slight modi-

*Received by the editors November 30, 2002; accepted for publication (in revised form) August 6, 2003; published electronically December 19, 2003. This work was supported by the Kayamori Foundation of Informational Science Advancement, a Grant-in-Aid of the Ministry of Education, Culture, Sports, Science and Technology of Japan, and the 21st Century COE Program on Information Science and Technology Strategic Core.

http://www.siam.org/journals/siopt/14-3/41900.html

†Graduate School of Information Science and Technology, University of Tokyo, and PRESTO, JST, Tokyo 113-8656, Japan (murota@mist.i.u-tokyo.ac.jp).

[1] "$M^\natural$-convex" should be read "M-natural-convex," and similarly for "$L^\natural$-convex."

fications. It is the first polynomial time algorithm for L-convex function minimization, but it is not practical, being based on the ellipsoid method. A steepest descent algorithm was proposed by Murota [19], with a subsequent improvement by Iwata [9] using a scaling technique. The steepest descent algorithm heavily depends on algorithms for submodular set-function minimization.

In this paper we investigate the complexity of steepest descent algorithms for M-convex functions and L-convex functions. With certain simple tie-breaking rules we can obtain complexity bounds that are polynomials in the dimension n of the variables and the size K of the effective domain. Combining the present complexity bound with a standard scaling approach results in an efficient algorithm for L-convex function minimization of complexity bounded by polynomials in n and $\log K$. This is faster than any other known algorithms for L-convex function minimization.

Some conventions are introduced. We consider functions defined on integer lattice points that may possibly take $+\infty$, i.e., $f : \mathbf{Z}^V \to \mathbf{R} \cup \{+\infty\}$ with a finite set V of cardinality n. The effective domain of f is denoted by

$$\operatorname{dom} f = \{x \in \mathbf{Z}^V \mid f(x) < +\infty\}, \tag{1.1}$$

and the ℓ_1-size of $\operatorname{dom} f$ by

$$K_f = \max\{||x - y||_1 \mid x, y \in \operatorname{dom} f\}, \tag{1.2}$$

where the ℓ_1-norm of a vector $x = (x(v) \mid v \in V)$ with components indexed by V is designated by

$$||x||_1 = \sum_{v \in V} |x(v)|.$$

For a subset X of V we denote by χ_X the characteristic vector of X; $\chi_X(v)$ equals one or zero according to whether v belongs to X or not. For $u \in V$ we denote $\chi_{\{u\}}$ by χ_u.

2. M-convex function minimization. M-convex functions are defined in terms of a generalization of the exchange axiom for matroids. We say that a function $f : \mathbf{Z}^V \to \mathbf{R} \cup \{+\infty\}$ with $\operatorname{dom} f \neq \emptyset$ is M-convex if it satisfies the exchange axiom

(M-EXC) For $x, y \in \operatorname{dom} f$ and $u \in \operatorname{supp}^+(x - y)$, there exists $v \in \operatorname{supp}^-(x - y)$ such that

$$f(x) + f(y) \geq f(x - \chi_u + \chi_v) + f(y + \chi_u - \chi_v). \tag{2.1}$$

The inequality (2.1) implicitly imposes the condition that $x - \chi_u + \chi_v \in \operatorname{dom} f$ and $y + \chi_u - \chi_v \in \operatorname{dom} f$ for the finiteness of the right-hand side. It follows from (M-EXC) that the effective domain of an M-convex function lies on a hyperplane $\{x \in \mathbf{R}^V \mid \sum_{v \in V} x(v) = r\}$ for some integer r.

Global optimality for an M-convex function is characterized by local optimality.

LEMMA 2.1 (see [17, 20, 21]). *For an M-convex function f and $x \in \operatorname{dom} f$, we have*

$$f(x) \leq f(y) \ \ (\forall\, y \in \mathbf{Z}^V) \iff f(x) \leq f(x - \chi_u + \chi_v) \ \ (\forall\, u, v \in V).$$

This local characterization of global minimality naturally suggests the following algorithm of steepest descent type [16, 19, 24].

STEEPEST DESCENT ALGORITHM FOR AN M-CONVEX FUNCTION f.
S0: Find a vector $x \in \operatorname{dom} f$.
S1: Find $u, v \in V$ $(u \neq v)$ that minimize $f(x - \chi_u + \chi_v)$.
S2: If $f(x) \leq f(x - \chi_u + \chi_v)$, then stop ($x$ is a minimizer of f).
S3: Set $x := x - \chi_u + \chi_v$ and go to S1.

Step S1 can be done with n^2 evaluations of function f. At the termination of the algorithm in step S2, x is a global optimum by Lemma 2.1. The function value f decreases monotonically with iterations. This property alone does not ensure finite termination in general, although it does if f is integer-valued and bounded from below.

The following is a key property of the steepest descent algorithm for M-convex functions, showing an upper bound on the number of iterations in terms of the distance to the optimal solution rather than in terms of the function value. We denote by x° the initial vector found in step S0.

LEMMA 2.2. *If f has a unique minimizer, say x^*, the number of iterations is bounded by $||x^\circ - x^*||_1/2$.*

Proof. Put $x' = x - \chi_u + \chi_v$ in step S2. By Lemma 2.3 below we have $x^*(u) \leq x(u)-1 = x'(u)$ and $x^*(v) \geq x(v)+1 = x'(v)$, which implies $||x'-x^*||_1 = ||x-x^*||_1-2$. Note that $||x^\circ - x^*||_1$ is an even integer. □

LEMMA 2.3 (see [24]; see also [21]). *Let $f : \mathbf{Z}^V \to \mathbf{R} \cup \{+\infty\}$ be an M-convex function with $\arg\min f \neq \emptyset$. For $x \in \operatorname{dom} f \setminus \arg\min f$, let $u, v \in V$ be such that*

$$f(x - \chi_u + \chi_v) = \min_{s,t \in V} f(x - \chi_s + \chi_t).$$

Then $u \neq v$ and there exists $x^* \in \arg\min f$ with

$$x^*(u) \leq x(u) - 1, \qquad x^*(v) \geq x(v) + 1.$$

When given an M-convex function f, which may have multiple minimizers, we consider a perturbation of the function so that we can use Lemma 2.2. Assume now that f has a bounded effective domain of ℓ_1-size K_f in (1.2). We arbitrarily fix a bijection $\varphi : V \to \{1, 2, \ldots, n\}$ to represent an ordering of the elements of V, put $v_i = \varphi^{-1}(i)$ for $i = 1, \ldots, n$, and define a function f_ε by

$$f_\varepsilon(x) = f(x) + \sum_{i=1}^{n} \varepsilon^i \, x(v_i),$$

where $\varepsilon > 0$. This function is M-convex, and, for a sufficiently small ε, it has a unique minimizer that is also a minimizer of f. Suppose that the steepest descent algorithm is applied to the perturbed function f_ε. Since

$$f_\varepsilon(x - \chi_u + \chi_v) = f(x - \chi_u + \chi_v) + \sum_{i=1}^{n} \varepsilon^i x(v_i) - \varepsilon^{\varphi(u)} + \varepsilon^{\varphi(v)}$$

this amounts to employing a tie-breaking rule:

(2.2) Take (u, v) that lexicographically minimizes $\Phi(u, v)$,

where

$$\Phi(u, v) = \begin{cases} (-1, \varphi(u), -\varphi(v)) & \text{if } \varphi(u) < \varphi(v), \\ (+1, -\varphi(v), \varphi(u)) & \text{if } \varphi(u) > \varphi(v), \end{cases}$$

in case of multiple candidates in step S1 of the steepest descent algorithm applied to f. Combining this observation with Lemma 2.1 yields the following complexity bound, where F_f denotes an upper bound on the time to evaluate f.

THEOREM 2.4. *For an M-convex function f with finite K_f, the number of iterations in the steepest descent algorithm with tie-breaking rule* (2.2) *is bounded by $K_f/2$. Hence, if a vector in* $\operatorname{dom} f$ *is given, the algorithm finds a minimizer of f in* $O(F_f \cdot n^2 K_f)$ *time.*

Although a number of algorithms of smaller theoretical complexity are already known for M-convex function minimization [24, 25, 26], the present analysis is intended to reveal the most fundamental fact about M-convex function minimization. The tie-breaking rule (2.2) as well as the steepest descent algorithm can be adapted to $M^\natural$-convex function minimization.

3. L-convex function minimization. L-convex functions are defined in terms of submodularity on integer lattice points. For integer vectors $p, q \in \mathbf{Z}^V$ we denote by $p \vee q$ and $p \wedge q$ the vectors of componentwise maximum and minimum of p and q, i.e.,

$$(p \vee q)(v) = \max(p(v), q(v)), \quad (p \wedge q)(v) = \min(p(v), q(v)) \qquad (v \in V).$$

We say that a function $g : \mathbf{Z}^V \to \mathbf{R} \cup \{+\infty\}$ with $\operatorname{dom} g \neq \emptyset$ is L-convex if it satisfies

$$\text{(SBF)} \qquad g(p) + g(q) \geq g(p \vee q) + g(p \wedge q) \qquad (\forall p, q \in \mathbf{Z}^V),$$

$$\text{(TRF)} \qquad \exists r \in \mathbf{R} \text{ such that } g(p + \mathbf{1}) = g(p) + r \qquad (\forall p \in \mathbf{Z}^V),$$

where $\mathbf{1} = (1, 1, \ldots, 1) \in \mathbf{Z}^V$. In this paper we assume $r = 0$, since otherwise g is not bounded from below and does not have a minimum.

Global optimality for an L-convex function is characterized by local optimality.

LEMMA 3.1 (see [18, 20, 21]). *For an L-convex function g with $r = 0$ in* (TRF) *and $p \in \operatorname{dom} g$, we have*

$$g(p) \leq g(q) \quad (\forall\, q \in \mathbf{Z}^V) \iff g(p) \leq g(p + \chi_X) \quad (\forall\, X \subseteq V).$$

This local characterization of global minimality naturally suggests the following algorithm of steepest descent type [19]. Recall our assumption $r = 0$ in (TRF).

STEEPEST DESCENT ALGORITHM FOR AN L-CONVEX FUNCTION g.

S0: Find a vector $p \in \operatorname{dom} g$.
S1: Find $X \subseteq V$ that minimizes $g(p + \chi_X)$.
S2: If $g(p) \leq g(p + \chi_X)$, then stop (p is a minimizer of g).
S3: Set $p := p + \chi_X$ and go to S1.

Step S1 amounts to minimizing a set-function

$$\rho_p(X) = g(p + \chi_X) - g(p)$$

over all subsets X of V. As a consequence of (SBF) this function is submodular, i.e.,

$$\rho_p(X) + \rho_p(Y) \geq \rho_p(X \cup Y) + \rho_p(X \cap Y) \qquad (\forall X, Y \subseteq V),$$

and can be minimized in strongly polynomial time (see, e.g., [10, 11, 14, 23]). At the termination of the algorithm in step S2, p is a global optimum by Lemma 3.1. The function value g decreases monotonically with iterations. This property alone does

not ensure finite termination in general, although it does if g is integer-valued and bounded from below.

We can guarantee an upper bound on the number of iterations by introducing a tie-breaking rule in step S1:

(3.1) Take the (unique) minimal minimizer X of ρ_p.

Let p° be the initial vector found in step S0. If g has a minimizer at all, it has, by (TRF), a minimizer p^* satisfying $p^\circ \le p^*$. Let p^* denote the smallest of such minimizers, which exists since $p^* \wedge q^* \in \arg\min g$ for $p^*, q^* \in \arg\min g$.

LEMMA 3.2. *In step* S1, $p \le p^*$ *implies* $p + \chi_X \le p^*$. *Hence the number of iterations is bounded by* $||p^\circ - p^*||_1$.

Proof. Put $Y = \{v \in V \mid p(v) = p^*(v)\}$ and $p' = p + \chi_X$. By submodularity we have

$$g(p^*) + g(p') \ge g(p^* \vee p') + g(p^* \wedge p'),$$

whereas $g(p^*) \le g(p^* \vee p')$ since p^* is a minimizer of g. Hence $g(p') \ge g(p^* \wedge p')$. Here we have $p' = p + \chi_X$ and $p^* \wedge p' = p + \chi_{X \setminus Y}$, whereas X is the minimal minimizer by the tie-breaking rule (3.1). This means that $X \setminus Y = X$, i.e., $X \cap Y = \emptyset$. Therefore, $p' = p + \chi_X \le p^*$. □

It is easy to find the minimal minimizer of ρ_p using the existing algorithms for submodular set-function minimization. For example, with Schrijver's algorithm [23] we can find the minimal minimizer with $O(n^8)$ function evaluations and $O(n^9)$ arithmetic operations. Assuming that the minimal minimizer of a submodular set-function can be computed with $O(\sigma(n))$ function evaluations and $O(\tau(n))$ arithmetic operations, and denoting by F_g an upper bound on the time to evaluate g, we can perform step S1 in $O(\sigma(n)F_g + \tau(n))$ time, where $(\sigma(n), \tau(n)) = (n^8, n^9)$ is a valid choice. We measure the size of the effective domain of g by

$$\hat{K}_g = \max\{||p - q||_1 \mid p, q \in \operatorname{dom} g,\ p(v) = q(v) \text{ for some } v \in V\}, \tag{3.2}$$

where it is noted that $\operatorname{dom} g$ itself is unbounded by (TRF).

THEOREM 3.3. *For an L-convex function g with finite $\hat{K}_g$, the number of iterations in the steepest descent algorithm with tie-breaking rule* (3.1) *is bounded by $\hat{K}_g$. Hence, if a vector in* $\operatorname{dom} g$ *is given, the algorithm finds a minimizer of g in* $O((\sigma(n)F_g + \tau(n))\hat{K}_g)$ *time.*

Proof. We have $||p^\circ - p^*||_1 \le \hat{K}_g$ since $p^\circ(v) = p^*(v)$ for some $v \in V$. Then the claim follows from Lemma 3.2. □

A function $g : \mathbf{Z}^V \to \mathbf{R} \cup \{+\infty\}$ is called $L^\natural$-convex if the function

$$\tilde{g}(p_0, p) = g(p - p_0\mathbf{1}) \qquad (p_0 \in \mathbf{Z}, p \in \mathbf{Z}^V) \tag{3.3}$$

is an L-convex function in $n + 1$ variables. Whereas $L^\natural$-convex functions are conceptually equivalent to L-convex functions by the relation (3.3), the class of $L^\natural$-convex functions in n variables is strictly larger than that of L-convex functions in n variables. The steepest descent algorithm for L-convex functions can be adapted to $L^\natural$-convex function minimization.

STEEPEST DESCENT ALGORITHM FOR AN $L^\natural$-CONVEX FUNCTION g.

S0: Find a vector $p \in \operatorname{dom} g$.
S1: Find $\varepsilon \in \{1, -1\}$ and $X \subseteq V$ that minimize $g(p + \varepsilon\chi_X)$.
S2: If $g(p) \le g(p + \varepsilon\chi_X)$, then stop ($p$ is a minimizer of g).
S3: Set $p := p + \varepsilon\chi_X$ and go to S1.

Step S1 amounts to minimizing a pair of submodular set functions

$$\rho_p^+(X) = g(p + \chi_X) - g(p), \qquad \rho_p^-(X) = g(p - \chi_X) - g(p).$$

Let X^+ be the minimal minimizer of ρ_p^+, and let X^- be the maximal minimizer of ρ_p^-. The tie-breaking rule for step S1 reads

$$(\varepsilon, X) = \begin{cases} (1, X^+) & \text{if } \min \rho_p^+ \le \min \rho_p^-, \\ (-1, X^-) & \text{if } \min \rho_p^+ > \min \rho_p^-. \end{cases} \tag{3.4}$$

This is a translation of the tie-breaking rule (3.1) for $\tilde{g}$ in (3.3) through the correspondence

g		$\tilde{g}$
$p \to p + \chi_X$	$\Longleftrightarrow$	$\tilde{p} \to \tilde{p} + (0, \chi_X)$
$p \to p - \chi_X$	$\Longleftrightarrow$	$\tilde{p} \to \tilde{p} + (1, \chi_{V \setminus X})$

,

where $\tilde{p} = (0, p) \in \mathbf{Z}^{1+n}$. Since $(1, \chi_{V \setminus X^-})$ cannot be minimal in the presence of $(0, \chi_{X^+})$, we choose $(1, X^+)$ in the case of $\min \rho_p^+ = \min \rho_p^-$.

In view of the complexity bound given in Theorem 3.3 we note that the size $\hat{K}_{\tilde{g}}$ of the effective domain of the associated L-convex function $\tilde{g}$ is bounded in terms of the size of $\operatorname{dom} g$. The ℓ_1-size and ℓ_∞-size of $\operatorname{dom} g$ are denoted, respectively, by K_g in (1.2) and

$$K_g^\infty = \max\{||p - q||_\infty \mid p, q \in \operatorname{dom} g\}.$$

LEMMA 3.4. $\hat{K}_{\tilde{g}} \le K_g + nK_g^\infty \le \min[\ (n+1)K_g,\ 2nK_g^\infty\ ]$.

Proof. Take $\tilde{p} = (p_0, p)$ and $\tilde{q} = (q_0, q)$ in $\operatorname{dom} \tilde{g}$ such that $\hat{K}_{\tilde{g}} = |p_0 - q_0| + ||p - q||_1$ and either (i) $p_0 = q_0$ or (ii) $p(v) = q(v)$ for some $v \in V$. We may assume $p_0 \ge q_0$ and $p \ge q$ since $\tilde{p} \vee \tilde{q}, \tilde{p} \wedge \tilde{q} \in \operatorname{dom} \tilde{g}$ and $||(\tilde{p} \vee \tilde{q}) - (\tilde{p} \wedge \tilde{q})||_1 = ||\tilde{p} - \tilde{q}||_1$. The vectors $p' = p - p_0\mathbf{1}$ and $q' = q - q_0\mathbf{1}$ belong to $\operatorname{dom} g$. In case (i), we have $\hat{K}_{\tilde{g}} = ||p - q||_1 = ||p' - q'||_1 \le K_g$. In case (ii), we have $p_0 - q_0 = q'(v) - p'(v)$ and

$$\begin{aligned} \hat{K}_{\tilde{g}} &= |p_0 - q_0| + ||p - q||_1 \\ &= (p_0 - q_0) + \sum_{u \in V} (p(u) - q(u)) \\ &= (p_0 - q_0) + \sum_{u \in V} (p'(u) - q'(u)) + n(p_0 - q_0) \\ &= \sum_{u \ne v} (p'(u) - q'(u)) - n(p'(v) - q'(v)) \\ &\le K_g + nK_g^\infty. \end{aligned}$$

Note finally that $K_g \le nK_g^\infty$ and $K_g^\infty \le K_g$. □

4. Discussion.

4.1. Scaling algorithm. Scaling is one of the common techniques in designing efficient algorithms. This is also the case with L- or M-convex function minimization. We deal with L-convex function minimization to demonstrate an implication of our result stated in Theorem 3.3.

A scaling algorithm to minimize an L-convex function g finds a minimizer of the scaled function $g_\alpha(q) = g(p^\circ + \alpha q)$ for $\alpha = \alpha^\circ, \alpha^\circ/2, \alpha^\circ/4, \alpha^\circ/8, \ldots$, starting with a

sufficiently large α° (a power of 2) until reaching $\alpha = 1$, where p° is an initial solution. For each α, g_α is an L-convex function, which can be minimized, e.g., by the steepest descent algorithm. The scaling algorithm reads as follows, where

$$\hat{K}_g^\infty = \max\{||p - q||_\infty \mid p, q \in \operatorname{dom} g,\ p(v) = q(v) \text{ for some } v \in V\}$$

and $r = 0$ in (TRF).

Scaling algorithm for an L-convex function g.

S0: Find a vector $p \in \operatorname{dom} g$, and set $\alpha := 2^{\lceil \log_2(\hat{K}_g^\infty/2n)\rceil}$.

S1: Find an integer vector q that minimizes $g(p + \alpha q)$ and set $p := p + \alpha q$.

S2: If $\alpha = 1$, then stop (p is a minimizer of g).

S3: Set $\alpha := \alpha/2$ and go to S1.

The success of this scaling approach hinges on the efficiency of the minimization in step S1. By a proximity theorem due to [12] (see Proposition 8.9 in [20] or Theorem 7.18 in [21]) there exists a minimizer q of $g(p + \alpha q)$ such that $\mathbf{0} \le q \le (n-1)\mathbf{1}$. Our complexity bound (Lemma 3.2 or Theorem 3.3) guarantees that the steepest descent algorithm with tie-breaking rule (3.1) finds the minimizer in step S1 in $O((\sigma(n)F_g + \tau(n))n^2)$ time. The number of executions of step S1 is bounded by $\lceil \log_2(\hat{K}_g^\infty/2n)\rceil$, and at the termination of the algorithm in step S2 with $\alpha = 1$, p is a minimizer of g by Lemma 3.1. Thus the result of the present paper guarantees the efficiency of the scaling approach based on steepest descent algorithm.

It is in order here to compare our algorithm with the scaling algorithm of [9], which is described in [20]. In [9] step S1 above is performed via submodular set-function minimization over a ring family on a ground set of cardinality $\le n^2$. This is based on the general fact (Birkhoff's representation theorem) that any distributive lattice can be represented as a boolean lattice over a ground set, and the size of the ground set is equal to the length of a maximal chain of the distributive lattice. Thus the minimization of the scaled function in step S1 can be carried out with $O(\sigma(n^2))$ evaluations of g. Although the complexity of this algorithm for step S1 is bounded by a polynomial in n, the algorithm is not easy to implement and will be slow in practice. Our steepest descent algorithm above is much simpler, both conceptually and algorithmically, and will be faster in practice, performing the minimization of the scaled function in step S1 with $O(\sigma(n)n^2)$ evaluations of g. Note that $\sigma(n)n^2$ is smaller in order than $\sigma(n^2)$ if $\sigma(n) = n^s$ with $s > 2$.

As for M-convex function minimization, a similar scaling approach works, provided that the scaled function $f_\alpha(y) = f(x + \alpha y)$ remains M-convex for any α and x, although this is not always the case; see [16]. See [25] and [26] for more sophisticated scaling algorithms for M-convex function minimization.

4.2. Integrally convex functions. Global optimality is characterized by local optimality also for integrally convex functions, of which M-convex and L-convex functions are special cases. Namely, it is known [4] that, for an integrally convex function f, a point x in $\operatorname{dom} f$ is a global minimizer of f if and only if $f(x) \le f(x + \chi_Y - \chi_Z)$ for all disjoint subsets $Y, Z \subseteq V$. This fact would naturally suggest the following generic scheme of steepest descent algorithms for minimizing an integrally convex function.

Steepest descent scheme for an integrally convex function f.

S0: Find a vector $x \in \operatorname{dom} f$.

S1: Find disjoint $Y, Z \subseteq V$ that minimize $f(x - \chi_Y + \chi_Z)$.

S2: If $f(x) \le f(x - \chi_Y + \chi_Z)$, then stop ($x$ is a minimizer of f).

S3: Set $x := x - \chi_Y + \chi_Z$ and go to S1.

The steepest descent algorithms for M-convex and L-convex functions in sections 2 and 3 both fit in this generic form. It is emphasized, however, that for a general integrally convex function no efficient algorithm for step S1 is available, whereas we do have polynomial time algorithms for M-convex and L-convex functions.

Acknowledgments. The author thanks Satoru Fujishige, Shiro Matuura, and Akihisa Tamura for helpful comments, and the anonymous referees for pointing out a flaw in the earlier version of the tie-breaking rule for M-convex function minimization.

REFERENCES

[1] W. J. Cook, W. H. Cunningham, W. R. Pulleyblank, and A. Schrijver, *Combinatorial Optimization*, John Wiley and Sons, New York, 1998.
[2] A. W. M. Dress and W. Wenzel, *Valuated matroid: A new look at the greedy algorithm*, Appl. Math. Lett., 3 (1990), pp. 33–35.
[3] J. Edmonds, *Submodular functions, matroids and certain polyhedra*, in Combinatorial Structures and Their Applications, R. Guy, H. Hanani, N. Sauer, and J. Schönheim, eds., Gordon and Breach, New York, 1970, pp. 69–87.
[4] P. Favati and F. Tardella, *Convexity in nonlinear integer programming*, Ricerca Operativa, 53 (1990), pp. 3–44.
[5] A. Frank, *An algorithm for submodular functions on graphs*, in Bonn Workshop on Combinatorial Optimization, Ann. Discrete Math. 16, North-Holland, Amsterdam, New York, 1982, pp. 97–120.
[6] S. Fujishige, *Theory of submodular programs: A Fenchel-type min-max theorem and subgradients of submodular functions*, Math. Program., 29 (1984), pp. 142–155.
[7] S. Fujishige, *Submodular Functions and Optimization*, Ann. Discrete Math. 47, North-Holland, Amsterdam, 1991.
[8] S. Fujishige and K. Murota, *Notes on L-/M-convex functions and the separation theorems*, Math. Program., 88 (2000), pp. 129–146.
[9] S. Iwata, Oral presentation at Workshop on Matroids, Matching, and Extensions, University of Waterloo, ON, Canada, 1999.
[10] S. Iwata, *A faster scaling algorithm for minimizing submodular functions*, in Integer Programming and Combinatorial Optimization, W. J. Cook and A. S. Schulz, eds., Lecture Notes in Comput. Sci. 2337, Springer-Verlag, New York, 2002, pp. 1–8.
[11] S. Iwata, L. Fleischer, and S. Fujishige, *A combinatorial, strongly polynomial-time algorithm for minimizing submodular functions*, J. ACM, 48 (2001), pp. 761–777.
[12] S. Iwata and M. Shigeno, *Conjugate scaling algorithm for Fenchel-type duality in discrete convex optimization*, SIAM J. Optim., 13 (2002), pp. 204–211.
[13] L. Lovász, *Submodular functions and convexity*, in Mathematical Programming—The State of the Art, A. Bachem, M. Grötschel, and B. Korte, eds., Springer-Verlag, Berlin, 1983, pp. 235–257.
[14] S. T. McCormick, *Submodular function minimization*, in Handbook on Discrete Optimization, K. Aardal, G. Nemhauser, and R. Weismantel, eds., Elsevier Science, Amsterdam, to appear.
[15] B. L. Miller, *On minimizing nonseparable functions defined on the integers with an inventory application*, SIAM J. Appl. Math., 21 (1971), pp. 166–185.
[16] S. Moriguchi, K. Murota, and A. Shioura, *Scaling algorithms for M-convex function minimization*, IEICE Trans. Fundamentals of Electronics, Communications and Computer Sciences, E85-A (2002), pp. 922–929.
[17] K. Murota, *Convexity and Steinitz's exchange property*, Adv. Math., 124 (1996), pp. 272–311.
[18] K. Murota, *Discrete convex analysis*, Math. Program., 83 (1998), pp. 313–371.
[19] K. Murota, *Algorithms in discrete convex analysis*, IEICE Trans. Systems and Information, E83-D (2000), pp. 344–352.
[20] K. Murota, *Discrete Convex Analysis—An Introduction*, Kyoritsu, Tokyo, 2001 (in Japanese).
[21] K. Murota, *Discrete Convex Analysis*, SIAM, Philadelphia, 2003.
[22] K. Murota and A. Shioura, *M-convex function on generalized polymatroid*, Math. Oper. Res., 24 (1999), pp. 95–105.
[23] A. Schrijver, *A combinatorial algorithm minimizing submodular functions in strongly polynomial time*, J. Combin. Theory Ser. B, 80 (2000), pp. 346–355.

[24] A. Shioura, *Minimization of an M-convex function*, Discrete Appl. Math., 84 (1998), pp. 215–220.

[25] A. Shioura, *Fast scaling algorithms for M-convex function minimization with application to the resource allocation problem*, Discrete Appl. Math., 134 (2003), pp. 303–316.

[26] A. Tamura, *Coordinatewise domain scaling algorithm for M-convex function minimization*, in Integer Programming and Combinatorial Optimization, W. J. Cook and A. S. Schulz, eds., Lecture Notes in Comput. Sci. 2337, Springer-Verlag, New York, 2002, pp. 21–35.

SIAM J. OPTIM.
Vol. 14, No. 3, pp. 708–731

ON ACCELERATED RANDOM SEARCH*

M. J. APPEL†, R. LABARRE‡, AND D. RADULOVIĆ§

Abstract. A new variant of pure random search (PRS) for function optimization is introduced. The basic finite-descent accelerated random search (ARS) algorithm is simple: the search is confined to shrinking neighborhoods of a previous record-generating value, with the search neighborhood reinitialized to the entire space when a new record is found. Local maxima are avoided by including an automatic restart feature which reinitializes the search neighborhood after some number of shrink steps have been performed.

One goal of this article is to provide rigorous mathematical comparisons of ARS to PRS. It is shown that the sequence produced by the ARS process converges, with probability one, to the maximum of a continuous objective function faster than that of the PRS process by adjustably large multiples of the time step (Theorem 1). Regarding an infinite-descent (no automatic restart) version of ARS, it is shown that if the objective function satisfies a local nonflatness condition, then the right tails of the distributions of inter-record times are exponentially smaller than those of PRS (Theorem 3).

Performance comparisons between ARS, PRS, and three quasi-Newton-type optimization routines are reported in attempting to find extrema of (i) each of a small collection of standard test functions of two variables, and (ii) d-dimensional polynomials with random roots. Also reported is a three-way performance comparison between ARS, PRS, and a simulated annealing algorithm in attempting to solve traveling salesman problems.

Key words. random search, stochastic optimization, global optimization

AMS subject classifications. 65K10, 90C15

DOI. 10.1137/S105262340240063X

1. Introduction.

Let f be a real-valued function with compact support $D \subset \mathbf{R}^d$. Pure random search (PRS), the quintessential Monte Carlo optimization technique for attempting to solve

$$\max_{x \in D} f(x), \tag{1.1}$$

consists of sampling a stream of independent and identically distributed (i.i.d.) random vectors $\{X_i\}_{i=1}^n$ with uniform distribution on D and then computing $M_n = \max\{f(X_i) : i = 1, \ldots, n\}$. PRS is very easy to implement. The sequence $\{M_n\}$ is guaranteed to converge, with probability one, to the essential supremum of f, ess sup $f \equiv \sup\{r : \Pr\{x : f(x) > r\} > 0\}$, if f is merely measurable. (The essential supremum is equivalent to the maximum in (1.1) when the latter exists.) Unfortunately, convergence is extremely slow in most cases of interest (e.g., where standard calculus-based methods are not applicable, where d is very large, etc.).

Much attention has been devoted to modifying PRS to improve its convergence rate. Our investigation can be put into the following framework: given an initial sequence $\{X_i\}_{i=1}^k$ of candidates for $\arg\max_{x \in D} f(x)$ and an associated sequence of probability distributions $\{P_i\}_{i=1}^k$ on D, a new candidate Y is obtained by first sam-

*Received by the editors January 7, 2002; accepted for publication (in revised form) June 5, 2003; published electronically December 19, 2003.

http://www.siam.org/journals/siopt/14-3/40063.html

†Mortgage Guaranty Insurance Corporation, Milwaukee, WI 53202 (martin_appel@mgic.com).

‡United Technologies Research Center, East Hartford, CT 06108 (labarre@utrc.utc.com).

§Department of Statistics, Yale University, New Haven, CT 06520-8290 (dragan.radulovic@yale.edu).

pling from a distribution P_{k+1} and then taking

$$X_{k+1} = \begin{cases} Y \text{ if } f(Y) > f(X_k), \\ X_k \text{ if } f(Y) \le f(X_k). \end{cases} \tag{1.2}$$

Some authors [15, 8, 2, 14] allow P_k to vary only slightly, if at all, so as to remain absolutely continuous with respect to the dominating measure (usually Lebesgue measure), thus preserving convergence; improvements are mainly due to refinements made to the decision function (1.2). There are approaches [31, 4] involving adaptive construction of P_k's which assign more mass to promising regions of the search space. The standard theme here is to direct a localized search either by using the gradient of the objective function or by partitioning the domain D and reassigning mass accordingly. Others [12, 26, 27] develop time-based triggers which shrink the domain of P_k by some factor, depending on how many records (i.e., k such that $X_{k+1} \neq X_k$) have been observed. Techniques belonging to the latter two classes are harder to analyze rigorously. In order to ensure convergence, additional assumptions must be made with respect to the surface of objective function, its gradient, etc. Proofs of convergence are very uncommon; even more rare are rigorous results showing an acceleration in the convergence rate relative to PRS.

The ARS algorithm. We propose the following algorithm for solving (1.1). We assume that D is the d-dimensional unit hypercube $[0,1]^d$. We denote by $\|\cdot\|$ the sup-norm on D. The closed ball of radius r centered at x, $\{y \in D : \|x - y\| \le r\}$, is denoted by $B(x,r)$. Thus, $D = D \cap B(x,1)$ for any x in D. We work with a real-valued, measurable objective function f on domain D.

Let a contraction factor $c > 1$ and a precision threshold $\rho > 0$ be given.

Step 0. Set $n = 1$ and $r_1 = 1$. Generate X_1 from a uniform distribution on D.

Step 1. Given $X_n \in D$ and $r_n \in (0,1]$, generate Y_n from a uniform distribution on $B(X_n, r_n)$.

Step 2. If $f(Y_n) > f(X_n)$, then let $X_{n+1} = Y_n$ and $r_{n+1} = 1$.
Else if $f(Y_n) \le f(X_n)$, then let $X_{n+1} = X_n$ and $r_{n+1} = r_n/c$.
If $r_{n+1} < \rho$, then $r_{n+1} = 1$.
Increment $n := n + 1$ and go to Step 1.

We refer to the algorithm just described as *finite descent ARS* (accelerated random search). We refer to the sequence $\{X_n\}$ as the sequence of *record generators* and the sequence $\{M_n = f(X_n)\}$ as the *record sequence*. The ARS algorithm first appeared in [1].

At first glance the algorithm we propose may seem counterintuitive, in that we shrink the search space if the new candidate Y is *not* an improvement; otherwise, we reinitialize the search space to include all of D, while keeping our last record generator. Due to the exponentially contracting radii, ARS will shrink the search space very fast and ultimately sample only in neighborhoods of local maxima.

Consider the following example. Let f be continuous on $[0,1]$, and set $c = 2$ and $\rho = 2^{-20}$. Given an initial X_0 in the unit interval, ARS samples random points from the balls $B(X_0, 1/2^i)$, $i = 1, 2, \ldots$, until either a candidate Y satisfies $f(Y) > f(X_0)$ or $i = 20$, in which case the algorithm restarts, sampling from the entire unit interval. Either way, ARS will eventually produce a sequence of record generators X_k such that X_k converges to some X^* in $\arg\max_{x \in D} f(x)$ and $f(X_k)$ converges to $f^* \equiv f(X^*)$.

Once X_k is close enough to X^*, the advantage of ARS vis-à-vis PRS is revealed. Suppose that $|X_k - X^*| \leq 10^{-3}$. In this case, PRS will need to sample approximately 10^3 points before it finally hits the relevant ball $B(X_k, 1/2^{10})$ (note $2^{-10} \approx 10^{-3}$). By contrast, ARS will sample from the relevant ball after only 10 iterations.

We note that the ARS algorithm may be readily extended to problems defined over a general metric space, for example, to combinatorial optimization problems and to optimization problems whose domains involve complicated constraints. The theoretical results contained herein extend as well after simple modifications.

From a practical standpoint, there is nothing to be gained by allowing the precision threshold ρ to take values smaller than machine precision. However, from a theoretical point of view it is interesting to consider the case where $\rho = 0$. We call the following algorithm *infinite descent ARS*.

Let a contraction factor $c > 1$ be given.

Step 0. Set $n = 1$ and $r_1 = 1$. Generate X_1 from a uniform distribution on D.

Step 1. Given $X_n \in D$ and $r_n \in (0, 1]$, generate Y_n from a uniform distribution on $B(X_n, r_n)$.

Step 2. If $f(Y_n) > f(X_n)$, then let $X_{n+1} = Y_n$ and $r_{n+1} = 1$.
Else If $f(Y_n) \leq f(X_n)$, then let $X_{n+1} = X_n$ and $r_{n+1} = r_n/c$.
Increment $n := n + 1$ and go to Step 1.

The only change here is the removal of the radius restart logic (if $r_{n+1} < \rho$, then $r_{n+1} = 1$). However, this has major ramifications for the algorithm, as convergence of the record sequence to the essential supremum of the objective function no longer occurs with probability one. For example, any locally constant function will, with positive probability, cause infinite descent ARS to stall at a local maximum.

We remark that ARS bears some resemblance to the classic simulated annealing (SA) algorithm [11] in that the typical ARS realization of some m points sampled uniformly from the shrinking balls $B(X_k, 1), B(X_k, 1/c), \ldots, B(X_k, 1/c^m)$ is similar statistically to m points sampled from a symmetric distribution with exponential tails centered at X_k. The latter is the net result of a sequence from SA's probabilistic scheme of accepting a candidate which does not offer an improvement with a probability which is log-linearly proportional to the difference between the "energy" of the current best point and the new candidate. One notable difference between ARS and SA is that the number of contractions prior to restart in ARS depends in an automatic way on the topology of the function surface, whereas SA relies on a prescribed "cooling schedule" for its stream of exponential probability models.

We also note that the ARS algorithm fits nicely into the meta-approach of Solis and Wets [29]. Their basic conceptual algorithm can sample randomly from uniform distributions on sup-norm balls centered at the current record generator whose radius shrinks by a constant factor if no improvement is found and stretches (to full unit value) otherwise. Mathematical rigor is limited there to proofs of convergence only, without results on rates of convergence.

In other related work, there is a fair amount of literature (see [17]) on pure adaptive search (PAS). The core concept in PAS is to draw the next point X_{k+1} uniformly from the region $R_k = \{x : f(x) > f(X_k)\}$. There are results which establish an exponential improvement of PAS over PRS in the case of Lipschitz continuous [32] and finite-valued objective functions [34]. An extension to SA has also been investigated [5, 21]. The drawback, of course, is that PAS is not implementable, since

the region R_k is rarely known. However, various schemes for sampling from R_k in an efficient or approximate way have been considered [21, 30, 33]. One idea is to construct a candidate approximant to R_k by covering with (possibly shrinking) balls a subset of search iterates consisting of those points whose objective function value is larger than a given value, and then sampling uniformly from this covering [9]. ARS may be thought of in this context since for a "nice" function f—for example, if f is strictly concave in a neighborhood of $\arg\max_{x\in D} f(x)$—there exists a constant $\lambda > 0$ which does not depend on k such that for k large enough, $\mathrm{vol}(B(X_k, r) \cap R_k)/\mathrm{vol}(B(X_k, r)) > \lambda$ for all r small. In the concave case in one dimension, we have $\lambda = 1/2$, which ensures that a sample from the ball has a 50% chance of being from the preferred region R_k. Attempts to combine PAS and its extensions with directional search techniques have been made [22, 3, 10]. With these methods, as with the previously mentioned methods involving time-based triggers, it would seem difficult to pinpoint conditions necessary to prevent the search from becoming trapped away from local extrema.

In the next section, we state our main theoretical results; proofs and complements are delayed until Appendix B. In section 3, we report the results of applications of ARS to a variety of numerical optimization problems. The paper is briefly summarized in section 4. Throughout the paper, and often in the proofs of the theoretical results, we use standard notation and terminology from probability theory and real analysis; a reader unfamiliar with these is urged to consult any of a number of standard graduate-level texts in probability theory; see, for example, [7].

2. Statement of theoretical results.

THEOREM 1. *Let f be a continuous function on the d-dimensional unit cube with finitely many global maxima. Let $\{M_n\}$ be the record sequence produced by finite descent ARS, and let $\{\widetilde{M}_n\}$ be the record sequence produced by PRS.*

Given a contraction factor $c > 1$ and a precision threshold $\rho \in (0, 1)$, let $m = |\ln(\rho)|/\ln(c)$. For each positive integer $C < c^m/(3m)$ there exists a positive integer n_C, depending only on C, such that for each $n > n_C$

$$E(M_n) \geq E\left(\widetilde{M}_{Cn}\right). \tag{2.1}$$

Here, $E(\cdot)$ stands for the expected value operator. The content of Theorem 1 is essentially that for a continuous objective function, one may pick a multiplier C from a range (depending only on c and ρ) of large constants, and that after a (possibly large) number of initial steps, finite descent ARS at n steps outperforms PRS at Cn steps, in the sense of L^1 (and hence with probability one if the two processes are run simultaneously on the same space, since f is bounded). We note that the assumption of continuity is somewhat more than what is needed for Theorem 1 to hold; all we really require is the existence of a certain "maximal absorbing" set. The reader is referred to the proof in Appendix B for details.

For infinite descent ARS, convergence to local or global maxima is no longer guaranteed. In fact, the algorithm may stall on some very smooth functions; that is, the event $\{X_{n+1} = X_n$, eventually$\}$ may occur with positive probability, where the common value of X_n beyond some index is only a local optimum. However, we do have convergence when the complementary event $\{X_{n+1} \neq X_n$ infinitely often (i.o.)$\}$ occurs.

THEOREM 2. *Let $\{X_n\}$ denote the record-generating sequence from infinite descent ARS. The record sequence converges, with probability one, to* ess sup f *on the event $\{X_{n+1} \neq X_n$ i.o.$\}$.*

Theorem 2 implies that $P(M_n \to \text{ ess sup} f) = P(X_{n+1} \neq X_n \text{ i.o.})$; the common value of the two probabilities is no longer necessarily one. In Appendix B (Proposition 7), we provide a characterization of $P(X_{n+1} \neq X_n \text{ i.o.}) = 1$ in terms of a product criterion. The sufficiency half of this criterion is similar to the global convergence theorem of Solis and Wets [29].

Our third result states that for a class of objective functions satisfying a local non-"flatness" condition, the right tails of the distributions of inter-record times in infinite descent ARS are exponentially smaller than those of PRS.

PROPERTY B. *For almost every* $x \in D$,

$$\sup_{r>0} P\{y : f(y) \leq f(x) \,|B(x,r)\} < 1. \tag{2.2}$$

THEOREM 3. *Let* f *satisfy Property B and let* $\{X_n\}$ *denote the record-generating sequence from infinite descent ARS. Let* $\{N_k\}$ *be the sequence of random times at which a new record is found,*

$$N_1 = \min\{n : X_{n+1} \neq X_n\}, \tag{2.3}$$

and, recursively,

$$N_{k+1} = \min\{n > N_k : X_{n+1} \neq X_n\}. \tag{2.4}$$

Let $\Delta_{k+1} = N_{k+1} - N_k$ *denote the inter-record times. Let* $\widetilde{X}_n$, $\widetilde{N}_k$, *and* $\widetilde{\Delta}_{k+1}$ *denote the analogous quantities from PRS. Then there are constants* $0 \leq \eta < 1$ *and* $K > 0$ *such that*

$$P(\Delta_k > n) \leq \eta^n P(\widetilde{\Delta}_k > n) \text{ for all } k \geq K. \tag{2.5}$$

3. Applications. We now report results from computational experiments with ARS. The first three sets involve performance comparisons between ARS, PRS, and three quasi-Newton-type optimization routines in attempting to find global optima of each of a small collection of test functions of two variables as well as a collection of polynomials on the d-dimensional unit hypercube with random roots. In the fourth set, we report the performance of ARS versus PRS and SA in attempting to find minimum-weight Hamiltonian paths through a complete graph (traveling salesman problem).

The quasi-Newton-type methods used were (i) *Mathematica*'s [13] default optimization routine, a modified Powell's method routine (denoted here by MDF); (ii) *Mathematica*'s quasi-Newton method (MQN); and (iii) Algorithm 573 [6], as downloaded from the Internet [16]. Algorithm 573, also known as N12SOL, is another Powell-type quasi-Newton method.

Since the candidate Y_{n+1}, generated immediately after finite descent ARS restarts, is uniformly distributed on the entire space $D = B(X_{n+1}, r_{n+1}) = B(X_{n+1}, 1)$, it follows that the subsequence of the record sequence that corresponds to these candidates provides us with a PRS sequence. Thus, with probability one, finite descent ARS will never permanently stall at a local but nonglobal maximum. However, in the experiments described below we used a modification of finite descent ARS in order to speed up our experimental work. Below, the term *modified ARS* refers to the following: we keep track of the number N_{restart} of times finite descent ARS restarts ($r_n < \rho$ and so $r_{n+1} = 1$). If $N_{\text{restart}} = n > L$, a prescribed positive integer, then we restart the entire ARS process, setting $r_{n+1} = 1$, $N_{\text{restart}} = 0$, and generating X_{n+1} from a uniform

TABLE 1
Standard test functions.

Test function	N_{ARS}	N_{MQN}		N_{MDF}		N_{573}
Rosenbrock	1256	188		210		119
Himmelblau	188	26		57		29
Freudenstein–Roth	1693	99		166		121
Jennrich–Sampson	438	65		138		58
Griewank	539	54		56		43
Rastrigin	1640	550	(1)	817	(5)	1017
Gaussian 1	2723	413		388	(12)	342
Gaussian 2	16977	545		485	(22)	475

distribution on D; X_n is declared a local maximum and is not used in subsequent iterations of the algorithm.

We note that modified ARS is very robust with respect to the choice of the parameter L. We chose $L = 4$ everywhere but in the traveling salesman problems. We have experimented with different (small) values and can report no observable change in performance. We worked with the dimension-dependent choice of $L = d^2$ in the traveling salesman problems in order to counter the rapidity with which ARS shrinks its search space there.

Standard test functions. We report results for a variety of test functions, the first six of which are standard in the optimization literature. Some mathematical formulae appear in Appendix A. All experiments were conducted using modified ARS with $L = 4$, contraction factor $c = \sqrt{2}$, and precision threshold $\rho = 10^{-4}$.

The whole point here is to demonstrate that ARS outperforms PRS by several orders of magnitude. For each problem, we first ran PRS for 1,000,000 iterations and recorded the final (minimum value) record attained. We then ran modified ARS until it found a record whose value was no greater than that found by PRS and recorded the number of iterations needed. We repeated the whole process 100 times. The same experiment was repeated using algorithms MDF, MQN, and 573, starting from random points, with one addition: since each of the quasi-Newton-type methods is really designed to look for local extrema, we aided them within each trial by repeatedly restarting from a new random starting point until either a better result than PRS was found or the number of such restarts exceeded a maximum number T_{restart}. We set $T_{\mathrm{restart}} = 50$ here. All internal stopping criteria and parameters for these quasi-Newton algorithms were taken as the subroutines' defaults.

The results are summarized in Table 1. The average number (over 100 trials) of function evaluations required by each of ARS, MDF, MQN, or 573 to find a better record than PRS is denoted below as N_{ARS}, N_{MDF}, N_{MQN}, and N_{573}, respectively. For example, in 100 trials on Rastrigin's function, ARS needed on average 1640 iterations to find a better record than PRS after 1,000,000 iterations, while Algorithm 573 needed 1017. The values in parentheses stand for the number of trials in which the given method failed to beat PRS after reaching the maximum of 50 restarts. For example, the MQN routine exceeded its allotted 50 restarts on one of the 100 trials; thus its average value of 550 is an average over 99 successes. Similarly, MDF's average of 817 is over 95 trials, as MDF exceeded its 50-restart allotment on 5 trials.

Clearly, ARS significantly outperformed PRS in each of these experiments, which is consistent with Theorem 1. We get a sense of the value of the constant C in the theorem by simply dividing N_{ARS} into 1,000,000. While all of these test functions are smooth with two-dimensional domains, there is some variety in terms of scaling issues,

TABLE 2
Brownian sheet.

Group	Avg PRSmax	Avg ARSmax	N_{ARS}	Max	Min
1	4.51E-02	4.60E-02	3049	12177	154
2	1.76E-01	1.77E-01	1638	4837	160
3	1.64E-01	1.64E-01	2650	8416	179
4	2.85E-02	2.91E-02	12881	38782	276
5	6.87E-02	6.90E-02	1104	2517	177

the existence of local extrema (Griewank and Rastrigin), near-flatness near global extrema (Himmelblau), and "spiky" functions with hard-to-find extrema (Gaussian functions).

Brownian sheet. Consider the function $f(s,t) = B^1(s) \cdot B^2(t)$, where B^1 and B^2 are independent Brownian motions on the unit interval. We simulate Brownian motion here by first generating i.i.d. standard normal variates $Z_1, \ldots, Z_n$ and then computing $B(t) = n^{-1/2} \sum_{i=1}^{[nt]} Z_i$, where the resolution was set to $n = 10^6$ and $[x]$ denotes the integer part of x. As Brownian motion is everywhere continuous but nowhere differentiable, any calculus-based method can be expected to perform poorly on it; the three quasi-Newton-type methods MQN, MDF, and 573 each failed to produce meaningful results here.

On each of 100 realizations of a Brownian sheet, we again ran PRS 1,000,000 times, recorded the (maximum value) record achieved, and then ran modified ARS until it reported a better (larger) result. All experiments were conducted using modified ARS with $L = 4$, contraction factor $c = 2^{1/2}$, and precision threshold $\rho = 10^{-4}$. The summary statistics over 5 groups of 20 runs appear in Table 2. Again, N_{ARS} is the average number of function evaluations used by ARS, while Max and Min here are the maximum and minimum number, respectively, of ARS function evaluations over the 20-run trial.

Polynomials with random roots. For fixed integers d and m we construct a set of random univariate polynomials on [0,1] as follows: for each $i = 1, \ldots, d$, let $Z_{i,j}$, $j = 1, \ldots, m-1$, be i.i.d. random variables with uniform distribution on $[0,1]$. Taking $0, Z_{i,1}, \ldots, Z_{i,m-1}, 1$ to be the zeros of the ith polynomial $p_i(\cdot)$, define $p_i(u) = u(u-1)\prod_{j=1}^{m-1}(u - Z_{i,j})$. For x in the d-dimensional unit hypercube we now define a test function $f(x) = d^m \cdot \prod_{i=1}^{d} p_i(x_i)$. Thus, f is a product of polynomials with random roots in $[0,1]$. The magnification factor d^m before the product is irrelevant to PRS and ARS, but we found that the quasi-Newton-type methods have difficulty in regions where the objective is near zero.

For each d, m pair, we performed the following experiment 100 times: we first generated a random polynomial-product test function, then ran PRS for $N_{\text{PRS}} = d^2(m)^2 10^4$ trials, followed by runs of each of modified ARS with $L = 4$, contraction factor $c = 2^{1/d}$, and precision threshold $\rho = (10 N_{\text{PRS}})^{-1/d}$, MQN, MDF, and Algorithm 573 until each respective record was larger than that found by PRS.

Table 3 shows the average number of trials required by each method to find a better record than PRS. As before, the values in parentheses indicate the number of trials in which the given method failed to beat PRS before reaching the maximum number of restarts, $T_{\text{restart}} = 50d$.

These results provide a dramatic demonstration of the acceleration in performance predicted by Theorem 1. For each d, m pair, ARS required orders of magnitude fewer iterations to achieve performance on par with PRS. Perhaps more surprising is the

TABLE 3
Polynomials with random roots.

d	m	N_{PRS}	N_{ARS}	N_{MQN}		N_{MDF}		N_{573}	
1	3	9000	115	16		26		12	
1	4	16000	103	19		23		22	
1	5	25000	120	25		22		31	
1	9	81000	216	61		31		48	
1	11	121000	166	87		37		72	
1	13	169000	194	76		22		38	
2	3	360000	258	43		89	(3)	54	
2	4	640000	262	93		119	(3)	142	
2	5	1000000	423	147	(2)	198	(1)	183	
2	6	1440000	518	335		329		226	
3	3	810000	419	306	(1)	217	(4)	202	(3)
3	4	1440000	540	1681	(19)	447	(1)	523	(11)
3	5	2250000	813	2118	(13)	841	(1)	753	(9)
3	6	3240000	765	3256	(65)	2012	(5)	1623	(23)
4	3	1440000	800	-		-		-	
4	4	2560000	1004	-		-		-	
4	5	4000000	1454	-		-		-	
4	6	5760000	3704	-		-		-	
5	3	2250000	1663	-		-		-	
5	4	4000000	1965	-		-		-	
5	5	6250000	3989	-		-		-	
5	6	9000000	5858	-		-		-	

comparison of ARS to the three calculus-based methods. For $d = 1$ and 2 the results are as we might have expected from Table 1. In the trials with $d = 3$, we begin to see a degradation in performance of the quasi-Newton-type methods relative to ARS, which worsens as the complexity (m) of the function surface increases. We believe that this behavior is due to the existence in these polynomials of numerous local maxima coupled with wide, nearly flat regions: $f(x) = 0$ whenever $x = Z_{i,j} \times [0,1]^{d-1}$, $i = 1, \ldots, d$, $j = 1, \ldots, m-1$. We recall (Table 1) the relative difficulty that the quasi-Newton-type methods had both with Rastrigin's function, with its dozens of local minima, and with the Gaussian functions.

We found that the quasi-Newton-type methods performed reasonably well when the random starting point was in a neighborhood of $\arg\max_{x \in D} f(x)$. But this of course becomes increasingly unlikely as dimension and degree grow—hence the numerous restarts, all attracted to local maxima. We therefore report no trials using algorithms MQN, MDF, or 573 for $d > 3$, as the performance degradation only worsened there.

Traveling salesman problems. The following example is intended to demonstrate the wider applicability of ARS, where all that is really required for implementation is that the objective function be computable and that its domain be a metric space from which one can simulate uniform random variates.

For a given dimension d, we produced a realization of a randomly weighted complete graph on d vertices by generating a $d \times d$ symmetric matrix with random (uniform on $[0,1]$) entries. We then searched for a minimum-weight (Hamiltonian) path through the graph using three techniques: PRS, modified ARS, and a version of the SA algorithm [11].

A path through the complete graph is uniquely represented by a permutation of the ordered string $1, 2, \ldots, d$. PRS was performed by generating a predetermined number (N_{PRS}) of random permutations, computing the weight (sum over all edge

TABLE 4
Traveling salesmen.

d	$N_{\mathrm{PRS}} = d^6$	N_{ARS}	N_{SA}
10	1000000	23565	72479
11	1771561	12306	11431
12	2985984	8619	22319
13	4826809	11403	23293
14	7529536	14936	22290
15	11390625	3380	15183
16	16777216	19402	43862
17	24137569	5368	12334
18	34012224	4693	6709
19	47045881	2431	8927
20	64000000	4233	6534
21	85766121	2028	5654

TABLE 5
More traveling salesmen.

d	$N_{\mathrm{PRS}} = 2((d-1)!)$	N_{ARS}	N_{SA}
8	10080	171	686
9	80640	989	162
10	725760	9531	87286
11	7257600	14377	3770
12	79833600	56148	219145
13	958003200	478162	741741

weights) of each, keeping the current lowest.

In order to apply ARS, we first placed a metric on the space of paths (permutations) by taking the distance between two paths π_1 and π_2 to be the smallest number of pairwise transpositions needed to transform π_1 into π_2. For example, the distance between $1, 2, 3$ and $3, 1, 2$ is 2. Although this is a discrete problem, one can argue that if two paths π_1 and π_2 are close (with respect to this distance), then the sums over all edge weights are also close.

We implemented SA according to the following algorithm: given a path π_1, sample a new path π_2 randomly from the ball $B(\pi_1, d/2)$ (with respect to the above metric). Let $w(\pi)$ denote the weight of a path π. If $w(\pi_2) < w(\pi_1)$, then keep π_2 as the new candidate; if not, compute the probability $p = \exp(-3(w(\pi_1) - w(\pi_2)))$, and then keep or reject π_2 as the new candidate with probability p or $1 - p$, respectively.

In our modified ARS, we took $L = d^2$, contraction factor $c = 1.3$, and precision threshold $\rho = 1$. As in the previous experiments, we ran ARS and the SA procedure until the reported minimum was smaller than the one obtained by PRS. For each dimension d we repeated the experiment 10 times and recorded the averages.

As before, ARS outperformed PRS by orders of magnitude (see Table 4). It is important to note that none of the three methods found an actual minimum-weight path. We simply claim that, after only a few thousand steps, ARS found a path with smaller weight than did PRS after millions of iterations. The same is also true of SA, although ARS did somewhat better here as well.

To see how ARS performs in searching for an actual minimum-weight path, we conducted a second experiment in which we ran PRS for $2((d-1)!)$ times. In this case, PRS found a true minimum-weight path with probability $1-(1-1/(d-1)!)^{2(d-1)!} \geq 1 - e^{-2} \approx 0.85$ (see Table 5).

Linear systems. It is of interest to note a conspicuous failure of ARS. We considered optimizing objective functions of the form $f(x) = -\|Ax-b\|$, where x and b lie in $\mathbf{R}^d$, A is a real $d \times d$ matrix, and $\|\cdot\|$ is an l^p vector norm. Typically, ARS performed poorly in this setting. It was not unusual for ARS to require more than 10^4 iterations in order to solve a 5×5 linear system. It is of small consolation to note that ARS did perform better than PRS (which typically needs more than 1,000,000 iterations). As one would expect, the quasi-Newton-type methods performed flawlessly here, since for any convex quadratic function, Newton's method converges in one iteration.

4. Summary. We have described *accelerated random search*, a simple, but effective, new variant of the classic Monte Carlo pure random search algorithm. We have included a literature context for our work. Some mathematically rigorous theoretical results have been provided, which clearly point to improvements in performance of ARS over PRS. In order to demonstrate the efficacy of the algorithm, we have included the results of a variety of numerical optimization experiments which feature performance comparisons of ARS to PRS, as well as to standard quasi-Newton-type optimization methods (smooth functions of several variables) and a simulated annealing algorithm (traveling salesman problems).

Appendix A. List of test functions.

Rosenbrock "banana" function [23]:
$f(x,y) = 100(y-x^2)^2 + (1-x)^2$, where $x, y \in [-5,5]$. The function has a unique global minimum of 0 at $(1,1)$.

Himmelblau's function [20]:
$f(x,y) = (x^2+y-11)^2 + (x+y^2-7)^2$, where $x, y \in [-5,5]$. The function takes its minimum value of 0 at four solution points, given by the intersection of the two conic sections $y = -x^2 + 11$ and $x = -y^2 + 7$.

Freudenstein–Roth function [24]:
$f(x,y) = ((x-13) + ((5-y)y-2)y)^2 + ((x-29) + ((y+1)y-14)y)^2$, where $x, y \in [-10,10]$. The global minimum of 0 occurs at $(5,4)$.

Jennrich–Sampson function [25]:
$f(x,y) = \sum_{k=1}^{10}(2+2k-(e^{kx}+e^{ky}))^2$, where $x, y \in (-1,1)$. The global minimum of 124.362 occurs at $(0.257825, 0.257825)$.

Griewank's function [28]:
$f(x,y) = 1 + \sum_{i=1}^{n} \frac{x_i^2}{d} - \prod_{i=1}^{n} \cos(\frac{x_i}{\sqrt{i}})$, where $x, y \in (-100,100)$, with $n = 2$ and $d = 10$. The global minimum of 0 occurs at $(0,0)$.

Rastrigin's function [19]:
$f(x,y) = (x^2+y^2) - \cos(18x) - \cos(18y)$, where $x, y \in [-1,1]$. The global minimum of -2 occurs at $(0,0)$.

Gaussian function 1:
let $g(x,y;h,m,s) = (h - ((x-m_1)^2 + (y-m_2)^2)) \cdot \exp(-s((x-m_1)^2 + (y-m_2)^2))$ and construct a test function

$$f(x,y) = g(x,y;h^1,m^0,s^1) + g(x,y;h^2,m^1,s^2) + g(x,y;h^2,m^2,s^2) + g(x,y;h^2,m^3,s^2) + g(x,y;h^2,m^4,s^2)$$

with $h^1 = 5$, $h^2 = 1$, $s^1 = 10$, $s^2 = 0.5$, $m^0 = (0,0)$, $m^1 = (0,5)$, $m^2 = (0,-5)$, $m^3 = (5,0)$, and $m^4 = (-5,0)$. The function f has a global maximum at $(0,0)$ with very small support. There are also four local maxima at $(\pm 5, \pm 5)$ with large supports.

Gaussian function 2: same as above but with $s^1 = 100$.

Appendix B. Proofs and theoretical complements. Throughout the following, 1_A denotes the binary indicator random variable for set membership in A. We denote Lebesgue measure on $\mathbf{R}^d$ by μ. Let h be a measurable mapping on a probability space $(\Omega, \mathcal{F}, P)$, and let C denote a condition (set membership, an inequality, etc.). We make frequent use of two notational abbreviations, which are commonly used in probability theory:

$$\{\omega \in \Omega : h(\omega) \text{ satisfies } C\} = \{h \text{ satisfies } C\},$$

$$P(\{\omega \in \Omega : h(\omega) \text{ satisfies } C\}) = P(h \text{ satisfies } C).$$

Proof of Theorem 1. We first prove two preliminary lemmas.

LEMMA 4. *Let $D \subset \mathbf{R}^d$ such that $0 < \mu(D) < \infty$ and such that there exists $r > 0$ for which $D \subset \cap_{x \in D} B(x, r)$. Let Z_0 be uniformly distributed on D (i.e., $P(Z_0 \in C) = \mu(C \cap D)/\mu(D)$) and let $\{Z_k\}_{k=1}^{\infty}$ and $\{X_k\}_{k=0}^{\infty}$ be such that for $k \geq 0$*

$$\begin{aligned} &X_k \in \sigma_k = \sigma(Z_0, Z_1, \ldots, Z_k) \\ &\qquad \textit{(the sigma-field generated by } Z_0, Z_1, \ldots, Z_k), \\ &X_k \in D, \end{aligned}$$

$$Z_{k+1} \sim \textit{uniform on } B(X_k, r).$$

Let $y \in D^c$ and for $k > 0$, let $Y_k = Z_k \cdot 1_{\{Z_k \in D\}} + y \cdot 1_{\{Z_k \notin D\}}$. The conditional distribution $\mathcal{L}(Y_k | Y_k \neq y)$ of Y_k, given that $Y_k \neq y$, is uniform on D, and the sequence $\{Y_k\}_{k=1}^{\infty}$ is i.i.d.

Proof of Lemma 4. First we observe that $\mu_B \equiv \mu(B(X_k, r))$ does not depend on X_k. For any set A not containing y we have

$$\begin{aligned} P(Y_k \in A) &= P(Z_k \in A \cap D) \\ &= E(P(Z_k \in A \cap D | \sigma(X_{k-1}))) \\ &\quad \text{(by the definition of } Z_k \text{ and since } D \subset \cap_{x \in D} B(x, r)) \\ &= E\left(\frac{\mu(A \cap D \cap B(X_{k-1}, r))}{\mu(B(X_{k-1}, r))}\right) \\ &= E\left(\frac{\mu(A \cap D)}{\mu_B}\right) \\ &= \frac{\mu(A \cap D)}{\mu_B}, \end{aligned} \tag{B.1}$$

which does not depend on k. Similarly, for any A which contains y

$$\begin{aligned} P(Y_k \in A) &= P(Y_k \in A \setminus \{y\}) + P(Y_k = y) \\ &= \frac{\mu(A \cap D)}{\mu_B} + 1 - \frac{\mu(D)}{\mu_B}, \end{aligned}$$

which again does not depend on k. This implies that $\{Y_k\}_{k=1}^{\infty}$ is i.i.d. and that $\mathcal{L}(Y_k | Y_k \neq y)$ is uniform on D.

We now demonstrate independence of the sequence $\{Y_k\}_{k=1}^{\infty}$. For any sequence

of sets $C_i \subset \mathbf{R}^d$, $i = 1, 2, \ldots$, we have

$$\begin{aligned}
P(Y_k \in C_i|\sigma_{k-1}) &= P(Y_k \in C_i \setminus \{y\}|\sigma_{k-1}) + P(Y_k \in C_i \cap \{y\}|\sigma_{k-1}) \\
&= P(Z_k \in D \cap C_i|\sigma_{k-1}) + 1_{\{y \in C_i\}} P(Z_k \in D^c|\sigma_{k-1}) \\
&\quad \text{(again, by the definition of } Z_k) \\
&= P(Z_k \in D \cap C_i|\sigma(X_{k-1})) + 1_{\{y \in C_i\}} \cdot P(Z_k \in D^c|\sigma(X_{k-1})) \\
&\quad \text{(using the same argument as in (B.1))} \\
&= \frac{\mu(C_i \cap D)}{\mu_B} + 1_{\{y \in C_i\}} \cdot \left(1 - \frac{\mu(D)}{\mu_B}\right) \\
&\equiv q_i.
\end{aligned}$$

Since each q_i, $i = 1, 2, \ldots$, does not depend on k, nor is it random, we have

$$\begin{aligned}
P(Y_k \in C_i) &= E\left[P(Y_k \in C_i|\sigma_{k-1})\right] \\
&= q_i. \qquad \text{(B.2)}
\end{aligned}$$

Therefore

$$\begin{aligned}
P\left(\cap_{k=1}^n \{Y_k \in C_k\}\right) &= E\left[P\left(\cap_{k=1}^n \{Y_k \in C_k\}|\sigma_{n-1}\right)\right] \\
&\quad \text{(since } Y_k \in \sigma_k) \\
&= E\left(\prod_{k=1}^{n-1} 1_{\{Y_k \in C_k\}} \cdot P\left(Y_n \in C_n|\sigma_{n-1}\right)\right) \\
&\quad \text{(by (B.2))} \\
&= E\left(q_n \cdot \prod_{k=1}^{n-1} 1_{\{Y_k \in C_k\}}\right) = q_n \cdot P\left(\cap_{k=1}^{n-1} \{Y_k \in C_k\}\right) \\
&\quad \text{(by repeating the above computation } n-1 \text{ times)} \\
&= \prod_{k=1}^{n} q_k \\
&= \prod_{k=1}^{n} P\left(Y_k \in C_k\right),
\end{aligned}$$

which implies independence. This proves Lemma 4.

Before stating the next lemma we make the following definitions. Let K be the largest integer such that $1/c^K > \rho$ and let

$$\begin{aligned}
&k_1 = 1, \\
&k_i = \min\{n : n > k_{i-1} \text{ and } X_n \neq X_{k_{i-1}}\} \text{ for } i = 1, 2, \ldots, \\
&I_n = \max\{m : k_m \leq n\}, \\
&\tau_i = k_{i+1} - k_i,
\end{aligned}$$

and

$$Y_{i,j} \sim \text{uniform on } B\left(X_{k_i}, \frac{r}{c^l}\right), \text{ where } j \leq \tau_i + 1 \text{ and } l = (j-1) \bmod (K+1).$$

For given $A \subset \mathbf{R}^d$, let

$$m_L(A) = \min\left\{ n : \sum_{i=1}^{I_n - 1} \left[\frac{\tau_i}{K+1}\right] 1_{\{X_{k_i} \in A\}} + \left[\frac{n - k_{I_n}}{K+1}\right] 1_{\{X_{k_{I_n}} \in A\}} = L \right\},$$

where $[x]$ denotes the largest integer less than or equal to x. We set $m_L(A) = \infty$ if the above minimum is not defined. Essentially, $m_L(A)$ tells us how far we should go with the sequence $\{X_1, X_2, \ldots, X_{m_L(A)}\}$ in order to have exactly L of $Y_{i,j}$'s sampled uniformly on the minimal balls $B(\nu_i, \frac{1}{c^K})$ with centers $\nu_i \in A$. For $X^* \in \arg\max_{x \in D} f(x)$ and fixed $r \in \mathbf{R}$, a set A such that

$$A \subset B(X^*, r), \;\; \mu(A) > 0,$$

and

$$f(x) \geq f(y) \text{ for all x} \in A \text{ and for all } y \in D \setminus B(X^*, r)$$

is called an *absorbing set for the radius* r. A set is called the *maximal absorbing set for the radius* r if it is the union of all the absorbing sets for the radius r. It should be noted that for some r and functions f, absorbing sets might not exist. We use the term "absorbing" to describe such a set A because once the ARS sequence lands in A, it never leaves the ball $B(X^*, r)$. That is, if $X_k \in A$, then $X_{k+n} \in B(X^*, r)$ for all n. If A is a maximal absorbing set, then $X_k \in A$ implies that $X_{k+n} \in A$ for all n.

LEMMA 5. *Let f be a continuous function on the d-dimensional hypercube D with unique maximum at X^*, and let $\{X_i\}_{i=1}^{\infty}$ be a sequence of record generators produced by the ARS algorithm. Then for a maximal absorbing set A for the radius $s = 1/(3\rho)$*

$$P(m_L(A) > 3KL) \to 0 \text{ as } L \to \infty.$$

Proof of Lemma 5. Since f is continuous with unique maximizer X^*, for every $0 < \varepsilon < \rho$ the maximal absorbing set A^ε associated with radius ε exists in $B(X^*, \varepsilon)$. Since ρ is fixed

$$\mu(A^\varepsilon)/\mu(B(X^*, \rho)) \to 0 \text{ as } \varepsilon \to 0.$$

First we will show that

$$P(\tau_i \leq K) \to 0 \text{ as } i \to \infty. \tag{B.3}$$

To see this, note that

$$\begin{aligned} P(\tau_i \leq K) &= P\left(\{X_{k_i+1} \neq X_{k_i}\} \cup \cdots \cup \{X_{k_i+K} \neq X_{k_i}\}\right) \\ &\leq \sum_{j=1}^{K} P(X_{k_i+j} \neq X_{k_i}) \\ &\leq \sum_{j=1}^{K} P\left(\{X_{k_i+j} \neq X_{k_i}\} \cap \{X_{k_i} \in A^\varepsilon\}\right) + K \cdot P(X_{k_i} \notin A^\varepsilon) \end{aligned} \tag{B.4}$$

for any $\varepsilon > 0$. For $X_{k_i} \in A^\varepsilon$, we have $X_{k_i+j} \neq X_{k_i}$ only if $Y_{k_i,j} \in A^\varepsilon$ and $f(Y_{k_i,j}) > f(X_{k_i})$. Therefore

$$\begin{aligned} P\left(\{X_{k_i+j} \neq X_{k_i}\} \cap \{X_{k_i} \in A^\varepsilon\}\right) &\leq P(Y_{k_i,j} \in A^\varepsilon) \\ &\leq \frac{\mu(A^\varepsilon)}{\mu(B(X^*, \rho))}, \end{aligned}$$

which can be made arbitrarily small by the choice of ε. For the second part of (B.4) we observe that since X_{k_i} converges, with probability one, to $X^* \in A^\varepsilon$ (since ARS restarts periodically) and $\mu(A^\varepsilon) > 0$, we have $P(X_{k_i} \in A^\varepsilon) \to 1$ as i tends to infinity. This proves (B.3). Now, by the definition of $m_L(A)$ we have

$$P(m_L(A) \le 3KL) = P\left(\sum_{i=1}^{I_{3KL}-1} \left[\frac{\tau_i}{K+1}\right] 1_{\{X_{k_i}\in A\}} + \left[\frac{3KL - k_{I_{3KL}}}{K+1}\right] 1_{\{X_{k_{I_{3KL}}}\in A\}} \ge L \right).$$

We observe that, for $\phi = \min\{j : X_{k_j} \in A\}$, the definition of maximal absorbing set yields

$$\begin{aligned}
&\sum_{i=1}^{I_{3KL}-1} \left[\frac{\tau_i}{K+1}\right] 1_{\{X_{k_i}\in A\}} + \left[\frac{3KL - k_{I_{3KL}}}{K+1}\right] 1_{\{X_{k_{I_{3KL}}}\in A\}} \\
&= \sum_{i=1}^{\phi} \left[\frac{\tau_i}{K+1}\right] 1_{\{X_{k_i}\in A\}} + \sum_{i=\phi}^{I_{3KL}-1} \left[\frac{\tau_i}{K+1}\right] 1_{\{X_{k_i}\in A\}} + \left[\frac{3KL - k_{I_{3KL}}}{K+1}\right] 1_{\{X_{k_{I_{3KL}}}\in A\}} \\
&= 1_{\{\phi < I_{3KL}-1\}} \cdot \left(\sum_{i=\phi}^{I_{3KL}-1} \left[\frac{\tau_i}{K+1}\right] + \left[\frac{3KL - k_{I_{3KL}}}{K+1}\right] \right).
\end{aligned}$$

Since $X_{k_i} \to X^*$ and $P(X_i = X^*) = 0$, we have $I_{3KL} \to \infty$ as $L \to \infty$, and therefore $1_{\{\phi < I_{3KL}-1\}} \to 1$ (all limits are with probability one). We will need the estimate

$$\begin{aligned}
&\sum_{i=\phi}^{I_{3KL}-1} \left[\frac{\tau_i}{K+1}\right] + \left[\frac{3KL - k_{I_{3KL}}}{K+1}\right] \\
&= \sum_{i=\phi}^{I_{3KL}-1} \left[\frac{\tau_i}{K+1}\right] 1_{\{\tau_i \ge K+1\}} + \left[\frac{3KL - k_{I_{3KL}}}{K+1}\right] 1_{\{(3KL - k_{I_{3KL}}) \ge K+1\}} \\
&\ge \frac{1}{2} \sum_{i=\phi}^{I_{3KL}-1} \frac{\tau_i}{K+1} 1_{\{\tau_i \ge K+1\}} + \frac{1}{2} \frac{3KL - k_{I_{3KL}}}{K+1} 1_{\{(3KL - k_{I_{3KL}}) \ge K+1\}} \\
&= \frac{1}{2} \left(\sum_{i=\phi}^{I_{3KL}-1} \frac{\tau_i}{K+1} + \frac{3KL - k_{I_{3KL}}}{K+1} \right) \\
&\quad - \frac{1}{2} \left(\sum_{i=\phi}^{I_{3KL}-1} \frac{\tau_i}{K+1} 1_{\{\tau_i < K+1\}} + \frac{3KL - k_{I_{3KL}}}{K+1} 1_{\{(3KL - k_{I_{3KL}}) < K+1\}} \right)
\end{aligned}$$

(since $\frac{\tau_i}{K+1} 1_{\{\tau_i < K+1\}} \le 1_{\{\tau_i < K+1\}}$ and likewise for $(3KL - k_{I_{3KL}})$)

$$\begin{aligned}
&\ge \frac{1}{2} \left(\sum_{i=\phi}^{I_{3KL}-1} \frac{\tau_i}{K+1} + \frac{3KL - k_{I_{3KL}}}{K+1} \right) - \frac{1}{2} \left(\sum_{i=\phi}^{I_{3KL}-1} 1_{\{\tau_i < K+1\}} + 1_{\{(3KL - k_{I_{3KL}}) < K+1\}} \right) \\
&\ge \frac{1}{2} \left(\sum_{i=\phi}^{I_{3KL}-1} \frac{\tau_i}{K+1} + \frac{3KL - k_{I_{3KL}}}{K+1} \right) - \frac{1}{2} \left(\sum_{i=\phi}^{I_{3KL}-1} 1_{\{\tau_i < K+1\}} \right) \\
&= \frac{1}{2(K+1)} \left(\sum_{i=1}^{I_{3KL}-1} \tau_i + (3KL - k_{I_{3KL}}) \right) - \frac{1}{2(K+1)} \sum_{i=1}^{\phi-1} \tau_i - \frac{1}{2} \left(\sum_{i=\phi}^{I_{3KL}-1} 1_{\{\tau_i < K+1\}} \right).
\end{aligned}$$

By the definition of $\tau_i = k_{i+1} - k_i$ we have $\sum_{i=1}^{I_{3KL}-1} \tau_i = k_{I_{3KL}} - 1$. This implies that the last quantity above is equal to

$$\frac{3KL-1}{2(K+1)} - \frac{1}{2(K+1)} \sum_{i=1}^{\phi-1} \tau_i - \frac{1}{2} \left(\sum_{i=\phi}^{I_{3KL}-1} 1_{\{\tau_i < K+1\}} \right).$$

Letting $W = \frac{1}{2(K+1)} (\sum_{i=1}^{\phi-1} \tau_i - 1)$ and summarizing, we have shown that

$$\begin{aligned} &P(m_L(A) \le 3KL) \\ &\ge P\left(\left\{ \frac{3KL}{2(K+1)} - W - \frac{1}{2} \left(\sum_{i=\phi}^{I_{3KL}-1} 1_{\{\tau_i < K+1\}} \right) \ge L \right\} \cap \{\phi < I_{3KL} - 1\} \right) \\ &\quad + P(\phi \ge I_{3KL} - 1). \end{aligned}$$

We already have observed that $P(\phi \ge I_{3KL}) \to 0$ as $L \to \infty$, so in order to show that $P(m_L(A) \le 3KL) \to 1$, it suffices to show that

$$P\left(\frac{3KL}{2(K+1)} - W - \frac{1}{2} \left(\sum_{i=\phi}^{I_{3KL}-1} 1_{\{\tau_i < K+1\}} \right) \ge L \right) \to 1$$

as $L \to \infty$. Since W does not depend on L, it is sufficient to show that

$$P\left(\left(\sum_{i=\phi}^{I_{3KL}-1} 1_{\{\tau_i < K+1\}} \right) \le L/2 \right) \to 1 \text{ as } L \to \infty.$$

This would easily follow if $\frac{1}{L} \sum_{i=\phi}^{I_{3KL}} 1_{\{\tau_i < K+1\}} \to 0$ in probability. An easy computation yields

$$\frac{1}{L} \sum_{i=\phi}^{I_{3KL}} 1_{\{\tau_i < K+1\}} \le \frac{1}{L} \sum_{i=1}^{3KL} 1_{\{\tau_i < K+1\}}$$

and

$$P\left(\frac{1}{L} \sum_{i=\phi}^{I_{3KL}} 1_{\{\tau_i < K+1\}} > \varepsilon \right) \le \frac{1}{\varepsilon L} \sum_{i=1}^{3KL} P(\tau_i < K+1),$$

which converges to 0 by (B.3). This proves Lemma 5.

Proof of Theorem 1. We prove the theorem only in the case of a unique maximum since the extension to finitely many maxima is straightforward. Without loss of generality we can assume that $0 < f(0) \le f(x)$ for all $x \in D$ and that $X^* \equiv \arg\max_{x \in D} f(x)$ is at least ρ away from the boundary of D. (This can always be accomplished by slightly enlarging the cube and by shifting the function.) We will use the notation $f^* \equiv f(X^*)$.

Since f is continuous and has a unique maximum, we can take the set A to be the maximal absorbing set for $s = 1/(3c^K)$. Let $m_L := m_L(A)$. (Here K and $m_L(A)$ are as in Lemma 5.) Let

$$\begin{aligned} \Lambda_L = &\{(i,j) \in \mathbf{N}^2 : 1 \le i \le (I_{m_L} - 1), 1 \le j \le \tau_i\} \\ &\cup \{(i,j) \in \mathbf{N}^2 : i = I_{m_L}, 1 \le j \le (n - k_{I_{M_L}})\}, \end{aligned}$$

and let $\widetilde{\Lambda}_L \subset \mathbf{N}^2$ consist of those (i,j) satisfying

$$X_{k_i} \in A \text{ , } i = 1, \ldots, (I_{m_L} - 1), \text{ and } j = l(K+1) \text{ with } l = 1, \ldots, \left[\frac{\tau_i}{K+1}\right]$$

or

$$X_{k_{I_{m_L}}} \in A \text{ , } i = I_{m_L}, \text{ and } j = l(K+1) \text{ with } l = 1, \ldots, \left[\frac{n - k_{I_{m_L}}}{K+1}\right].$$

In other words, $\widetilde{\Lambda}_L$ consists of (i,j) for which $Y_{i,j}$ are sampled from the minimal balls $B(\nu, 1/c^K)$ with center $\nu \in A$. Since $\widetilde{\Lambda}_L \subset \Lambda_L$ we have

$$\begin{aligned} f(X_{m_L}) &= \max\{f(X_1), \ldots, f(X_{m_L})\} \\ &= \max\{f(Y_{i,j}) : (i,j) \in \Lambda_L\} \\ &\geq \max\{f(Y_{i,j}) : (i,j) \in \widetilde{\Lambda}_L\}. \end{aligned} \tag{B.5}$$

By the choice of m_L, the cardinality of $\widetilde{\Lambda}_L$ is exactly L. Therefore we can reorder the $Y_{i,j}$ variables as Z_i, $i = 1, \ldots, L$, yielding

$$f(X_{m_L}) \geq \max\{f(Z_i) : i = 1, \ldots, L\}.$$

We also observe that, for each $i = 1, \ldots, L$, Z_i is uniformly distributed on $B(\eta_i, \frac{1}{c^K})$, where η_i are random variables measurable with respect to $\sigma(Z_1, Z_2, \ldots, Z_{i-1})$ such that $\eta_i \in A$ for all ω. Let η_0 be any point in A and $Z_0 \sim$ uniform on $B(\eta_0, \frac{1}{c^K})$. Since by assumption $f(0) \leq f(x)$ for all $x \in D$, we have

$$\max\{f(Z_i) : i = 1, \ldots, L\} \geq \max\{f(Z_i \cdot 1_{\{Z_i \in A\}}) : i = 1, \ldots, L\}.$$

Combining the above inequalities, we see that

$$f(X_{m_L}) \geq \max\{f(Z_i \cdot 1_{\{Z_i \in A\}}) : i = 1, \ldots, L\}. \tag{B.6}$$

Clearly the sequence $\{Z_i \cdot 1_{\{Z_i \in A\}}\}_{i=1}^L$ satisfies the conditions of Lemma 4 (by letting $y = 0$ and $Y_i = Z_i \cdot 1_{\{Z_i \in A\}}$) and is therefore i.i.d. We would like to compare the right-hand side of (B.6) with the maximum obtained by PRS. Since by assumption the argmax X^* is at least ρ away from the boundary of D, we have

$$\mu\left(B\left(X^*, \frac{1}{c^K}\right)\right) = \frac{1}{\lambda^K}, \tag{B.7}$$

where $\lambda = c^d$ (this is true only for sup-norm balls; modifications for other norms are straightforward). Let $\{\widetilde{Z}_i\}_{i=1}^\infty$ be i.i.d. uniform on D,

$$N_L = \max\{f(\widetilde{Z}_i \cdot 1_{\{\widetilde{Z}_i \in A\}}) : i = 1, \ldots, \lambda^K L\}$$

and

$$M_L = \max\{f(Z_i \cdot 1_{\{Z_i \in A\}}) : i = 1, \ldots, L\}.$$

We will show that

$$P(M_L > t) > P(N_L > t) \text{ for all } t \in (0, f^*) \text{ and all } L \in \mathbf{N}, \tag{B.8}$$

which in turn implies that $E(M_L) > E(N_L)$, since $0 < f(x) < f^*$ for all $x \neq X^*$. Let

$$A_t = \{x \in D : f(x) > t\}\ ,\ B_t = A_t \cap A.$$

For $t > f(0)$ we have

$$\begin{aligned} P(M_L > t) &= P(Z_i \cdot 1_{\{Z_i \in A\}} \in A_t \text{ for some } i \leq L) \\ &\quad \text{(since } 0 \notin A_t\text{)} \\ &= P(Z_i \cdot 1_{\{Z_i \in A\}} \in A_t \cap A \text{ for some } i \leq L) \\ &= P(Z_i \in B_t \text{ for some } i \leq L) \\ &= 1 - (P(Z_1 \in (B_t)^c))^L \\ &= 1 - (1 - P(Z_1 \in B_t))^L \\ &= 1 - (1 - PM(t))^L, \end{aligned}$$

where $PM(t) = P(Z_1 \in B_t)$. By a similar argument and with $PN(t) = P(\widetilde{Z}_1 \in B_t)$, we have

$$P(N_L > t) = 1 - (1 - PN(t))^{\lambda^K L}. \tag{B.9}$$

By the choice of λ and expression (B.7) we have

$$PN(t) = \frac{\mu(B_t)}{\mu(D)} = \frac{\mu(B_t)}{\mu(B(X^*, \frac{1}{c^K}))\lambda^K} = \frac{PM(t)}{\lambda^K}.$$

Hence, (B.9) is equal to

$$1 - \left(1 - \frac{PM(t)}{\lambda^K}\right)^{\lambda^K L}.$$

Since

$$1 - (1-a)^L > 1 - \left(1 - \frac{a}{s}\right)^{sL}$$

for all $a \in (0,1)$, $L > 0$, and $s > 1$, we obtain the inequality (B.8) by setting $a = PM(t)$ and $s = \lambda^K$. Combining (B.8) and (B.6) gives

$$E\left(f\left(X_{m_L}\right)\right) > E\left(\max_{i \leq \lambda^K L}\left\{f\left(\widetilde{Z}_i \cdot 1_{\{\widetilde{Z}_i \in A\}}\right)\right\}\right). \tag{B.10}$$

Since $f(x) > 0$ and A is a maximal absorbing set, we can eliminate $1_{\{\widetilde{Z}_i \in A\}}$ from (B.10). That is, for L large enough

$$\max_{i \leq \lambda^K L}\left\{f\left(\widetilde{Z}_i \cdot 1_{\{\widetilde{Z}_i \in A\}}\right)\right\} = \max_{i \leq \lambda^K L}\left\{f\left(\widetilde{Z}_i\right)\right\}.$$

Thus, for any $0 < \delta < 1$

$$E\left(f\left(X_{m_L}\right)\right) \geq E\left(\max_{i \leq \delta\lambda^K L}\{f(\widetilde{Z}_i)\}\right) \text{ for } L \text{ large enough.} \tag{B.11}$$

Finally, Lemma 5 allows us to replace $f(X_{m_L})$ with $f(X_{3KL})$, from which we have

$$E\left(f\left(X_{3KL}\right)\right) \geq E\left(\max_{i \leq \delta \lambda^K L}\{f(\widetilde{Z}_i)\}\right) \text{ for } L \text{ large enough.}$$

Theorem 1 now follows by letting $n = 3KL$.

Proof of Theorem 2. We first prove the following lemma. Let $\{X_n\}$ be a sequence of integrable random variables defined on a probability space $(\Omega, \mathcal{F}, P)$, let $\{\sigma_n\}$ be a sequence of subsigma fields, and let N be an integer stopping time $\{N = n\} \in \sigma_n$ for all n. Define

$$\sigma_N \equiv \sigma\{C : C \cap \{N = n\} \in \sigma_n \text{ for each } n\}.$$

LEMMA 6. *The random variables X_N and X_n satisfy*

$$E\left(X_N |\ \sigma_N\right) = E\left(X_n |\ \sigma_n\right) \text{ a.s. on } \{N = n\}. \tag{B.12}$$

Proof of Lemma 6. For any A in σ_N,

$$\begin{aligned} E[E(X_N |\ \sigma_N) \cdot 1_{\{A \cap \{N < \infty\}\}}] &= E[X_N \cdot 1_{\{A \cap \{N < \infty\}\}}] \\ &= \sum_n E[X_N \cdot 1_{\{A \cap \{N = n\}\}}] \\ &= \sum_n E[E(X_N |\ \sigma_N) \cdot 1_{\{A \cap \{N = n\}\}}] \\ &= E\left[\left(\sum_n E[E(X_N |\ \sigma_N) 1_{\{N=n\}}\right) \cdot 1_{\{A \cap \{N < n\}\}}]\right]. \end{aligned}$$

Since A is arbitrary, we have

$$E(X_N |\ \sigma_N) = \sum_n E(X_n |\ \sigma_n) \cdot 1_{\{N=n\}} \text{ a.s. on } \{N < \infty\},$$

which is equivalent to the statement of the lemma.

Moving on to a proof of Theorem 2, let $L = \sup_n M_n$. By monotonicity, $M_n \uparrow L$ as $n \to \infty$. We assume without loss of generality that $P(X_{n+1} = X_n \text{ eventually}) < 1$ and that $\{X_{n+1} \neq X_n \text{ i.o.}\} = \{X_{n+1} = X_n \text{ eventually}\}^c$ occurs. Let

$$N_1 = \min\{n : X_{n+1} \neq X_n\} \tag{B.13}$$

and, recursively,

$$N_{k+1} = \min\{n > N_k : X_{n+1} \neq X_n\}. \tag{B.14}$$

Note that N_k is not a stopping time; however, $N_k^+ = N_k + 1$ is. The random indices $N_k \uparrow \infty$ as $k \to \infty$. By compactness, there is a convergent subsequence $X_{N_{k_i}^+}$ of $X_{N_k^+}$ which converges to some X as $i \to \infty$. By monotonicity, $M_{N_{k_i}^+} \uparrow L$ as $i \to \infty$, as well.

For $n = 1, 2, \ldots$, let $\{\sigma_n\} = \sigma(X_1, \ldots, X_n)$. Let A be any (Lebesgue) measurable set and consider the event $A_n = \{Y_n \in A\}$. By definition,

$$P(A_n \mid \sigma(X_n)) = \frac{\mu(A \cap B(X_n, R))}{\mu(B(X_n, R))}, \text{ with probability one.} \tag{B.15}$$

Clearly,

$$P(A_n \mid \sigma_n) = P(A_n \mid \sigma(X_n)), \text{ with probability one.} \tag{B.16}$$

A "strong Markov" version of (B.16) also holds. Indeed, for any stopping time N an expansion along possible values of N yields

$$P(A_N \mid \sigma_N) = P(A_N \mid \sigma(X_N)), \text{ with probability one, on } \{N < \infty\}. \tag{B.17}$$

An argument analogous to the proof of Lemma 6 shows that

$$P(A_N \mid \sigma_N) = P(A_n \mid \sigma_n), \text{ with probability one, on } \{N = n\}. \tag{B.18}$$

It follows that

$$\frac{\mu(A \cap B(X_N, 1))}{\mu(B(X_N, 1))} = P(A_n \mid \sigma_n), \text{ with probability one, on } \{N = n\}. \tag{B.19}$$

Observe that for any x in D and $s > 0$

$$s^d \le \mu(B(x, s)) \le (2s)^d \wedge 1. \tag{B.20}$$

Let $0 < \delta < 1$, and let Z_δ be distributed uniformly on $B(X, 1-\delta)$ and independent of the record-generating sequence $\{X_n\}$. For any Lebesgue measurable set A, if n and I are large enough so that $N^+_{k_i} = n$ for some $i \ge I$, then by (B.19) and (B.20)

$$\begin{aligned} P(Z_\delta \in A) &= \frac{\mu(A \cap B(X, 1-\delta))}{\mu(B(X, 1-\delta))} \\ &\le \frac{2^d}{(1-\delta)^d} \sum_{i \ge I} \frac{\mu(A \cap B(X_{N^+_{k_i}}, 1))}{2^d} \cdot 1_{\{N^+_{k_i} = n\}} \\ &\le \frac{2^d}{(1-\delta)^d} \sum_i P(A_n \mid \sigma_n) \cdot 1_{\{N^+_{k_i} = n\}}, \end{aligned} \tag{B.21}$$

where again $A_n = \{Y_n \in A\}$. Take $A = \{x \in D : f(x) > L\}$; note that A is random and measurable (pointwise). Since $A_n = \emptyset$, we find that $P(A_n \mid \sigma_n) = 0$; hence, by (B.21) we must have $P(Z_\delta \in A) = \mu(A \cap B(X, 1-\delta)) = 0$. We conclude the proof by letting $\delta \downarrow 0$.

We are able to characterize the almost-sure occurrence of an infinite stream of record generators in terms of the following assumption on the topology of f under μ.

PROPERTY A. *For almost every $x \in D$,*

$$\prod_{i=1}^{M} P\left(f \le f(x) \,\middle|\, B(x, c^{-i})\right) \to 0 \text{ as } M \to \infty. \tag{B.22}$$

PROPOSITION 7. *Let $\{X_n\}$ be the record-generating sequence from infinite descent ARS applied to the objective function f. Then $P\{X_{n+1} \ne X_n \text{ i.o.}\} = 1$ if and only if f satisfies Property* A.

We know that for any sequence $\{p_i\}$ in $[0, 1]$,

$$\prod_{i=1}^{M} (1 - p_i) \to 0 \text{ if and only if } \sum_{i=1}^{M} p_i \to \infty.$$

Thus Property A is equivalent to the statement that $\sum_i P(f > f(x)|B(x, c^{-i}))$ diverges for almost every x in D. We note the similarity of Property A to assumption H2 of Solis and Wets [29].

Proof of Proposition 7. We first suppose that f satisfies Property A. Conditioning on the mth record generator in the ARS sequence, we have

$$\begin{aligned} Q_M(m) &\equiv P\left(\cap_{n=m}^{M}\{X_{n+1} = X_n\}|\ \sigma\left(X_m\right)\right) \\ &= \prod_{i=1}^{M-m} P\left(f \le f(X_m)\,\middle|B(X_m, c^{-i-J})\right) \to 0, \text{ with probability one, as } M \to \infty, \end{aligned}$$

with $J \in \{0, 1, \ldots, m\}$ randomly chosen. By the bounded convergence theorem, for each m

$$\begin{aligned} P\left(\cap_{n\ge m}\{X_{n+1} = X_n\}\right) &= E\left[P\left(\cap_{n\ge m}\{X_{n+1} = X_n\}|\ \sigma\left(X_m\right)\right)\right] \\ &= E\left[\lim_{M\to\infty} Q_M(m)\right] \\ &= 0, \end{aligned}$$

and so by Boole's inequality,

$$\begin{aligned} P\left(X_{n+1} = X_n,\ \text{eventually}\right) &= P\left(\cup_m \cap_{n\ge m}\left\{X_{n+1} = X_n\right\}\right) \\ &\le \sum\nolimits_m P\left(\cap_{n\ge m}\{X_{n+1} = X_n\}\right) \\ &= 0. \end{aligned}$$

On the other hand, assume $P(X_{n+1} = X_n \text{ eventually}) = 0$. Then $P(\cap_{n\ge m}\{X_{n+1} = X_n\}) = 0$ for each m, and so $Q_M(m) \downarrow 0$ as $M \to \infty$ in L^1, hence in probability, and therefore with probability one, by monotonicity. Assume that Property A does not hold: there is a measurable set A such that $\mu(A) > 0$ and

$$\limsup_M \prod_{i=1}^{M} P\left(f \le f(x)\,\middle|B(x, c^{-i})\right) > 0$$

for all x in A. Now for each m

$$\begin{aligned} Q_M(m) &= \prod_{i=1}^{M-m} P\left(f \le f(X_m)\,\middle|B(X_m, c^{-(i+J)})\right) \\ &= \prod_{i=J+1}^{M+J-m} P\left(f \le f(X_m)\,\middle|B(X_m, c^{-i})\right) \\ &\ge \prod_{i=1}^{M} P\left(f \le f(X_m)\,\middle|B(X_m, c^{-i})\right), \end{aligned}$$

since $0 \le J \le m$. Since $Q_M(m) \downarrow 0$, the last quantity tends to zero as M tends to infinity, so it must be that X_m is not in A. Since we have assumed an infinite sequence of new records and hence restarts, it follows that there is an infinite sequence $\{Y_n\}$ of candidates from Step 1 of the infinite descent ARS algorithm which are i.i.d. and uniformly distributed on D. Since none of these lies in A, with probability one, we must conclude that $\mu(A) = 0$, which contradicts the assumption that Property A does not hold. This proves Proposition 7.

The convergence to zero of the product in Property A (or, equivalently, divergence of the associated series) controls the rate at which the probabilities $P(f > f(x)|B(x,c^{-i}))$ can decay in order for (B.22) to hold. If we assume slightly more, namely, Property B ((2.2) above), then we obtain Theorem 3. Property B is essentially a restriction against local flatness. To see this, note first that if f is almost surely constant on a ball B, then $P(f \leq f(x)|B(x,r)) = 1$ for all $r > 0$ small enough so that $B(x,r)$ lies in B. On the other hand, if there is an $r > 0$ such that $P(f \leq f(x)|B(x,r)) = 1$ for all x in a ball B, then fix such an x. For any $y \in B(x,r) \cap B$ the conditions imposed yield $f(y) = f(x)$, with probability one; that is, f is almost surely constant on $B(x,r) \cap B$.

Proof of Theorem 3. Let $N_k^+ = N_k + 1$ and $A \in \sigma_{N_k^+}$. Then

$$\begin{aligned}
P(\Delta_{k+1} = n, A) &= \sum_{i \geq k} P(N_{k+1} = n+i, A, N_k = i) \\
&= \sum_{i \geq k} P(X_{n+i+1} \neq X_{i+1}, X_{l+i} = X_{i+1}, l = 2, \ldots, n, A, N_k = i) \\
&= \sum_{i \geq k} P(f(Y_{n+i+1}) > f(X_{i+1}) \geq f(Y_{l+i}), l = 2, \ldots, n, A, N_k^+ = i+1) \\
&= \sum_{i \geq k} E[1_{\{N_k = i\}} \cdot 1_A \cdot P(f(Y_{n+i+1}) > f(X_{i+1}) \geq f(Y_{l+i}), \\
&\qquad l = 2, \ldots, n, A |\ \sigma_{i+1})],
\end{aligned}$$

where $\{Y_{l+i}\}_{l=2,\ldots,n+1}$ are independent with Y_{l+i} distributed uniformly on $B(X_i, c^{-l})$. Here we use Lemma 6, the double expectation formula conditioning on σ_{i+1}, and the fact that N_k^+ is a stopping time. Let

$$\begin{aligned}
q_{i+1,l} &= P(f(Y_{l+i}) \leq f(X_{i+1}) |\ \sigma_{i+1}) \\
&= \frac{\mu(\{f \leq f(X_{i+1})\} \cap B(X_{i+1}, c^{-l}))}{\mu(B(X_{i+1}, c^{-l}))} \\
&= P\left(f \leq f\left(X_{i+1}\right) \middle| B\left(X_{i+1}, c^{-l}\right)\right).
\end{aligned}$$

Then

$$P\left(\Delta_{k+1} = n, A\right) = \sum_{i \geq k} E\left[1_{\{N_k = i\}} \cdot 1_A \cdot (1 - q_{i+1,n+1}) \prod_{l=2}^{n} q_{i+1,l}\right],$$

and since A is arbitrary it follows that for the ARS algorithm we have

$$P\left(\Delta_{k+1} = n \,|\, \sigma_{N_k}\right) = (1 - q_{i+1,n+1}) \prod_{l=2}^{n} q_{i+1,l} \text{ a.s. on } \{N_k = i\}, \tag{B.23}$$

and by summation

$$P\left(\Delta_{k+1} > n \,|\, \sigma_{N_k}\right) = \prod_{l=2}^{n+1} q_{i+1,l} \text{ a.s. on } \{N_k = i\}. \tag{B.24}$$

Next, consider again PRS for maxima of f. Let $\{\widetilde{X}_n\}$ be an i.i.d. sequence of random points in D with common uniform distribution and let $\widetilde{\sigma}_n = \sigma(\widetilde{X}_1, \ldots, \widetilde{X}_n)$. For a

stopping time $\widetilde{N}$ with respect to the filtration $\{\widetilde{\sigma}_n\}$, define $\widetilde{\sigma}_{\widetilde{N}} = \{C : C \cap \{\widetilde{N} = n\} \in \widetilde{\sigma}_n \text{ for all } n\}$. Let

$$\widetilde{N}_1 = \min\{n : f(X_{n+1}) > M_n\}$$

and, recursively,

$$\widetilde{N}_{k+1} = \min\left\{n > \widetilde{N}_k : f(X_{n+1}) > M_n\right\}.$$

Again, $\widetilde{N}_k$ is not a stopping time but $\widetilde{N}_k^+ = \widetilde{N}_k + 1$ is. Let $\widetilde{\Delta}_{k+1} = \widetilde{N}_{k+1} - \widetilde{N}_k$ and

$$\begin{aligned}\widetilde{q}_{i+1} &= P\left(f\left(\widetilde{X}_{i+2}\right) \le M_{i+1} \mid \widetilde{\sigma}_{i+1}\right)\\ &= \mu(f \le M_{i+1}).\end{aligned}$$

Since the $\widetilde{X}$-process is i.i.d., it is not surprising that $\widetilde{\Delta}_{k+1}$ follows a geometric distribution conditional on $\widetilde{N}_k$. Let $A \in \widetilde{\sigma}_{\widetilde{N}_k^+}$. Then

$$\begin{aligned}P\left(\widetilde{\Delta}_{k+1} = n, A\right) &= \sum_{i \ge k} P\left(\widetilde{N}_{k+1} = n + i, A, \widetilde{N}_k = i\right)\\ &= \sum_{i \ge k} P\left(f\left(\widetilde{X}_{n+i+1}\right) > M_{n+i} = M_{i+1} \ge f\left(\widetilde{X}_l\right),\right.\\ &\qquad \left. l = i+2, \ldots, n+i, A, \widetilde{N}_k^+ = i+1\right)\\ &= \sum_{i \ge k} E\left[1_{\{\widetilde{N}_k = i\}} \cdot 1_A \cdot (1 - \widetilde{q}_{i+1})\widetilde{q}_{i+1}^{n-1}\right],\end{aligned}$$

where we condition on σ_{i+1} and use the double expectation formula, Lemma 6, and the i.i.d. assumption. Since A is arbitrary, we have

$$P\left(\widetilde{\Delta}_{k+1} = n \mid \widetilde{\sigma}_{\widetilde{N}_k}\right) = (1 - \widetilde{q}_{i+1})\widetilde{q}_{i+1}^{n-1} \text{ a.s. on } \left\{\widetilde{N}_k = i\right\}$$

and, of course,

$$P\left(\widetilde{\Delta}_{k+1} > n \mid \widetilde{\sigma}_{\widetilde{N}_k}\right) = \widetilde{q}_{i+1}^n \text{ a.s. on } \left\{\widetilde{N}_k = i\right\} \tag{B.25}$$

by summation. The record sequence $\{\widetilde{M}_i\}$ of PRS converges almost surely to ess sup f; thus, for given $a < 1$, we have $\widetilde{q}_i > a$ for all i large enough. From Property B, there is a $b < 1$ such that $q_{i,l} \le b$ for all i and l. To finish the proof we take $a > b$, set $\eta = b/a$, compare (B.24) with (B.25), and take expected values.

Acknowledgments. The authors would like to thank two anonymous referees and Prof. J. E. Dennis for their insight and patience in helping them refine their ideas and presentation. The authors would also like to thank Prof. J. Hartigan for pointing out that ARS bears some resemblance to simulated annealing. Most of this work was done while the authors worked together in the Informistics group at United Technologies Research Center in East Hartford, CT.

REFERENCES

[1] M. J. Appel and D. Radulovic, *Accelerated random search*, in Proceedings of the 16th IMACS World Congress 2000 on Scientific Computing, Lausanne, Switzerland, 2000, CD-ROM.

[2] N. Baba, T. Shoman, and Y. Sawaragi, *A modified convergence theorem for a random optimization algorithm*, Inform., 13 (1977), pp. 159–166.

[3] H. C. P. Barbee, C. G. E. Boender, A. H. G. Rinnooy Kan, C. L. Scheffer, R. L. Smith, and J. Telgan, *Hit-and-run algorithms for the identification of nonredundant linear inequalities*, Math. Programming, 37 (1987), pp. 184–207.

[4] R. Brunelli and G. P. Tecchiolli, *Stochastic minimization with adaptive memory*, J. Comput. Appl. Math., 57 (1995), pp. 329–343.

[5] D. W. Bulger and G. R. Wood, *Hesitant adaptive search for global optimization*, Math. Programming, 81 (1998), pp. 89–102.

[6] J. E. Dennis, D. M. Gay, and R. E. Welsch, *Algorithm* 573. Nl2SOL—*An Adaptive Nonlinear Least-Squares Algorithm*, ACM Trans. Math. Softw., 7 (1981), pp. 369–383.

[7] R. Durrett, *Probability: Theory and Examples*, 2nd ed., ITP/Duxbury, Pacific Grove, CA, 1996.

[8] M. Gaviano, *Some general results on the convergence of random search algorithms in minimization problems*, in Towards Global Optimisation, L. Dixon and G. Szego, eds., North-Holland, Amsterdam, 1975, pp. 149–157.

[9] E. M. T. Hendrix and O. Klepper, *On uniform covering, adaptive random search and raspberries*, J. Global Optim., 18 (2000), pp. 143–163.

[10] D. E. Kaufman and R. L. Smith, *Directional choice for accelerated convergence in hit-and-run sampling*, Oper. Res., 46 (1998), pp. 84–95.

[11] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi, *Optimization by simulated annealing*, Science, 220 (1983), pp. 671–680.

[12] J. P. Lawrence III and K. Steiglitz, *Randomized pattern search*, IEEE Trans. Comput., C-21 (1972), pp. 382–385.

[13] *Mathematica* 3.0, Copyright 1988–1996 Wolfram Research, Inc.

[14] J. Matyas, *Random optimization*, Autom. Remote Control, 26 (1965), pp. 246–253.

[15] N. Metropolis, A. Rosenbluth, M. Rosenbluth, A. Teller, and E. Teller, *Equations of state calculations by fast computing machines*, J. Chem. Phys., 21 (1953), pp. 1087–1091.

[16] *Netlib Repository*, http://www.netlib.org/toms/573.

[17] N. R. Patel, R. Smith, and Z. B. Zabinsky, *Pure adaptive search in Monte Carlo optimization*, Math. Programming, 43 (1988), pp. 317–328.

[18] W. L. Price, *A controlled random search procedure for global optimization*, Comput. J., 20 (1979), pp. 367–370.

[19] L. A. Rastrigin, *Systems of Extremal Control*, Nauka, Moscow, 1974 (in Russian).

[20] G. V. Reklaitis, A. Ravindran, and K. M. Ragsdell, *Engineering Optimization Methods*, John Wiley, New York, 1983.

[21] H. E. Romeijn and R. L. Smith, *Simulated annealing and adaptive search in global optimization*, J. Global Optim., 5 (1994), pp. 101–126.

[22] H. E. Romeijn, Z. B. Zabinsky, D. L. Graesser, and S. Neogi, *New reflection generator for simulated annealing in mixed integer/continuous global optimization*, J. Optim. Theory Appl., 101 (1999), pp. 403–427.

[23] H. H. Rosenbrock, *An automatic method for finding the greatest or least value of a function*, Comput. J., 3 (1960/1961), pp. 175–184.

[24] F. Freudenstein and B. Roth, *Numerical solutions of systems of nonlinear equations*, J. ACM, 10 (1963), pp. 550–556.

[25] R. I. Jennrich and P. F. Sampson, *Application of stepwise regression to nonlinear estimation*, Technometrics, 10 (1968), pp. 63–72.

[26] G. Schrack and M. Choit, *Optimized relative step size random searches*, Math. Programming, 10 (1968), pp. 230–244.

[27] M. A. Schumer and K. Steiglitz, *Adaptive step size random search*, IEEE Trans. Automat. Control, AC-13 (1968), pp. 270–276.

[28] A. Simoes and E. Costa, *Using genetic algorithms with sexual and asexual transposition: A comparative study*, in Proceedings of the IEEE Congress on Evolutionary Computation, IEEE Press, San Diego, CA, 2000.

[29] F. J. Solis and R. Wets, *Minimization by random search techniques*, Math. Oper. Res., 6 (1981), pp. 19–30.

[30] R. L. Smith, *Efficient Monte Carlo procedures for generating points uniformly distributed over bounded regions*, Oper. Res., 32 (1984), pp. 1296–1308.

[31] Z. B. Tang, *Optimal sequential sampling policy of partitioned random search and its approximation*, J. Optim. Theory Appl., 98 (1998), pp. 431–448.
[32] Z. B. Zabinsky and R. L. Smith, *Pure adaptive search in global optimization*, Math. Programming, 53 (1992), pp. 323–338.
[33] Z. B. Zabinsky, R. L. Smith, J. F. McDonald, H. E. Romeijn, and D. E. Kaufman, *Improving hit-and-run for global optimization*, J. Global Optim., 3 (1993), pp. 171–192.
[34] Z. B. Zabinsky, G. R. Wood, M. A. Steel, and W. P. Baritompa, *Pure adaptive search for finite global optimization*, Math. Programming, 69 (1995), pp. 443–448.

SIAM J. OPTIM.
Vol. 14, No. 3, pp. 732–742

MINIMIZATION OF ERROR FUNCTIONALS OVER VARIABLE-BASIS FUNCTIONS*

PAUL C. KAINEN†, VĚRA KŮRKOVÁ‡, AND MARCELLO SANGUINETI§

Abstract. Generalized Tikhonov well-posedness is investigated for the problem of minimization of error functionals over admissible sets formed by variable-basis functions, i.e., linear combinations of a fixed number of elements chosen from a given basis without a prespecified ordering. For variable-basis functions of increasing complexity, rates of decrease of infima of error functionals are estimated. Upper bounds are derived on such rates which do not exhibit the curse of dimensionality with respect to the number of variables of admissible functions. Consequences are considered for Boolean functions and decision trees.

Key words. error functionals, approximate optimization, generalized Tikhonov well-posedness, rate of decrease of infima, complexity of admissible functions

AMS subject classifications. 49K40, 41A46, 41A25

DOI. 10.1137/S1052623402401233

1. Introduction.

Functionals defined as distances from (target) sets are called *error functionals*. Minimization of such functionals occurs in optimization tasks arising in various areas such as decision processes, system identification, machine learning, and pattern recognition.

In various applications, admissible solutions over which error functionals are minimized are functions depending on a large number of variables: for example, when routing strategies have to be devised for large-scale communication and transportation networks, when an optimal closed-loop control law has to be determined for a dynamical system with high-dimensional output measurement vector and a large number of decision stages, etc. In the last decades, complex optimization problems of this kind have been approximately solved by searching suboptimal solutions over admissible sets of functions computable by neural networks [4], [21], [22], [25], [28], [29]. Neural networks can be studied in a more general context of variable-basis functions, which also include other nonlinear families of functions such as free-node splines and trigonometric polynomials with free frequencies [17]. Families of variable-basis functions are formed by linear combinations of a fixed number of elements chosen from a given basis without a prespecified ordering [16], [17].

When admissible functions depend on a large number of variables, implementation of some procedures of approximate optimization may be infeasible due to the "curse of dimensionality" [3]. For example, when optimization is performed over linear combinations of fixed-basis functions, the number of basis functions required to

*Received by the editors January 21, 2002; accepted for publication (in revised form) February 7, 2003; published electronically December 19, 2003. Collaboration between the first and second authors was partially supported by an NAS COBASE grant, and that between the second and third authors by Scientific Agreement Italy-Czech Republic, Area MC 6, Project 22.

http://www.siam.org/journals/siopt/14-3/40123.html

†Department of Mathematics, Georgetown University, Washington, D.C. 20057-1233 (kainen@georgetown.edu). This author was partially supported by a travel grant from Georgetown University.

‡Institute of Computer Science, Academy of Sciences of the Czech Republic, Pod Vodárenskou věží 2, 182 07, Prague 8, Czech Republic (vera@cs.cas.cz). This author was partially supported by GA ČR grants 201/00/1482 and 201/02/0428.

§Department of Communications, Computer, and System Sciences (DIST), University of Genoa, Via Opera Pia 13, 16145 Genoa, Italy (marcello@dist.unige.it). This author was partially supported by the CNR-Agenzia 2000 project and by the Italian Ministry of Education, University and Research.

guarantee a desired optimization accuracy may grow exponentially fast with the number of variables of admissible solutions [23, pp. 232–233], [29]. However, experience has shown that neural networks with a small number of computational units may perform well in optimization tasks where admissible solutions depend on a large number of variables [21], [22], [25], [28], [29].

In this paper, we investigate generalized Tikhonov well-posedness of the problems of minimization of error functionals over admissible sets formed by variable-basis functions, and we estimate rates of decrease of infima of such problems with increasing complexity of admissible sets. For such an investigation, we derive various conditions on target and admissible sets guaranteeing convergence of minimizing sequences. We show that these conditions are satisfied by target sets defined by suitable interpolation and smoothness conditions and admissible sets formed by functions computable by families of variable-basis functions that include commonly used classes of neural networks. Rates of decrease are estimated for infima of error functionals over neural networks with increasing number of computational units. We derive upper bounds on such rates.

The paper is organized as follows. In section 2, we introduce basic concepts and definitions used throughout the paper. Section 3 states conditions on sets of target functions and admissible solutions that guarantee convergence of minimizing sequences. Section 4 applies the tools developed in section 3 to minimization of error functionals over neural networks and variable-basis functions, and section 5 gives estimates of rates of decrease of infima of such functionals with increasing number of computational units.

2. Preliminaries. In this paper, by a normed linear space $(X, \|.\|)$ we mean a real normed linear space. We write only X when it is clear which norm is used. For a positive integer d, a set $\Omega \subseteq \Re^d$, where $\Re$ denotes the set of real numbers, and $p \in [1, \infty)$, by $(L_p(\Omega), \|.\|_p)$ is denoted the space of measurable, real-valued functions on Ω such that $\int_\Omega |f(x)|^p \, dx < \infty$ endowed with the L_p-norm. $(\mathcal{C}(\Omega), \|.\|_\mathcal{C})$ denotes the space of real-valued continuous functions on Ω with the supremum norm.

By $\mathcal{B}(\{0,1\}^d)$ is denoted the space of real-valued Boolean functions, i.e., functions from $\{0,1\}^d$ to $\Re$. This space is endowed with the standard inner product defined for $f, g \in \mathcal{B}(\{0,1\}^d)$ as $f \cdot g = \sum_{x \in \{0,1\}^d} f(x)g(x)$, which induces the l_2-norm $\|f\|_{l_2} = \sqrt{f \cdot f}$. The space $(\mathcal{B}(\{0,1\}^d), \|.\|_{l_2})$ is isomorphic to the 2^d-dimensional Euclidean space $\Re^{2^d}$ with the l_2-norm.

For $M \subseteq (X, \|.\|)$, $\mathrm{cl}(M)$ denotes the closure of M in the topology induced by the norm $\|.\|$. For $f \in X$, we write $\|f - M\| = \inf_{g \in M} \|f - g\|$. A ball of radius r centered at $h \in (X, \|.\|)$ is denoted by $B_r(h, \|.\|) = \{f \in X : \|f - h\| \leq r\}$. We write $B_r(\|.\|)$ for $B_r(0, \|.\|)$ and merely B_r when it is clear which norm is used.

For brevity, sequences are denoted by $\{h_i\}$ instead of $\{h_i : i \in \mathcal{N}_+\}$, where $\mathcal{N}_+$ is the set of positive integers. When there is no ambiguity, the same notation is used for a sequence and its subsequences. A sequence converges subsequentially if it has a convergent subsequence.

Following [8], we denote by (M, Φ) the problem of infimizing a functional $\Phi : M \to \Re$ over $M \subseteq X$. M is called the set of *admissible solutions* or the *admissible set*. A sequence $\{g_i\}$ of elements of M is called Φ-*minimizing over* M if $\lim_{i \to \infty} \Phi(g_i) = \inf_{g \in M} \Phi(g)$. The set of argminima of the problem (M, Φ) is denoted by $\mathrm{argmin}(M, \Phi) = \{h \in M : \Phi(h) = \inf_{g \in M} \Phi(g)\}$. The problem (M, Φ) is *Tikhonov well-posed in the generalized sense* [8, p. 24] if $\mathrm{argmin}(M, \Phi)$ is not empty and each Φ-minimizing sequence over M converges subsequentially to an element of

$\operatorname{argmin}(M, \Phi)$.

For C a nonempty subset of X, the *error functional* measuring the distance from C is denoted by e_C and defined for any $h \in X$, as $e_C(h) = \|h - C\|$. We call C the *target set* or the set of *target functions*. By the triangle inequality, $e_C = e_{cl(C)}$. For a singleton $C = \{h\} \subset X$, we write e_h instead of $e_{\{h\}}$.

For error functionals, the definition of generalized Tikhonov well-posedness can be restated as follows.

PROPOSITION 2.1. *Let M and C be nonempty subsets of a normed linear space $(X, \|.\|)$. Then (M, e_C) is Tikhonov well-posed in the generalized sense if and only if every sequence in M that minimizes e_C converges subsequentially to an element of M.*

Proof. Let $\{g_i\}$ be a subsequence of an e_C-minimizing sequence converging to $g^o \in M$. By continuity of e_C [26, p. 391], $\inf_{g \in M} e_C(g) = \lim_{i \to \infty} e_C(g_i) = e_C(\lim_{i \to \infty} g_i) = e_C(g^0)$. Thus, $g^o \in \operatorname{argmin}(M, e_C)$ and so (M, e_C) is Tikhonov well-posed in the generalized sense. The "only if" statement follows directly from the definition of generalized Tikhonov well-posedness. ∎

Recall that a nonempty subset M of a normed linear space is *compact* if every sequence has a convergent subsequence, is *precompact* if $\operatorname{cl}(M)$ is compact, and is *boundedly compact* if its intersection with any ball is precompact (equivalently, every bounded sequence in M is subsequentially convergent). Note that this definition of boundedly compact set does not require M to be closed. M is *approximatively compact* [26, pp. 368, 382] if, for all $h \in X$, every sequence in M that minimizes the distance to h converges subsequentially to an element of M.

By Proposition 2.1, the notion of an approximatively compact set can be reformulated in terms of optimization theory as a set M such that, for every $h \in X$, the problem (M, e_h) is Tikhonov well-posed in the generalized sense. A subset M of a normed linear space X is *proximinal* (or an *existence set*) if for any $h \in X$ there exists $g \in M$ such that $\|h - M\| = \|h - g\|$. In decreasing degree of strength, a subset of a normed linear space may be compact, boundedly compact, approximatively compact, and proximinal. Each implies the next, with the exception that bounded compactness implies approximative compactness only for closed sets; proximinal implies closed [26, pp. 368, 382–383].

3. Minimization of error functionals under weakened compactness. Generalized Tikhonov well-posedness can be interpreted as a type of weakened compactness of admissible sets. The following theorem shows that for error functionals it is closely related to the concept of approximative compactness.

THEOREM 3.1. *Let M and C be nonempty subsets of a normed linear space $(X, \|.\|)$. Each of the following conditions guarantees that (M, e_C) is Tikhonov well-posed in the generalized sense:*

(i) *M is approximatively compact and C is precompact;*
(ii) *M is approximatively compact and bounded and C is boundedly compact;*
(iii) *M is boundedly compact and closed and C is bounded.*

Proof. Let $\{g_i\}$ be an e_C-minimizing sequence over M. By Proposition 2.1, it is sufficient to show that $\{g_i\}$ converges subsequentially to $g^o \in M$.

(i) Since $e_C = e_{cl(C)}$, it is sufficient to consider $\operatorname{cl}(C)$. As $\operatorname{cl}(C)$ is compact, it is proximinal and so there exists a sequence $\{f_i\} \subseteq \operatorname{cl}(C)$ such that for every i, $e_C(g_i) = e_{cl(C)}(g_i) = \|f_i - g_i\|$. Again by compactness, the sequence $\{f_i\}$ converges subsequentially to $f^o \in \operatorname{cl}(C)$. Replacing $\{f_i\}$ and $\{g_i\}$ with the corresponding subsequences, for every $\varepsilon > 0$ we get $i_0 \in \mathcal{N}_+$ such that for all $i \geq i_0$, $\|f_i - f^o\| < \varepsilon/2$.

As $\{g_i\}$ is e_C-minimizing over M, there exists $i_1 \geq i_0$ such that for all $i \geq i_1$, $e_C(g_i) \leq \inf_{g\in M} e_C(g) + \varepsilon/2$. So, for all $i \geq i_1$, $e_{f^o}(g_i) \leq \|g_i - f_i\| + \|f_i - f^o\| = e_C(g_i) + \|f_i - f^o\| < \inf_{g\in M} e_C(g) + \varepsilon \leq \inf_{g\in M} e_{f^o}(g) + \varepsilon$. Hence, $\{g_i\}$ is an e_{f^o}-minimizing sequence over M. By approximative compactness of M, there exists $g^o \in M$ such that $\{g_i\}$ converges subsequentially to g^o.

(ii) As $\mathrm{cl}(C)$ is boundedly compact and closed, it is proximinal and so there exists a sequence $\{f_i\} \subseteq \mathrm{cl}(C)$ such that for every i, $e_C(g_i) = \|f_i - g_i\|$. By the triangle inequality, $\|f_i\| \leq \|f_i - g_i\| + \|g_i\|$. Both sequences, $\{\|g_i\|\}$ and $\{\|f_i - g_i\|\}$, are bounded: the first one by boundedness of M and the second one as $\{\|f_i - g_i\|\}$ is convergent (since $\lim_{i\to\infty} \|g_i - f_i\| = \lim_{i\to\infty} e_C(g_i) = \inf_{g\in M} e_C(g)$). By closedness and bounded compactness of $\mathrm{cl}(C)$, there exists $f^o \in \mathrm{cl}(C)$, to which $\{f_i\}$ converges subsequentially, and so we can proceed as in the second part of the proof of (i).

(iii) As C is bounded, there exists $r > 0$ such that $C \subseteq B_r$. Let $a = \inf\{\|f - g\| : f \in C, g \in M\}$. Then there exist $i_0 \in \mathcal{N}_+$ and $b > 0$ such that for all $i \geq i_0$, $e_C(g_i) < a + b$ and so there exist $i_1 \geq i_0$, $f_i \in C$, and $b' \geq b$ such that for all $i \geq i_1$, $\|g_i - f_i\| < a + b'$. By the triangle inequality, $\|g_i\| \leq \|g_i - f_i\| + \|f_i\| < a + b' + r$. Thus for all $i \geq i_1$, $\{g_i\} \subseteq B_{a+b'+r} \cap M$ and so $\{g_i\}$ has a bounded subsequence. As M is boundedly compact and closed, this subsequence converges subsequentially to $g^o \in M$. □

Table 3.1 summarizes conditions on M and C assumed in Theorem 3.1 which guarantee that (M, e_C) is Tikhonov well-posed in the generalized sense.

TABLE 3.1

Conditions on M and C guaranteeing Tikhonov well-posedness in the generalized sense of (M, e_C). Y = yes, N = no (by "no" we mean "there exists a counterexample").

	C precompact	C boundedly compact	C bounded
M approximatively compact	Y	N	N
M boundedly compact and closed	Y	N	Y
M approximatively compact and bounded	Y	Y	N

The first entry in the first column holds by Theorem 3.1(i), while the other two entries in the same column hold since there the conditions on M are stronger than those required in the first entry. In the second column, Theorem 3.1(ii) justifies the "yes" entry, while "yes" in the third column holds by Theorem 3.1(iii).

Both "no" entries in the second column are shown by the following counterexample. In the Euclidean space $\Re^2$, let C be the x-axis and M the graph of the exponential function. Then M and C are boundedly compact and closed and hence approximatively compact. But no e_C-minimizing sequence in M has a convergent subsequence.

The "no" entries in the third column are demonstrated by the following example. Let $(l_2, \|.\|_{l_2})$ be the Hilbert space of square-summable sequences and let $\{e_i\}$ be its orthonormal basis. Let L denote the orthogonal complement of a unit vector (say, e_1) and let $M = L \cap B_1(\|.\|_{l_2})$. As every closed convex subset of a uniformly convex Banach space is approximatively compact [5, p. 25], M is a bounded approximatively compact set. Let $C = w\, e_1 + M$, where w is any nonzero real number. Then C is closed and bounded. The sequence $\{e_2, e_3, \ldots\}$ in M satisfies, for all $j \geq 2$, $\|e_j - C\|_{l_2} = |w|$, and so it is e_C-minimizing over M but has no convergent subsequence.

Theorem 3.1 will be used in the next section to investigate generalized Tikhonov well-posedness of (M, e_C) for admissible sets M computable by variable-basis functions and, as a particular case, by neural networks.

4. Convergence of minimizing sequences formed by variable-basis functions. In this section, X is a linear space of real-valued functions on a subset of $\Re^d$. Let G be a subset of X. The families $span_n\, G = \{\sum_{i=1}^n w_i g_i :\ w_i \in \Re,\ g_i \in G\}$ and $conv_n\, G = \{\sum_{i=1}^n w_i g_i :\ w_i \in [0,1],\ \sum_{i=1}^n w_i = 1,\ g_i \in G\}$ are called *variable-basis functions* [16], [17]. Sets $span_n\, G$ model situations in which admissible functions are represented as linear combinations of any n-tuple of functions from G, with unconstrained coefficients in the linear combinations. In many applications such coefficients are constrained by a bound on a norm of the coefficients vector $(w_1, \ldots, w_n)$. When such a norm is the l_1-norm, the corresponding functions belong to the set $\{\sum_{i=1}^n w_i g_i :\ w_i \in \Re,\ g_i \in G, \sum_{i=1}^n |w_i| \le c\}$, where $c > 0$ is a given bound on the l_1-norm. It is easy to see that this set is contained in $conv_n G'$, where $G' = \{rg : |r| \le c\,,\ g \in G\}$. As any two norms on $\Re^n$ are equivalent, every norm-based constraint on the coefficients of linear combinations defines a set contained in a set of the form $conv_n\, G'$.

Depending on the choice of X and G, one can obtain a variety of admissible sets that include functions computable by neural networks, splines with free nodes, trigonometric polynomials with free frequencies, etc. For simplicity, we shall consider functions defined on $[0,1]^d$. Let $A \subseteq \Re^q$, $\phi : A \times [0,1]^d \to \Re$ be a function of two vector variables, and $G_\phi = \{\phi(a, \cdot) : a \in A\}$. By suitable choices of A and ϕ, one can represent by G_ϕ sets of functions computable by various types of so-called *neural networks* with computational unit ϕ. If $A = S^{d-1} \times \Re$, where $S^{d-1} = \{e \in \Re^d : \|e\| = 1\}$ is the set of unit vectors in $\Re^d$, and $\phi((e,b),x) = \vartheta(e \cdot x + b)$, where ϑ denotes the Heaviside function, defined as $\vartheta(t) = 0$ for $t < 0$ and $\vartheta(t) = 1$ for $t \ge 0$, then we shall denote such a set G_ϕ by H_d, as it is the set of characteristic functions of closed half-spaces of $\Re^d$ restricted to $[0,1]^d$. Functions in H_d are called *Heaviside perceptrons*; functions in $span_n\, H_d$ and $conv_n\, H_d$ are called *Heaviside perceptron networks*.

If $A = [-c,c]^d \times [-c,c]$ and $\phi((v,b),x) = \psi(v \cdot x + b)$, where $\psi : \Re \to \Re$ is called *activation function*, b is called *bias*, and the components of v are called *weights*, then G_ϕ, denoted by $P_d(\psi, c)$, is the set of functions on $[0,1]^d$ computable by *ψ-perceptrons* with both biases and weights bounded by c. $P_d(\psi)$ denotes the corresponding set with no bounds on the parameters values. Functions in $span_n\, P_d(\psi, c)$, $conv_n\, P_d(\psi, c)$, $span_n\, P_d(\psi)$, and $conv_n\, P_d(\psi)$ are called *ψ-perceptron networks*. The most common activation functions in perceptrons are *sigmoidals*, i.e., bounded measurable functions $\sigma : \Re \to \Re$ with $\lim_{t\to-\infty} \sigma(t) = 0$ and $\lim_{t\to+\infty} \sigma(t) = 1$ such as the logistic sigmoid $\sigma(t) = 1/(1 + \exp(-t))$ and the Heaviside function. If the activation function ψ is positive and even, $A = [-c,c]^d \times [-c,c]$, and $\phi((v,b),x) = \psi(b\|x - v\|)$, where $\|.\|$ is a norm on $\Re^d$, b is called *width*, and v is called *centroid*, then G_ϕ, denoted by $F_d(\psi, c)$, is the set of functions on $[0,1]^d$ computable by *ψ-radial-basis-functions* (*ψ-RBF*) with both widths and centroids bounded by c (a typical activation function for RBF units is the Gaussian function $\psi(t) = e^{-t^2}$). $F_d(\psi)$ denotes the corresponding set with no bounds on the parameters values. Functions in $span_n\, F_d(\psi, c)$, $conv_n\, F_d(\psi, c)$, $span_n\, F_d(\psi)$, and $conv_n\, F_d(\psi)$ are called *ψ-RBF networks*. The number n of computational units in ψ-perceptron networks and ψ-RBF networks can be considered as a measure of the network "complexity," as the number of network parameters depends on n linearly.

The following proposition applies Theorem 3.1 to admissible sets computable by

neural networks. We use the notation of the preceding three paragraphs.

PROPOSITION 4.1. *Let $(X, \|.\|)$ be a normed linear space and C, M be nonempty subsets. The problem (M, e_C) is Tikhonov well-posed in the generalized sense if any of the following conditions hold:*

(i) *C is bounded and $M = conv_n G_\phi$ or $M = span_n G_\phi$, where n is a positive integer and G_ϕ is finite-dimensional;*

(ii) *$(X, \|.\|) = (\mathcal{C}([0,1]^d), \|.\|_\mathcal{C})$, C is bounded, and $M = conv_n P_d(\psi, c)$ or $M = conv_n F_d(\psi, c)$, where $c > 0$, ψ is bounded and continuous, and d, n are positive integers;*

(iii) *$(X, \|.\|) = (L_p([0,1]^d), \|.\|_p)$, $p \in [1, \infty)$, C is precompact, and $M = span_n H_d$, or else C is bounded, $M = conv_n H_d$, and d, n are positive integers.*

Proof. (i) If G_ϕ is finite-dimensional (e.g., if the set A of parameters of ϕ is finite), then it is straightforward that $span_n G_\phi$ is boundedly compact and closed. So we conclude by Theorem 3.1(iii).

(ii) By Theorem 3.1(iii), it is sufficient to check that in all these cases M is boundedly compact and closed. For $G = P_d(\psi, c)$ and $G = F_d(\psi, c)$ with $c > 0$ and ψ bounded and continuous, compactness of $conv_n G$ in $(\mathcal{C}([0,1]^d), \|.\|_\mathcal{C})$ has been proved in [12].

(iii) If C is precompact and $M = span_n H_d$, then by Theorem 3.1(i) it is sufficient to check that $span_n H_d$ is approximatively compact. Approximative compactness of $span_n H_d$ in $(L_p([0,1]^d), \|.\|_p)$, $p \in [1, \infty)$, was shown in [11]. If C is bounded and $M = conv_n H_d$, then by Theorem 3.1(iii) it is sufficient to prove that $conv_n H_d$ is boundedly compact and closed. Compactness of $G = H_d$ in $(L_2([0,1]^d), \|.\|_2)$ was proved in [9] and inspection of the argument shows that it also holds for L_p-spaces with $p \in [1, \infty)$. Since the convex hull of a compact set G is compact and $conv_n G$ is closed in $conv\, G$, compactness of $conv_n H_d$ follows from compactness of H_d. □

Note that for neural networks with differentiable activation functions (e.g., perceptrons with logistic sigmoid or RBF with the Gaussian activation function) the sets $span_n G_\phi$ are not approximatively compact in $(\mathcal{C}([0,1]^d), \|.\|_\mathcal{C})$ or in $(L_p([0,1]^d), \|.\|_p)$, because they are not even closed. (It was shown in [20] for perceptron networks, and the arguments used there can be extended to Gaussian RBF networks.)

Proposition 4.1 can be combined with various conditions guaranteeing precompactness of the target set C, such as interpolation and smoothness conditions, which model neural network learning from data described by input/output pairs and constraints given by physical considerations or feasibility of implementation.

PROPOSITION 4.2. *Let d, n be positive integers and let C be a nonempty set of continuous functions defined on $[0,1]^d$ satisfying the following two conditions:*

(1) *there exists $a > 0$ such that on $(0,1)^d$ all first-order partial derivatives of all elements of C are continuous and bounded by a in absolute value;*

(2) *there exist $x_0 \in (0,1)^d$ and $b > 0$ such that for all $f \in C$, $|f(x_0)| \le b$.*

Then for every $c > 0$ and ψ bounded and continuous, $(conv_n P_d(\psi, c), e_C)$ and $(conv_n F_d(\psi, c), e_C)$ are Tikhonov well-posed in the generalized sense in $(\mathcal{C}([0,1]^d), \|.\|_\mathcal{C})$ and $(conv_n H_d, e_C)$ and $(span_n H_d, e_C)$ are Tikhonov well-posed in the generalized sense in $(L_p([0,1]^d), \|.\|_p)$, $p \in [1, \infty)$.

Proof. Since precompactness in the space $(\mathcal{C}([0,1]^d), \|.\|_\mathcal{C})$ implies precompactness in $(L_p([0,1]^d), \|.\|_p)$, $p \in [1, \infty)$, by Proposition 4.1(ii) and (iii) it is sufficient to check that C satisfies the assumptions of the Ascoli–Arzelà theorem [1, Theorem 1.30], i.e., that the elements of C are equibounded and equicontinuous on $(0,1)^d$. Equicontinuity follows from the mean value theorem [6, p. 79] and the Cauchy–Schwarz inequality,

which together imply that for all $f \in C$, all $x \in (0,1)^d$, and all h such that for every $t \in [0,1]$, $x+th \in (0,1)^d$, there exists $\tau \in (0,1)$ such that $|f(x+th)-f(x)| = |\nabla f(x+\tau h)\cdot h| \le \|\nabla f(x+\tau h)\|\,\|h\| \le a\sqrt{d}\,\|h\|$. By applying the inequality just derived, for every $f \in C$ and every $x \in (0,1)^d$ we have $|f(x)-f(x_0)| \le a\sqrt{d}\,\|x-x_0\| \le a\,d$. Hence, $f(x) \in [-b-a\,d, b+a\,d]$ and so functions in C are equibounded on $(0,1)^d$. Thus C is precompact in $(\mathcal{C}([0,1]^d), \|.\|_{\mathcal{C}})$ and the statements follow from Proposition 4.1(ii) and (iii). □

Precompactness in $(L_p([0,1]^d), \|.\|_p)$ can also be derived using L_p versions of the Ascoli–Arzelà theorem (see, e.g., [1, Theorem 2.21]). The conditions of smoothness and interpolation required by Proposition 4.2 may be incompatible, i.e., C could be empty. In this case, one must either increase the size of the intervals Y_j or increase the bound on the derivatives.

5. Rates of decrease of infima with increasing complexity of admissible sets of variable-basis functions. In applications, the rate of decrease of infima of an error functional over $conv_n\, G$ and $span_n\, G$ should be fast enough to achieve a reasonable accuracy even for small values of n. We shall derive estimates of such rates using a result from approximation theory by Maurey [24], Jones [10], and Barron [2].

Here we reformulate these estimates in terms of a norm tailored to a given subset G of a normed linear space $(X, \|.\|)$. Such a norm, called *G-variation* and denoted by $\|.\|_G$ (although it also depends on the normed linear space), was introduced in [13] as the Minkowski functional of the set $\mathrm{cl}\, conv\,(G \cup -G)$, where closure is with respect to $\|.\|$, i.e.,

$$\|f\|_G = \inf\left\{c > 0 \,:\, c^{-1} f \in \mathrm{cl}\, conv\,(G \cup -G)\right\}.$$

G-variation is a norm on the subspace $\{f \in X : \|f\|_G < \infty\} \subseteq X$; for its properties see [15], [17], and [18]. In [16] and [18] it was shown that when G is an orthonormal basis of a separable Hilbert space, G-variation is equal to the *l_1-norm with respect to* G, defined for $f \in X$ as $\|f\|_{1,G} = \sum_{g\in G} |f \cdot g|$. For $t > 0$, we define

$$G(t) = \{wg : g \in G, w \in \Re, |w| \le t\}.$$

We now restate in terms of G-variation the Maurey–Jones–Barron theorem [24], [10], [2] and its extension to L_p-spaces [7].

THEOREM 5.1. *Let $(X, \|.\|)$ be a normed linear space, G a bounded nonempty subset, and $s_G = \sup_{g\in G} \|g\|$. For every $f \in X$ and every positive integer n, the following hold:*

(i) *if $(X, \|.\|)$ is a Hilbert space, then*

$$\|f - span_n G\| \le \|f - conv_n\, G(\|f\|_G)\| \le \frac{\|f\|_G\, s_G}{\sqrt{n}};$$

(ii) *if $(X, \|.\|) = (L_p([0,1]^d), \|.\|_p)$, $p \in (1,\infty)$, then*

$$\|f - span_n\, G\| \le \|f - conv_n\, G(\|f\|_G)\| \le \frac{2^{1/\bar{p}+1}\|f\|_G\, s_G}{n^{1/\bar{q}}},$$

where $q = p/(p-1)$, $\bar{p} = \min(p,q)$, and $\bar{q} = \max(p,q)$;

(iii) *if $(X, \|.\|)$ is a separable Hilbert space and G is an orthonormal basis, then*

$$\|f - span_n\, G\| \le \|f - conv_n\, G(\|f\|_G)\| \le \frac{\|f\|_G\, s_G}{2\sqrt{n}}.$$

For the proof of Theorem 5.1(i) and (ii) see [13] and [14], respectively; for the proof of Theorem 5.1(iii) see [16, Theorem 3] and [18, Theorem 2.7].

As a corollary of Theorem 5.1, we obtain the following upper bounds on rates of decrease of infima of error functionals over $span_n\, G$, with n increasing.

COROLLARY 5.2. *Let $(X, \|.\|)$ be a normed linear space with G, C subsets such that $r = \inf_{f\in C} \|f\|_G$ and $s_G = \sup_{g\in G} \|g\|$ are finite. For every positive integer n, the following hold:*

(i) *if $(X, \|.\|)$ is a Hilbert space, then*

$$\inf_{g\in span_n G} e_C(g) \le \inf_{g\in conv_n\, G(r)} e_C(g) \le \frac{r}{\sqrt{n}} s_G;$$

(ii) *if $(X, \|.\|) = (L_p([0,1]^d), \|.\|_p)$, $p \in (1,\infty)$, then*

$$\inf_{g\in span_n G} e_C(g) \le \inf_{g\in conv_n\, G(r)} e_C(g) \le \frac{r\, 2^{1/\bar{p}+1}}{n^{1/\bar{q}}} s_G;$$

(iii) *if $(X, \|.\|)$ is a separable Hilbert space and G is an orthonormal basis, then*

$$\inf_{g\in span_n G} e_C(g) \le \inf_{g\in conv_n\, G(r)} e_C(g) \le \frac{r}{2\sqrt{n}} s_G.$$

Proof. (i) For each $t > r$, choose $f_t \in C$ such that $r \le \|f_t\|_G < t$. By Theorem 5.1(i), for every n we have $\|f_t - conv_n\, G(t)\| \le t\, s_G/\sqrt{n}$. Thus, $\inf_{g\in conv_n G(t)} e_C(g) \le \inf_{g\in conv_n\, G(t)} \|g - f_t\| \le t\, s_G/\sqrt{n}$. Since $G(r) = \bigcap\{G(t) : t > r\}$, we obtain $\inf_{g\in span_n\, G}\, e_C(g) \le \inf_{g\in conv_n\, G(r)} e_C(g) \le r\, s_G/\sqrt{n}$.

Parts (ii) and (iii) are proved similarly to (i) using Theorem 5.1(ii) and (iii), respectively. □

When applied to spaces of functions of d variables, the bounds from Theorem 5.1 and Corollary 5.2 show that for functions in balls of fixed radii in G-variation the curse of dimensionality does not occur. However, the shape of such balls may depend on the number of variables [14], [17], [18].

In the following we apply Corollary 5.2 to admissible sets of Boolean functions in $(\mathcal{B}(\{0,1\}^d), \|.\|_{l_2})$. We give conditions on target sets C, which guarantee rates of minimization of e_C of order $\mathcal{O}(1/\sqrt{n})$ for any number of variables d, for admissible sets of functions in $(\mathcal{B}(\{0,1\}^d), \|.\|_{l_2})$ computable by perceptron neural networks with the signum activation function, defined as sgn(t) $= -1$ or $+1$ according to whether $t < 0$ or $t \ge 0$. Let $\bar{H}_d$ denote the set of functions on $\{0,1\}^d$ computable by signum perceptrons, i.e., $\bar{H}_d = \{f \in \mathcal{B}(\{0,1\}^d) : f(x) = \mathrm{sgn}(v \cdot x + b),\, v \in \Re^d,\, b \in \Re\}$.

We estimate variation with respect to signum perceptrons using variation with respect to the *Fourier orthonormal basis* defined as $F_d = \{f_u : u \in \{0,1\}^d,\, f_u(x) = 2^{-d/2}(-1)^{u\cdot x}\}$ [27]. Every real-valued Boolean function can be represented as $f(x) = 2^{-d/2} \sum_{u\in\{0,1\}^d} \hat{f}(u)(-1)^{u\cdot x}$, where the Fourier coefficients $\hat{f}(u)$ are given by $\hat{f}(u) = 2^{-d/2} \sum_{x\in\{0,1\}^d} f(x)(-1)^{u\cdot x}$. If we interpret the output 1 as -1 and 0 as 1, then the elements of the Fourier basis F_d correspond to the generalized parity functions. The l_1-norm with respect to the Fourier basis, defined as $\|f\|_{1,F_d} = \|\hat{f}\|_{l_1} = \sum_{u\in\{0,1\}^d} |\hat{f}(u)|$, is called the *spectral norm*.

The next proposition gives an upper bound on the rate of decrease of infima of error functionals over perceptron neural networks, in terms of the smallest spectral norm of elements of the target set C.

PROPOSITION 5.3. *Let d be a positive integer, let $r > 0$, and let C be a bounded subset of $(\mathcal{B}(\{0,1\}^d), \|.\|_{l_2})$. Then for every positive integer n, $(span_{dn+1} \bar{H}_d, e_C)$ and $(conv_{dn+1} \bar{H}_d(r), e_C)$ are Tikhonov well-posed in the generalized sense, and for $a = \inf\{\|h\|_{1,F_d} : h \in C\}$, we have*

$$\min_{g \in span_{dn+1} \bar{H}_d} e_C(g) \leq \min_{g \in conv_{dn+1} \bar{H}_d(a)} e_C(g) \leq \frac{a}{2\sqrt{n}}.$$

Proof. For every $u, x \in \{0,1\}^d$, $(-1)^{u \cdot x} = \frac{1+(-1)^d}{2} + \sum_{j=1}^{d} (-1)^j \operatorname{sgn}(u \cdot x - j + \frac{1}{2})$; so every function of the Fourier basis F_d can be expressed as a linear combination of at most $d+1$ signum perceptrons [18]. Hence, any linear combination of n elements of F_d belongs to $span_{dn+1} \bar{H}_d$. As for any orthonormal basis of a separable Hilbert space, G-variation is equal to l_1-norm with respect to G [16], [18], we have $\|f\|_{F_d} = \|f\|_{1,F_d}$, and the statement follows from Proposition 4.1(i) and Corollary 5.2(iii). □

The next two propositions describe target sets for which minimization of error functionals over admissible sets computable by Boolean signum perceptron networks does not exhibit the curse of dimensionality. The first result considers target sets whose elements can be expressed as linear combinations of a "small" number of generalized parities.

PROPOSITION 5.4. *Let d, n, and m be positive integers, let $m \leq 2^d$, and let C be a subset of $(\mathcal{B}(\{0,1\}^d), \|.\|_{l_2})$ such that C contains a function f with at most m Fourier coefficients nonzero and with $\|f\|_{l_2} \leq 1$. Then $(span_{dn+1} \bar{H}_d, e_C)$ and $(conv_{dn+1} \bar{H}_d(\sqrt{m}), e_C)$ are Tikhonov well-posed in the generalized sense and*

$$\min_{g \in span_{dn+1} \bar{H}_d} e_C(g) \leq \min_{g \in conv_{dn+1} \bar{H}_d(\sqrt{m})} e_C(g) \leq \frac{1}{2}\sqrt{\frac{m}{n}} \ .$$

Proof. Let $f \in C$ be such that $f = \sum_{i=1}^{m} w_i g_i$, where $g_i \in F_d$ are the Fourier coefficients of f. Then $\|f\|_{F_d} = \|f\|_{1,F_d} = \sum_{i=1}^{m} |w_i|$. By the Cauchy–Schwarz inequality, $\sum_{i=1}^{m} |w_i| \leq \|w\|_2 \|u\|_2$, where $w = (w_1, \ldots, w_m)$ and $u = (u_1, \ldots, u_m)$, with $u_i = \operatorname{sgn}(w_i)$. As $\|w\|_2 = \|f\|_{l_2} \leq 1$ and $\|u\|_2 = \sqrt{m}$, we have $\|f\|_{1,F_d} \leq \sqrt{m}$. Hence, $\inf\{\|h\|_{1,F_d} : h \in C\} \leq \sqrt{m}$ and the statement follows by Proposition 5.3. □

For C satisfying the assumptions of Proposition 5.4, if e_C is minimized over the set of d-variable Boolean functions computable by networks with $dn+1$ signum perceptrons with $n \geq \frac{m}{4\varepsilon^2}$, then the minimum is bounded from above by ε; the number $\frac{dm}{4\varepsilon^2} + 1$ of perceptrons needed for accuracy ε grows linearly with d.

A *decision tree* (e.g., [19]) is a binary tree with labeled nodes and edges. The *size* of a decision tree is the number of its leaves. A function $f : \{0,1\}^d \to \Re$ is representable by a decision tree if there exists such a tree with internal nodes labeled by variables $x_1, \ldots, x_d$, all pairs of edges outgoing from a node labeled by 0's and 1's, and all leaves labeled by real numbers, so that f can be computed as follows: The computation starts at the root and, after reaching an internal node labeled by x_i, continues along the edge whose label coincides with the actual value of the variable x_i; finally a leaf is reached and its label is equal to $f(x_1, \ldots, x_d)$. Let $DT(s)$ be the set of all functions which are representable by a decision tree of size s.

For any real-valued function f (not identically equal to zero) on a finite set, define the *resolution* of f, $\rho(f)$, to be the ratio of the largest absolute value to the smallest nonzero absolute value. So for a nowhere-zero function f on $\{0,1\}^d$, the resolution of f is $\max_{x \in \{0,1\}^d} |f(x)| / \min_{x \in \{0,1\}^d} |f(x)|$.

PROPOSITION 5.5. *Let d, s be positive integers, let C be a subset of the unit ball in $(\mathcal{B}(\{0,1\}^d), \|.\|_{l_2})$, and suppose C contains a nowhere-zero function in $DT(s)$.*

Then $(span_{dn+1}\,\bar{H}_d, e_C)$ *and* $(conv_{dn+1}\,\bar{H}_d(sb), e_C)$ *are Tikhonov well-posed in the generalized sense and*

$$\min_{g\in span_{dn+1}\,\bar{H}_d} e_C(g) \le \min_{g\in conv_{dn+1}\,\bar{H}_d(sb)} e_C(g) \le \frac{sb}{2\sqrt{n}},$$

where $b = \inf\{\rho(f)\|f\|_{l_2} : f \in C, f(x) \neq 0\ \forall x \in \{0,1\}^d, f \in DT(s)\}$.

Proof. By [18, Theorem 3.4] (which extends [19, Lemma 5.1]), if f is in C and is representable by a decision tree of size s, then $\frac{\|f\|_{1,F_d}}{\|f\|_{l_2}} \le s\rho(f)$, so we get $\|f\|_{1,F_d} \le s\,b$. We conclude by Proposition 5.3. □

Inverting the estimate of Proposition 5.5, we see that $d\left(\frac{sb}{2\varepsilon}\right)^2 + 1$ perceptrons are sufficient for an accuracy ε.

REFERENCES

[1] R. A. Adams, *Sobolev Spaces*, Academic Press, New York, 1975.

[2] A. R. Barron, *Universal approximation bounds for superpositions of a sigmoidal function*, IEEE Trans. Inform. Theory, 39 (1993), pp. 930–945.

[3] R. Bellman, *Dynamic Programming*, Princeton University Press, Princeton, NJ, 1957.

[4] D. P. Bertsekas and J. N. Tsitsiklis, *Neuro-Dynamic Programming*, Athena Scientific, Belmont, MA, 1996.

[5] D. Braess, *Nonlinear Approximation Theory*, Springer-Verlag, Berlin, 1986.

[6] R. Courant, *Differential and Integral Calculus*, Vol. II, Wiley, New York, 1988.

[7] C. Darken, M. Donahue, L. Gurvits, and E. Sontag, *Rates of approximation results motivated by robust neural network learning*, in Proceedings of the 6th Annual ACM Conference on Computational Learning Theory, ACM, New York, 1993, pp. 303–309.

[8] A. L. Dontchev and T. Zolezzi, *Well-Posed Optimization Problems*, Lecture Notes in Math. 1543, Springer-Verlag, Berlin, 1993.

[9] L. Gurvits and P. Koiran, *Approximation and learning of convex superpositions*, J. Comput. System Sci., 55 (1997), pp. 161–170.

[10] L. K. Jones, *A simple lemma on greedy approximation in Hilbert space and convergence rates for projection pursuit regression and neural network training*, Ann. Statist., 20 (1992), pp. 608–613.

[11] P. C. Kainen, V. Kůrková, and A. Vogt, *Best approximation by linear combinations of characteristic functions of half-spaces*, J. Approx. Theory, 122 (2003), pp. 151–159.

[12] V. Kůrková, *Approximation of functions by perceptron networks with bounded number of hidden units*, Neural Networks, 8 (1995), pp. 745–750.

[13] V. Kůrková, *Dimension-independent rates of approximation by neural networks*, in Computer-Intensive Methods in Control and Signal Processing. The Curse of Dimensionality, K. Warwick and M. Kárný, eds., Birkhäuser Boston, Cambridge, MA, 1997, pp. 261–270.

[14] V. Kůrková, *High-dimensional approximation by neural networks*, in Advances in Learning Theory: Methods, Models and Applications, J. Suykens et al., eds., IOS Press, Amsterdam, 2003, pp. 69–88.

[15] V. Kůrková, P. C. Kainen, and V. Kreinovich, *Estimates of the number of hidden units and variation with respect to half-spaces*, Neural Networks, 10 (1997), pp. 1061–1068.

[16] V. Kůrková and M. Sanguineti, *Bounds on rates of variable-basis and neural-network approximation*, IEEE Trans. Inform. Theory, 47 (2001), pp. 2659–2665.

[17] V. Kůrková and M. Sanguineti, *Comparison of worst case errors in linear and neural network approximation*, IEEE Trans. Inform. Theory, 48 (2002), pp. 264–275.

[18] V. Kůrková, P. Savický, and K. Hlaváčková, *Representations and rates of approximation of real–valued Boolean functions by neural networks*, Neural Networks, 11 (1998), pp. 651–659.

[19] E. Kushilevitz and Y. Mansour, *Learning decision trees using the Fourier spectrum*, SIAM J. Comput., 22 (1993), pp. 1331–1348.

[20] M. Leshno, V. Y. Lin, A. Pinkus, and S. Schocken, *Multilayer feedforward networks with a nonpolynomial activation can approximate any function*, Neural Networks, 6 (1993), pp. 861–867.

[21] T. Parisini and R. Zoppoli, *Neural networks for feedback feedforward nonlinear control systems*, IEEE Trans. Neural Networks, 5 (1994), pp. 436–449.

[22] T. Parisini and R. Zoppoli, *Neural approximations for multistage optimal control of nonlinear stochastic systems*, IEEE Trans. Automat. Control, 41 (1996), pp. 889–895.
[23] A. Pinkus, *n-Widths in Approximation Theory*, Springer-Verlag, Berlin, Heidelberg, 1985.
[24] G. Pisier, *Remarques sur un résultat non publié de B. Maurey*, in Séminaire d'Analyse Fonctionnelle 1980–81, Exposé V, École Polytechnique, Centre de Mathématiques, Palaiseau, France, 1981, pp. V.1–V.12.
[25] T. J. Sejnowski and C. R. Rosenberg, *Parallel networks that learn to pronounce English text*, Complex Systems, 1 (1987), pp. 145–168.
[26] I. Singer, *Best Approximation in Normed Linear Spaces by Elements of Linear Subspaces*, Springer-Verlag, Berlin, 1970.
[27] H. J. Weaver, *Applications of Discrete and Continuous Fourier Analysis*, Wiley, New York, 1983.
[28] R. Zoppoli and T. Parisini, *Learning techniques and neural networks for the solution of N-stage nonlinear nonquadratic optimal control problems*, in Systems, Models and Feedback: Theory and Applications, A. Isidori and T. J. Tarn, eds., Birkhäuser Boston, Cambridge, MA, 1992, pp. 193–210.
[29] R. Zoppoli, M. Sanguineti, and T. Parisini, *Approximating networks and extended Ritz method for the solution of functional optimization problems*, J. Optim. Theory Appl., 112 (2002), pp. 403–440.

SIAM J. OPTIM.
Vol. 14, No. 3, pp. 743–756

MINIMIZING NONCONVEX NONSMOOTH FUNCTIONS VIA CUTTING PLANES AND PROXIMITY CONTROL*

A. FUDULI†, M. GAUDIOSO‡, AND G. GIALLOMBARDO§

Abstract. We describe an extension of the classical cutting plane algorithm to tackle the unconstrained minimization of a nonconvex, not necessarily differentiable function of several variables.

The method is based on the construction of both a lower and an upper polyhedral approximation to the objective function and is related to the use of the concept of proximal trajectory.

Convergence to a stationary point is proved for weakly semismooth functions.

Key words. nonsmooth optimization, cutting planes, bundle methods, proximal trajectory

AMS subject classifications. 90C26, 65K05

DOI. 10.1137/S1052623402411459

1. Introduction.

Most of the numerical methods for solving nonsmooth optimization problems aim at minimizing convex functions of several variables, and convex analysis is in fact the background theory [9, 22]. Although generalized gradient theory [2] and codifferentiable functions theory [4] provide an interesting framework for dealing with nonsmooth nonconvex functions, apparently they have not yet been fully exploited from the numerical point of view.

Most of the existing algorithms for nonsmooth optimization fall into the class of subgradient and space dilatation–type algorithms [24], bundle methods [7, 10, 18], or minmax-type algorithms [5, 21] (convexity is not necessary in the latter).

In particular, the bundle methods family is based on the cutting plane method, first described in [1, 11], where the convexity of the objective function is the fundamental assumption. In fact the extension of the cutting plane method to the nonconvex case is not straightforward. A basic observation is that, in general, first order information no longer provides a lower approximation to the objective function independently of the nonsmoothness assumption.

Thus, the optimization of the cutting plane approximation does not necessarily give an optimistic estimate of the obtainable reduction in the objective function. Moreover, such a model might even fail to interpolate the objective function at the points where its value is known.

On the other hand it is apparent that a number of ideas valid in the convex nonsmooth framework are valuable also in the treatment of the nonconvex case.

For example, search directions obtained as the opposite of a convex combination of gradients, relative to points close to each other, appear often to enjoy good de-

*Received by the editors July 17, 2002; accepted for publication (in revised form) July 30, 2003; published electronically January 30, 2004. This work was partially supported by the Italian Ministero dell'Istruzione, dell'Università e della Ricerca Scientifica, under FIRB project *Large Scale Nonlinear Optimization* (RBNE01WBBB).

http://www.siam.org/journals/siopt/14-3/41145.html

†Dipartimento di Ingegneria dell'Innovazione, Università di Lecce, Via Monteroni, 73100 Lecce (LE), Italia (antonio.fuduli@unile.it).

‡Dipartimento di Elettronica Informatica e Sistemistica, Università della Calabria, 87036 Rende (CS), Italia (gaudioso@deis.unical.it).

§Judge Institute of Management, University of Cambridge, Cambridge CB2 1AG, UK, and Dipartimento di Elettronica Informatica e Sistemistica, Università della Calabria, 87036 Rende (CS), Italia (g.giallombardo@jims.cam.ac.uk, giallo@deis.unical.it).

scent properties for nonconvex functions too, especially when the contour lines have a narrow valley shape.

Thus it appears reasonable to claim that nonconvex nonsmooth minimization can benefit from the experience of convex optimization, but the approaches valid in the latter case cannot be trivially extended.

Most of the authors who have extended bundle methods to the nonconvex case have considered piecewise affine models embedding possible downward shifting of the affine pieces [15, 19, 23]. However, the amount of the shifting appears somehow arbitrary.

In this paper we present an iterative algorithm which is still based on first order approximations to the objective function.

The main difference with other known methods is that our algorithm makes a distinction between affine pieces that exhibit a *convex* or a *concave* behavior relative to the current point in the iterative procedure. Furthermore, the use of downward shifting is restricted to some particular cases.

The following notation is adopted throughout the paper. We denote by $\|\cdot\|$ the Euclidean norm in $\mathbb{R}^n$, by $a^T b$ the inner product of the vectors a and b, and by e a vector of ones of appropriate dimension. The generalized gradient of a Lipschitz function $f : \mathbb{R}^n \mapsto \mathbb{R}$ at any point x is denoted by $\partial f(x)$.

2. The model. Consider the following unconstrained minimization problem:

$$\min_{x\in\mathbb{R}^n} \; f(x),$$

where $f : \mathbb{R}^n \mapsto \mathbb{R}$ is not necessarily differentiable.

We assume that f is locally Lipschitz; i.e., it is Lipschitz on every bounded set. Since f is locally Lipschitz, then it is differentiable almost everywhere. It is well known [2] that, under the above hypotheses, there is defined at each point x the generalized gradient (or Clarke's gradient or subdifferential)

$$\partial f(x) = \operatorname{conv}\{g \;|g \in \mathbb{R}^n, \nabla f(x_k) \to g,\; x_k \to x,\; x_k \notin \Omega_f\},$$

where Ω_f is the set (of zero measure) where f is not differentiable. An extension of the generalized gradient is the *Goldstein ϵ-subdifferential* $\partial_\epsilon^G f(x)$ defined as

$$\partial_\epsilon^G f(x) = \operatorname{conv}\{\partial f(y) \;|\|y - x\| \le \epsilon\}.$$

We assume also that we are able to calculate at each point x both the objective function value and a subgradient $g \in \partial f(x)$, i.e., an element of the generalized gradient.

Now we describe the basic idea of our method, focusing on the differences with respect to the methods tailored on the convex case. We denote by x_j the current estimate of the minimum in an iterative procedure and by g_j any subgradient of f at x_j. The bundle of available information is the set of elements

$$(x_i, f(x_i), g_i, \alpha_i, a_i), \quad i \in I,$$

where x_i, $i \in I$, are the points touched in the procedure, g_i is a subgradient of f at x_i, α_i is the linearization error between the actual value of the objective function at x_j and the linear expansion generated at x_i and evaluated at x_j, i.e.,

$$\alpha_i \stackrel{\triangle}{=} f(x_j) - f(x_i) - g_i^T(x_j - x_i),$$

and

$$a_i \stackrel{\triangle}{=} \|x_j - x_i\| .$$

We recall that the classical cutting plane method [1, 11] minimizes at each iteration the cutting plane function $f_j(x)$ defined as

$$f_j(x) = \max_{i \in I} \left\{ f(x_i) + g_i^T (x - x_i) \right\} .$$

The minimization of $f_j(x)$ can be put in linear programming form as

$$\left\{ \begin{array}{ll} \min\limits_{\eta, x} & \eta \\ & \eta \geq f(x_i) + g_i^T (x - x_i), \quad i \in I, \end{array} \right. \tag{2.1}$$

which is equivalent to solving

$$\left\{ \begin{array}{ll} \min\limits_{v, d} & v \\ & v \geq g_i^T d - \alpha_i, \quad i \in I, \end{array} \right. \tag{2.2}$$

where d is the "displacement" from x_j, i.e., $d \stackrel{\triangle}{=} x - x_j$. In what follows we will refer to the point x_j as the "stability center."

It is worth noting that in the nonconvex case α_i may be negative, since the first order expansion at any point does not necessarily support from below the epigraph of the function.

Thus we partition the set I into two sets I_+ and I_-, defined as follows:

$$I_+ \stackrel{\triangle}{=} \{i | \alpha_i \geq 0\}, \qquad I_- \stackrel{\triangle}{=} \{i | \alpha_i < 0\}. \tag{2.3}$$

The bundles defined by the index sets I_+ and I_- are characterized by points that somehow exhibit, respectively, a "convex behavior" and a "concave behavior" relative to x_j. We observe that I_+ is never empty as at least the element $(x_j, f(x_j), g_j, 0, 0)$ belongs to the bundle.

The basic idea of our approach is to treat differently the two bundles in the construction of a piecewise affine model.

We define the following piecewise affine functions:

$$\Delta^+(d) \stackrel{\triangle}{=} \max_{i \in I_+} \left\{ g_i^T d - \alpha_i \right\}$$

and

$$\Delta^-(d) \stackrel{\triangle}{=} \min_{i \in I_-} \left\{ g_i^T d - \alpha_i \right\} .$$

In fact $\Delta^+(d)$ is intended as an approximation of the difference function

$$h(d) \stackrel{\triangle}{=} f(x_j + d) - f(x_j),$$

which interpolates it at $d = 0$ (since the index j belongs to I_+).

On the other hand $\Delta^-(d)$ is a locally "pessimistic" approximation to the difference function $h(d)$. When $I_- \neq \emptyset$, since we have $\Delta^+(0) < \Delta^-(0)$, it appears reasonable to consider the approximation $\Delta^+(d)$ significant as far as

$$\Delta^+(d) \leq \Delta^-(d).$$

In other words we introduce a kind of trust region model $\mathcal{S}$ defined as

$$\mathcal{S} = \{d | \Delta^+(d) \leq \Delta^-(d)\}.$$

In addition we introduce proximity control [13] into our approach by defining the "proximal trajectory" [6] of $\Delta^+(d)$ as the optimal solution d_γ to the following convex quadratic program, parameterized in the nonnegative scalar γ, where the constraints ensure that $d \in \mathcal{S}$:

$$QP(\gamma) \qquad \begin{cases} z_\gamma = \min\limits_{v,d} & \gamma v + \dfrac{1}{2}\|d\|^2 \\ & v \geq g_i^T d - \alpha_i, \quad i \in I_+, \\ & v \leq g_i^T d - \alpha_i, \quad i \in I_-. \end{cases}$$

We observe that $z_\gamma \leq 0$, as the couple $(v, d) = (0, 0)$ is feasible; we have consequently that the optimal value of v cannot be positive.

The dual of the program $QP(\gamma)$ can be written in the form

$$DP(\gamma) \qquad \begin{cases} w_\gamma = \min\limits_{\lambda \geq 0, \mu \geq 0} & \dfrac{1}{2}\|G_+\lambda - G_-\mu\|^2 + \alpha_+^T\lambda - \alpha_-^T\mu \\ & e^T\lambda - e^T\mu = \gamma, \end{cases}$$

where G_+ and G_- are matrices whose columns are, respectively, the vectors g_i, $i \in I_+$, and g_i, $i \in I_-$. Analogously, the terms α_i, $i \in I_+$, and α_i, $i \in I_-$, are grouped in the vectors α_+ and α_-, respectively.

The optimal primal solution (v_γ, d_γ) is related to the optimal dual solution $(\lambda_\gamma, \mu_\gamma)$ by the following formulae:

$$d_\gamma = -G_+\lambda_\gamma + G_-\mu_\gamma, \tag{2.4a}$$

$$v_\gamma = -\frac{1}{\gamma}\left(\|d_\gamma\|^2 + \alpha_+^T\lambda_\gamma - \alpha_-^T\mu_\gamma\right). \tag{2.4b}$$

We remark that the proximal trajectory emanates from the stability center x_j.

Before giving a formal description of the algorithm, we state some simple properties of problem $QP(\gamma)$.

LEMMA 2.1. *Let $\gamma_1 > \gamma_2 > 0$. Then the following relations hold:*

(i) $z_{\gamma_1} \leq z_{\gamma_2}$;
(ii) $v_{\gamma_1} \leq v_{\gamma_2}$;
(iii) $\|d_{\gamma_1}\| \geq \|d_{\gamma_2}\|$.

Proof. (i) From the definitions of z_γ, v_γ, and d_γ, and taking into account $\gamma_1 > \gamma_2 > 0$, it follows that

$$z_{\gamma_1} = \gamma_1 v_{\gamma_1} + \frac{1}{2}\|d_{\gamma_1}\|^2 \leq \gamma_1 v_{\gamma_2} + \frac{1}{2}\|d_{\gamma_2}\|^2 \leq \gamma_2 v_{\gamma_2} + \frac{1}{2}\|d_{\gamma_2}\|^2 = z_{\gamma_2}.$$

(ii) Assume $v_{\gamma_1} > v_{\gamma_2}$. Then, since $\gamma_1 > \gamma_2$, it holds that

$$0 < (\gamma_1 - \gamma_2)(v_{\gamma_1} - v_{\gamma_2}) = \gamma_1 v_{\gamma_1} + \gamma_2 v_{\gamma_2} - (\gamma_1 v_{\gamma_2} + \gamma_2 v_{\gamma_1}).$$

By adding and subtracting to the right-hand side

$$\frac{1}{2}\|d_{\gamma_1}\|^2 + \frac{1}{2}\|d_{\gamma_2}\|^2$$

we would have

$$0 < \left[\left(\gamma_1 v_{\gamma_1} + \frac{1}{2}\|d_{\gamma_1}\|^2\right) - \left(\gamma_1 v_{\gamma_2} + \frac{1}{2}\|d_{\gamma_2}\|^2\right)\right]$$

$$+ \left[\left(\gamma_2 v_{\gamma_2} + \frac{1}{2}\|d_{\gamma_2}\|^2\right) - \left(\gamma_2 v_{\gamma_1} + \frac{1}{2}\|d_{\gamma_1}\|^2\right)\right],$$

which is a contradiction, since, by the definitions, the right-hand side is the sum of two nonpositive quantities.

(iii) Assume $\|d_{\gamma_1}\| < \|d_{\gamma_2}\|$. Then (ii) implies

$$\gamma_2 v_{\gamma_1} + \frac{1}{2}\|d_{\gamma_1}\|^2 < \gamma_2 v_{\gamma_2} + \frac{1}{2}\|d_{\gamma_2}\|^2,$$

which contradicts the optimality of $(v_{\gamma_2}, d_{\gamma_2})$. ∎

LEMMA 2.2. *For any $\gamma > 0$ the following relations hold:*

(i) $\|d_\gamma\| \le 2\gamma\|g_j\|$;

(ii) $z_\gamma \ge -\frac{1}{2}\gamma^2\|g_j\|^2$;

(iii) $|v_\gamma| \ge \frac{1}{2\gamma}\|d_\gamma\|^2$.

Proof. (i) Since $z_\gamma \le 0$ we have

$$(v_\gamma, d_\gamma) \in \mathcal{D} \triangleq \left\{(v,d) \mid \gamma v + \frac{1}{2}\|d\|^2 \le 0\right\}.$$

The property follows by noting that the objective function of $QP(\gamma)$ is minorized by

$$\gamma g_j^T d + \frac{1}{2}\|d\|^2. \tag{2.5}$$

(ii) The property follows by noting that $-\frac{1}{2}\gamma^2\|g_j\|^2$ is the minimum value of the minorizing function (2.5).

(iii) The property follows as a consequence of $z_\gamma \le 0$. ∎

3. The algorithm. In this section we describe an algorithm based on repeatedly solving problem $QP(\gamma)$, or, equivalently, $DP(\gamma)$. The core of the algorithm is the "main iteration," i.e., the set of steps where the stability center remains unchanged.

Two exits from the "main iteration" may occur:

(i) termination of the whole algorithm due to the satisfaction of an approximate stationarity condition;

(ii) update of the stability center due to the satisfaction of a sufficient decrease condition.

The initialization of the algorithm requires a starting point $x_0 \in \mathbb{R}^n$. The initial stability center y is set equal to x_0. The initial bundle is made up of just one element $(y, f(y), g(y), 0, 0)$, where $g(y) \in \partial f(y)$, so that I_- is the empty set, while I_+ is a singleton. The following global parameters are to be set:

- the stationarity tolerance $\delta > 0$ and the proximity measure $\epsilon > 0$;
- the descent parameter $m \in (0,1)$ and the cut parameter $\rho \in (m,1)$;

- the reduction parameter $r \in (0,1)$ and the increase parameter $R > 1$.

A short description of the algorithm is the following.

ALGORITHM OUTLINE.

1. Initialization.
2. Execute the "main iteration."
3. Update the bundle of information with respect to the new stability center and return to 2.

In what follows we describe in detail the "main iteration" without indexing it for the sake of notational simplicity.

The following local parameters are set each time the "main iteration" is entered:

- the proximity measure $\theta > 0$;
- the safeguard parameters γ_{min} and γ_{max}, $0 < \gamma_{min} < \gamma_{max}$.

We remark that in general the "main iteration" maintains the (updated) bundle of information from previous iterations. Updating the bundle is necessary since the quantities α_i and a_i are dependent on the stability center.

ALGORITHM 3.1 (main iteration).

0. *If* $\|g(y)\| \le \delta$, *then* STOP *(stationarity achieved).*
Set

$$\gamma_{min} := \frac{r\epsilon}{2\|g(y)\|}, \quad \gamma_{max} := R\gamma_{min}, \quad \theta := r\gamma_{min}\delta.$$

1. *Construct the proximal trajectory* d_γ *for increasing values of* γ *and choose* $\hat{\gamma}$ *equal to the minimum value of* $\gamma \in [\gamma_{min}, \gamma_{max}]$ *such that*

$$f(y + d_\gamma) > f(y) + mv_\gamma$$

if such γ *does exist. Otherwise set* $\hat{\gamma} := \gamma_{max}$. *If* $\|d_{\hat{\gamma}}\| > \theta$, *go to* 3.

2. *Set*

$$I_+ := I_+ \setminus \{i \in I_+ \mid a_i > \epsilon\}$$

and

$$I_- := I_- \setminus \{i \in I_- \mid a_i > \epsilon\}.$$

Calculate

$$g^* = \min_{g \in \mathrm{conv}\{g_i \mid i \in I_+\}} \|g\|.$$

If $\|g^*\| \le \delta$, *then* STOP *(stationarity achieved).*
Else set $\gamma_{max} := \gamma_{max} - r(\gamma_{max} - \gamma_{min})$ *and go to* 1.

3. *Set* $x_{\hat{\gamma}} := y + d_{\hat{\gamma}}$, *calculate* $g_{\hat{\gamma}} \in \partial f(x_{\hat{\gamma}})$, *and set*

$$\alpha_{\hat{\gamma}} := f(y) - f(x_{\hat{\gamma}}) + g_{\hat{\gamma}}^T d_{\hat{\gamma}}.$$

4. (a) *If* $\alpha_{\hat{\gamma}} < 0$ *and* $\|d_{\hat{\gamma}}\| > \epsilon$, *then insert the element* $(x_{\hat{\gamma}}, f(x_{\hat{\gamma}}), g_{\hat{\gamma}}, \alpha_{\hat{\gamma}}, \|d_{\hat{\gamma}}\|)$ *into the bundle for an appropriate value of* $i \in I_-$ *and set* $\hat{\gamma} := \hat{\gamma} - r(\hat{\gamma} - \gamma_{min})$.

(b) *Else, if* $g_{\hat{\gamma}}^T d_{\hat{\gamma}} \ge \rho v_{\hat{\gamma}}$, *then insert the element* $(x_{\hat{\gamma}}, f(x_{\hat{\gamma}}), g_{\hat{\gamma}}, \max(0, \alpha_{\hat{\gamma}}), \|d_{\hat{\gamma}}\|)$ *into the bundle for an appropriate value of* $i \in I_+$.

(c) *Else find a scalar* $t \in (0,1)$ *such that* $g(t) \in \partial f(y + td_{\hat{\gamma}})$ *satisfies the condition* $g(t)^T d_{\hat{\gamma}} \ge \rho v_{\hat{\gamma}}$ *and insert the element* $(y + td_{\hat{\gamma}}, f(y + td_{\hat{\gamma}}), g(t), \max(0, \alpha_t), t\|d_{\hat{\gamma}}\|)$ *into the bundle for an appropriate value of* $i \in I_+$, *where* $\alpha_t = f(y) - f(y + td_{\hat{\gamma}}) + tg(t)^T d_{\hat{\gamma}}$.

5. *If* $\|d_{\hat{\gamma}}\| \leq \theta$, *go to* 2. *If*

$$f(x_{\hat{\gamma}}) \leq f(y) + mv_{\hat{\gamma}}, \tag{3.1}$$

set the new stability center $y := x_{\hat{\gamma}}$ *and* EXIT *from the main iteration.*

6. *Solve* $QP(\hat{\gamma})$, *or, equivalently,* $DP(\hat{\gamma})$, *obtain both the primal and the dual optimal solution* $(v_{\hat{\gamma}}, d_{\hat{\gamma}})$ *and* $(\lambda_{\hat{\gamma}}, \mu_{\hat{\gamma}})$, *and go to* 3.

Some explanations are in order. The stationarity test at step 0 prevents the "main iteration" from being executed if enough information is already available to assess the stationarity of y.

The construction of the proximal trajectory at step 1 may be discretized by repeatedly solving $QP(\gamma)$ for increasing values of γ, or by adopting techniques of the type described in [6] (see also [14]).

The rationale of the test executed at step 2 is that the occurrence of a "small" (in norm) displacement d_γ corresponding to a "large" value of γ denotes either that a stationary point has been reached or that the model is inconsistent. We discriminate between these two cases by considering the distance measures a_i (bundle deletion at step 2). We observe that the choice of $\hat{\gamma}$ defines implicitly a constraint on the norm of $d_{\hat{\gamma}}$ (see Lemma 2.2(i)). On the other hand $\|d_{\hat{\gamma}}\| \leq \theta$ is never a consequence of the choice of a too small $\hat{\gamma}$. In fact we note that if $\|g(y)\| > \delta$, it holds that

$$\|d_{\gamma_{min}}\| \leq 2\gamma_{min}\|g(y)\| = \frac{2\|g(y)\|}{r\delta}\theta,$$

with the right-hand side strictly greater than θ.

We remark that the insertion of a bundle index into I_+ or I_- at step 4 is not simply based on the sign of α_i. In fact, in case $\alpha_i < 0$ and $a_i \leq \epsilon$, the index i is inserted into I_+, and not into I_- as would be expected, and α_i is set equal to zero; that is, the related affine piece is shifted downward of a quantity equal to $|\alpha_i|$ (see also [23]). This is aimed at letting all elements of the Goldstein ϵ-subdifferential at y contribute to the construction of the polyhedral approximation $\Delta^+(d)$, and also guarantees that the model interpolates the objective function at y. Furthermore the reduction of $\hat{\gamma}$, whenever a bundle index is inserted into I_-, is aimed at avoiding the same point solution $x_{\hat{\gamma}}$ being generated infinitely many times. To explain case (c) at step 4 we observe that the downward shifting of an affine piece, when $\hat{\alpha} < 0$, does not always cut out the point solution of $QP(\hat{\gamma})$ generated at the previous iteration. A sufficient condition for such a cut to be effective is that $g_{\hat{\gamma}}^T d_{\hat{\gamma}} \geq \rho v_{\hat{\gamma}}$. If such a condition is not verified, we resort to a line search–type procedure which allows us to find a point $y + td_{\hat{\gamma}}$, with $t \in (0, 1)$, satisfying $g(t)^T d_{\hat{\gamma}} \geq \rho v_{\hat{\gamma}}$, where $g(t) \in \partial f(y + td_{\hat{\gamma}})$ (see also [23]).

Notice that the search direction $d_{\hat{\gamma}}$ is calculated only at steps 1 and 6. This means that in passing through step 4(a), the reduction of $\hat{\gamma}$ does not cause an immediate change in $d_{\hat{\gamma}}$, and indeed the search direction used at step 5 is the one available right before such a reduction.

Finally we observe that every time the stability center is updated, the parameters α_i and a_i are to be updated for each element of the bundle as well, which may result in changing the assignment of the corresponding index i from I_+ to I_- and vice versa.

4. Convergence. In this section we prove the termination of the algorithm at a point satisfying an approximate stationarity condition. In particular we prove that, for any given $\epsilon > 0$ and $\delta > 0$, it is possible to set the input parameters such that,

after a finite number of "main iteration" executions, the algorithm stops at a point y satisfying the condition

$$\|g^*\| \le \delta \quad \text{with } g^* \in \partial_\epsilon^G f(y).$$

Throughout the section we make the following assumptions:

(A1) f is weakly semismooth;

(A2) the set $\mathcal{F}_0 = \{x \in \mathbb{R}^n \mid f(x) \le f(x_0)\}$ is compact.

We recall that a function $f : \mathbb{R}^n \mapsto \mathbb{R}$ is weakly semismooth at x (see [16, 20, 23]) if it is Lipschitz around x and

$$\lim_{t \downarrow 0} g(t)^T d$$

exists for all $d \in \mathbb{R}^n$, where $g(t) \in \partial f(x+td)$. In particular, if f is weakly semismooth at x, the directional derivative $f'(x,d)$ of f along the direction d exists for all $d \in \mathbb{R}^n$ and

$$f'(x,d) = \lim_{t \downarrow 0} g(t)^T d.$$

Moreover, f is weakly semismooth on $\mathbb{R}^n$ if it is weakly semismooth at each $x \in \mathbb{R}^n$.

Before proving finite termination of the "main iteration" we introduce the following lemma.

LEMMA 4.1. *Let $\{(v_{\hat\gamma}^{(k)}, d_{\hat\gamma}^{(k)})\}_{k \in \mathcal{K}}$ be a subsequence generated within a single "main iteration" such that*

$$\|d_{\hat\gamma}^{(k)}\| > \theta$$

and

$$f(y + d_{\hat\gamma}^{(k)}) - f(y) > m v_{\hat\gamma}^{(k)},$$

with the algorithm looping from step 3 to step 6. Then the following hold:

(i) *there exists an index $\bar k$ such that for each $k \ge \bar k$, $k \in \mathcal{K}$, every new bundle index is inserted into I_+ and $\hat\gamma$ remains unchanged;*

(ii) *step 4(c) of the algorithm is well posed; i.e., there exist two nonnegative scalars $t_1^{(k)}$ and $t_2^{(k)}$, $0 \le t_1^{(k)} < t_2^{(k)} < 1$, such that for any $t \in [t_1^{(k)}, t_2^{(k)}]$ the condition*

$$g(t)^T d_{\hat\gamma}^{(k)} \ge \rho v_{\hat\gamma}^{(k)}$$

is satisfied for every $g(t) \in \partial f(y + t d_{\hat\gamma}^{(k)})$;

(iii) *whenever a new bundle index is inserted into I_+ the condition*

$$g_k^T d_{\hat\gamma}^{(k)} \ge \rho v_{\hat\gamma}^{(k)}$$

holds, where g_k is the subgradient corresponding to the new bundle element.

Proof. (i) We observe that an infinite sequence of bundle index insertions into I_- cannot take place, as a consequence of the reduction of $\hat\gamma$ at step 4(a) of the algorithm. In particular, no bundle index can be inserted into I_- as soon as $\hat\gamma$ falls below the threshold $\frac{\epsilon}{2\|g(y)\|}$.

(ii) Since the directional derivative $f'(y + t^{(k)} d_{\hat\gamma}^{(k)}, d_{\hat\gamma}^{(k)})$ exists for any $t^{(k)} \ge 0$, from the mean value theorem (see [3], Chap. 3, Prop. 3.1) it follows that

$$f(y + d_{\hat\gamma}^{(k)}) - f(y) = c \tag{4.1}$$

for some $c \in [f'_{\inf}, f'_{\sup}]$, where

$$f'_{\inf} \stackrel{\triangle}{=} \inf_{0 \le t^{(k)} \le 1} f'(y + t^{(k)} d^{(k)}_{\hat\gamma}, d^{(k)}_{\hat\gamma}) \quad \text{and} \quad f'_{\sup} \stackrel{\triangle}{=} \sup_{0 \le t^{(k)} \le 1} f'(y + t^{(k)} d^{(k)}_{\hat\gamma}, d^{(k)}_{\hat\gamma}).$$

Moreover, taking into account that the sufficient decrease condition is not satisfied, i.e.,

$$\rho v^{(k)}_{\hat\gamma} < m v^{(k)}_{\hat\gamma} < f(y + d^{(k)}_{\hat\gamma}) - f(y),$$

by (4.1) and the definition of $f'_{\sup}$ there exists a scalar $\bar t^{(k)} \in (0,1)$ such that

$$\rho v^{(k)}_{\hat\gamma} < f'(y + \bar t^{(k)} d^{(k)}_{\hat\gamma}, d^{(k)}_{\hat\gamma}).$$

Thus the thesis follows as a consequence of the weakly semismoothness assumption.

(iii) We observe that the condition $g_k^T d^{(k)}_{\hat\gamma} \ge \rho v^{(k)}_{\hat\gamma}$ is ensured either by construction or by the fact that

$$g_k^T d^{(k)}_{\hat\gamma} \ge g_k^T d^{(k)}_{\hat\gamma} - \alpha^{(k)}_{\hat\gamma} = f(y + d^{(k)}_{\hat\gamma}) - f(y) > m v^{(k)}_{\hat\gamma} > \rho v^{(k)}_{\hat\gamma}$$

whenever $\alpha^{(k)}_{\hat\gamma} \ge 0$. ∎

Now we can prove finite termination of the "main iteration."

LEMMA 4.2. *The* "main iteration" *terminates after a finite number of steps.*

Proof. To prove finiteness of the "main iteration" it is necessary to demonstrate that in a finite number of steps either the stop at step 2 or the exit at step 5 is achieved.

We start by proving that the algorithm cannot pass infinitely many times through step 2. Assume by contradiction that such a case occurs, and let us index by $k \in \mathcal{K}$ all the quantities referred to in the kth passage. We have

$$\|d^{(k)}_{\hat\gamma}\| \le \theta$$

and

$$\|g^{*^{(k)}}\| > \delta.$$

Observe that $\hat\gamma \le \gamma_{max}$ and that by construction γ_{max} falls in a finite number of steps below the threshold $\frac{\epsilon}{2\|g(y)\|}$. Thus, from Lemma 2.2(i), it follows that asymptotically $\|d^{(k)}_{\hat\gamma}\| \le \epsilon$, which in turn implies that the indices of the new bundle elements are asymptotically inserted into I_+ and are never removed.

Moreover, the bundle insertion rules at step 4 allow us to insert an index into I_- only if $\|d_{\hat\gamma}\| > \epsilon$, and this implies that whenever a passage at step 2 occurs, all the elements with index $i \in I_-$ are removed.

From the above considerations, taking into account (2.4a) and the constraint $e^T\lambda - e^T\mu = \hat\gamma$ in the dual problem $DP(\hat\gamma)$, it follows that there exists an index $\bar k \in \mathcal{K}$ such that for all $k \ge \bar k$ the direction $d^{(k)}_{\hat\gamma}$ can be expressed in the form

$$d^{(k)}_{\hat\gamma} = -\hat\gamma g^{(k)},$$

with $g^{(k)} \in \operatorname{conv}\{g_i \mid i \in I^{(k)}_+\}$. But since $\|d^{(k)}_{\hat\gamma}\| \le \theta$ and $\|g^{*^{(k)}}\| > \delta$, we have

$$\theta \ge \|d^{(k)}_{\hat\gamma}\| = \hat\gamma \|g^{(k)}\| \ge \gamma_{min} \|g^{*^{(k)}}\| > \frac{\theta}{\delta}\delta = \theta,$$

reaching a contradiction.

So far we have proved that an infinite number of passages through step 2 cannot occur. To complete the proof of termination we need to show that it is impossible to have infinitely many times $\|d_{\hat{\gamma}}\| > \theta$ and the descent condition (3.1) not satisfied, with the algorithm looping between steps 3 and 6.

Indexing again by $k \in \mathcal{K}$ the kth passage through such a loop, we observe that, by Lemma 4.1(i), there exists an index $\bar{k}$ such that for every $k \geq \bar{k}$ the index of each new bundle element is put in I_+ with $\hat{\gamma}$ remaining unchanged. Under such a condition, for $k \geq \bar{k}$ the sequence $\{z_{\hat{\gamma}}^{(k)}\}$ is monotonically nondecreasing, bounded, and hence convergent. Moreover, since the sequence $\{d_{\hat{\gamma}}^{(k)}\}$ is bounded in norm, it admits a convergent subsequence, say $\{d_{\hat{\gamma}}^{(k)}\}_{k \in \mathcal{K}' \subseteq \mathcal{K}}$.

The above considerations imply also that the sequence $\{v_{\hat{\gamma}}^{(k)}\}_{k \in \mathcal{K}' \subseteq \mathcal{K}}$ is convergent to a nonpositive limit, say $\bar{v}$. Now assume that $\bar{v} < 0$, let i and j be two successive indices in $\mathcal{K}'$, and let $\beta_i = \max\{0, \alpha_i\}$, with $\alpha_i = f(y) - f(y + d_{\hat{\gamma}}^{(i)}) + g_i^T d_{\hat{\gamma}}^{(i)}$ and $g_i \in \partial f(y + d_{\hat{\gamma}}^{(i)})$. We have

$$v_{\hat{\gamma}}^{(j)} \geq g_i^T d_{\hat{\gamma}}^{(j)} - \beta_i, \tag{4.2}$$

$$f(y + d_{\hat{\gamma}}^{(i)}) - f(y) > m v_{\hat{\gamma}}^{(i)},$$

and

$$g_i^T d_{\hat{\gamma}}^{(i)} \geq \rho v_{\hat{\gamma}}^{(i)}.$$

We note that

$$g_i^T d_{\hat{\gamma}}^{(i)} - \beta_i \geq \rho v_{\hat{\gamma}}^{(i)}. \tag{4.3}$$

This inequality is trivial for $\beta_i = 0$. If, on the other hand, $\beta_i = \alpha_i$, then taking into account that $\rho > m$, it holds that

$$g_i^T d_{\hat{\gamma}}^{(i)} - \beta_i = f(y + d_{\hat{\gamma}}^{(i)}) - f(y) > m v_{\hat{\gamma}}^{(i)} > \rho v_{\hat{\gamma}}^{(i)}.$$

Combining (4.2) and (4.3) we obtain

$$v_{\hat{\gamma}}^{(j)} - \rho v_{\hat{\gamma}}^{(i)} \geq g_i^T (d_{\hat{\gamma}}^{(j)} - d_{\hat{\gamma}}^{(i)}),$$

and passing to the limit

$$(1 - \rho)\bar{v} \geq 0,$$

which contradicts $\bar{v} < 0$. Hence we conclude that $\bar{v} = 0$, which, by Lemma 2.2(iii), contradicts the fact that $\|d_{\hat{\gamma}}^{(k)}\| > \theta$ for all $k \in \mathcal{K}$. □

Remark. Since $\gamma_{min} = \frac{r\epsilon}{2\|g(y)\|}$ and $\theta = r\gamma_{min}\delta$ it follows that

$$\theta \geq \frac{r^2 \epsilon \delta}{2L_0}, \tag{4.4}$$

where L_0 is the Lipschitz constant of f on the set $\mathcal{F}_0$.

Now we are ready to prove the overall finiteness of the algorithm.

THEOREM 4.3. *For any $\epsilon > 0$ and $\delta > 0$, the algorithm stops in a finite number of "main iterations" at a point satisfying the approximate stationarity condition*

$$\|g^*\| \le \delta \quad \text{with } g^* \in \partial_\epsilon^G f(y). \tag{4.5}$$

Proof. The approximate stationarity condition (4.5) is exactly the stopping condition tested at step 2 of the "main iteration." Now suppose that it is not verified for an infinite number of "main iteration" executions. From Lemma 4.2 it follows that infinitely many times the descent condition is satisfied. Let $y^{(k)}$ be the stability center at the kth passage through "main iteration"; then $\|d_{\hat\gamma}^{(k)}\| > \theta^{(k)}$,

$$f(y^{(k+1)}) \le f(y^{(k)}) + m v_{\hat\gamma}^{(k)},$$

and

$$f(y^{(k+1)}) - f(y^{(0)}) \le m \sum_{i=0}^{k} v_{\hat\gamma}^{(i)}.$$

Now consider that by (4.4) $\|d_{\hat\gamma}^{(i)}\|$ is bounded away from zero. Then from Lemma 2.2(iii) it follows that $v_{\hat\gamma}^{(i)}$ is bounded away from zero as well. Therefore, by passing to the limit we obtain

$$\lim_{k\to\infty} f(y^{(k+1)}) - f(y^{(0)}) \le -\infty,$$

which is a contradiction, since f is bounded from below as a consequence of assumptions (A1) and (A2). ∎

5. Practical implementation and numerical results. The algorithm described in section 3 cannot be immediately implemented, since it may require unbounded storage. In fact it does not encompass any mechanism to control the growth of the bundle size. Also the convergence properties described in section 4 are derived under the hypothesis that the bundle size can grow indefinitely. Thus, before passing to the implementation issues, it is necessary to take into account explicitly that the bundle has finite size and to show that convergence is retained under such a hypothesis. A possible way to tackle the problem is to introduce an aggregation technique scheme of the type devised by Kiwiel [12] and widely used in bundle methods [10]. In particular let $\hat x$ be the point generated at step 3 of the "main iteration," obtained by solving $QP(\hat\gamma)$ or $DP(\hat\gamma)$. If we define the aggregate quantities

$$g_+ \triangleq \frac{G_+\lambda_{\hat\gamma}}{e^T\lambda_{\hat\gamma}}, \qquad \alpha^+ \triangleq \frac{\alpha_+^T\lambda_{\hat\gamma}}{e^T\lambda_{\hat\gamma}}$$

and, in case $\mu_{\hat\gamma} \ne 0$,

$$g_- \triangleq \frac{G_-\mu_{\hat\gamma}}{e^T\mu_{\hat\gamma}}, \qquad \alpha^- \triangleq \frac{\alpha_-^T\mu_{\hat\gamma}}{e^T\mu_{\hat\gamma}},$$

it is easy to verify that the aggregate problem

$$QP^a(\hat{\gamma}) \qquad \begin{cases} \min\limits_{v,d} & \hat{\gamma}v + \dfrac{1}{2}\|d\|^2 \\ & v \geq g_+^T d - \alpha^+, \\ & v \geq g_i^T d - \alpha_i, \quad i \in \bar{I}_+, \\ & v \leq g_-^T d - \alpha^-, \\ & v \leq g_i^T d - \alpha_i, \quad i \in \bar{I}_-, \end{cases}$$

has the same optimal solution $(v_{\hat{\gamma}}, d_{\hat{\gamma}})$ as $QP(\hat{\gamma})$, where $\bar{I}_+$ and $\bar{I}_-$ are arbitrary subsets of I_+ and I_-, respectively. Of course, in case $I_- = \emptyset$ or $\mu_{\hat{\gamma}} = 0$, the formulation of the aggregate problem does not contain the constraint $v \leq g_-^T d - \alpha^-$ and $(v_{\hat{\gamma}}, d_{\hat{\gamma}})$ is still optimal.

On the basis of the above observations it is possible to embed an aggregation scheme into the algorithm. Suppose that at a certain execution of the "main iteration," the quadratic program $QP(\hat{\gamma})$ (or $DP(\hat{\gamma})$) is solved, and the corresponding optimal dual vector $(\lambda_{\hat{\gamma}}, \mu_{\hat{\gamma}})$ is calculated. Then, once the quantities g_+, α^+, g_-, α^- have been calculated as well, it is possible to construct the aggregate problem $QP^a(\hat{\gamma})$ by inserting the aggregated constraints into $QP(\hat{\gamma})$ and deleting part of its bundle elements. Thus, next time the quadratic program must be solved, it can be obtained by inserting the new constraint, corresponding to the new bundle element calculated at step 3 of the "main iteration," into the aggregated problem $QP^a(\hat{\gamma})$. Of course, such an aggregation task will only be carried out each time a given maximal bundle dimension is reached.

The convergence of the algorithm is not affected by the aggregation mechanism. Indeed the key argument is that the monotonicity of the sequence $\{z_{\hat{\gamma}}^{(k)}\}$, necessary in the proof of Lemma 4.2, is still guaranteed.

The algorithm, encompassing the aggregation scheme, has been implemented in double precision Fortran-77 under a Windows ME system. The code, called NCVX, has been tested on a set [17] of 25 problems available on the web at the URL http://www.cs.cas.cz/~luksan/test.html. All test problems, except the Rosenbrock problem, are nonsmooth.

We have not implemented the construction of the proximal trajectory at step 1 of the "main iteration," and we have always set $\hat{\gamma} = 10\gamma_{min}$. Each test has returned the same number of function evaluations as the number of subgradient evaluations. In fact the condition at step 1 of the algorithm has always been satisfied by the initial choice of $\hat{\gamma}$ and step 4(c) has never been entered.

The input parameters have been set as follows: $\epsilon = 0.1$, $\delta = 10^{-4}$, $m = 0.2$, $\rho = 0.5$, $r = 0.5$, $R = 10^3$. In Table 5.1 we report the computational results in terms of N_f function evaluations. By f^* and f we indicate the minimum value of the objective function and the function value reached by the algorithm when the stopping criterion is met, respectively.

At each iteration we solve the dual program $DP(\gamma)$ by using the subroutine DQPROG provided by the IMSL library and based on M. J. D. Powell's implementation of the Goldfarb and Idnani [8] dual quadratic programming algorithm.

In testing the algorithm, we have always adopted the same set of input parameters,

TABLE 5.1
NCVX: Computational results.

Problem				NCVX	
#	Problem	n	f^*	N_f	f
1	Rosenbrock	2	0	70	5.009e-07
2	Crescent	2	0	22	8.022e-06
3	CB2	2	1.9522245	18	1.9522245
4	CB3	2	2	15	2.0000001
5	DEM	2	-3	21	-2.9999999
6	QL	2	7.2	28	7.2000005
7	LQ	2	-1.4142136	9	-1.4142135
8	Mifflin1	2	-1	127	-0.9999977
9	Mifflin2	2	-1	13	-1.0000000
10	Rosen–Suzuki	4	-44	29	-44.000000
11	Shor	5	22.600162	44	22.600162
12	Maxquad	10	-0.8414083	56	-0.8414078
13	Maxq	20	0	293	1.660e-07
14	Maxl	20	0	44	1.110e-15
15	Goffin	50	0	148	1.142e-13
16	El-Attar	6	0.5598131	152	0.5598163
17	Wolfe	2	-8	21	-7.9999998
18	MXHILB	50	0	33	1.768e-05
19	L1HILB	50	0	104	6.978e-07
20	Colville1	5	-32.348679	47	-32.348679
21	Gill	10	9.7857721	164	9.7857746
22	HS78	5	-2.9197004	159	-2.9196589
23*	TR48	48	-638565	353	-638565.00
24	Shell Dual	15	32.348679	1497	32.349404
25	Steiner2	12	16.703838	196	16.703838

with no tuning based on any specific test problem, aiming at checking algorithm robustness more than efficiency. For problem 23 (marked by "$*$" in Table 5.1) we have set $m = 0.8$, as with standard $m = 0.2$ the quadratic subprogram solver failed due to the accumulation of rounding errors.

REFERENCES

[1] E. W. CHENEY AND A. A. GOLDSTEIN, *Newton's method for convex programming and Tchebycheff approximation*, Numer. Math., 1 (1959), pp. 253–268.

[2] F. CLARKE, *Optimization and Nonsmooth Analysis*, John Wiley and Sons, New York, 1983.

[3] V. F. DEMYANOV AND A. RUBINOV, *Quasidifferential Calculus*, Optimization Software Inc., New York, 1986.

[4] V. F. DEMYANOV AND A. RUBINOV, *Constructive Nonsmooth Analysis*, Peter Lang, Frankfurt am Main, Germany, 1995.

[5] G. DI PILLO, L. GRIPPO, AND S. LUCIDI, *A smooth method for the finite minimax problem*, Math. Program., 60 (1993), pp. 187–214.

[6] A. FUDULI AND M. GAUDIOSO, *The Proximal Trajectory Algorithm for Convex Minimization*, Tech. Report 7/98, Laboratorio di Logistica, Dipartimento di Elettronica Informatica e Sistemistica, Università della Calabria, Italy, 1998.

[7] M. GAUDIOSO, *Nonsmooth optimization*, in Handbook of Applied Optimization, M. G. C. Resende and P. Pardalos, eds., Oxford University Press, New York, 2002, pp. 299–310.

[8] D. GOLDFARB AND A. IDNANI, *A numerically stable dual method for solving strictly convex quadratic program*, Math. Program., 27 (1983), pp. 1–33.

[9] J.-B. HIRIART-URRUTY AND C. LEMARÉCHAL, *Convex Analysis and Minimization Algorithms. Vol.* I, Springer-Verlag, Berlin, 1993.

[10] J.-B. HIRIART-URRUTY AND C. LEMARÉCHAL, *Convex Analysis and Minimization Algorithms. Vol.* II, Springer-Verlag, Berlin, 1993.

[11] J. E. KELLEY, JR., *The cutting-plane method for solving convex programs*, J. Soc. Indust. Appl.

Math., 8 (1960), pp. 703–712.
[12] K. C. Kiwiel, *An aggregate subgradient method for nonsmooth convex minimization*, Math. Program., 27 (1983), pp. 320–341.
[13] K. C. Kiwiel, *Proximity control in bundle methods for convex nondifferentiable minimization*, Math. Program., 46 (1990), pp. 105–122.
[14] K. C. Kiwiel, *Finding normal solutions in piecewise linear programming*, Appl. Math. Optim., 32 (1995), pp. 235–254.
[15] K. C. Kiwiel, *Restricted step and Levenberg–Marquardt techniques in proximal bundle methods for nonconvex nondifferentiable optimization*, SIAM J. Optim., 6 (1996), pp. 227–249.
[16] C. Lemaréchal, *A view of line-searches*, in Optimization and Optimal Control, Lecture Notes in Control and Inform. Sci. 30, A. Auslender, W. Oettli, and J. Stoer, eds., Springer-Verlag, Berlin, New York, 1981, pp. 59–78.
[17] L. Lukšan and J. Vlček, *Test Problems for Nonsmooth Unconstrained and Linearly Constrained Optimization*, Tech. Report 798, Institute of Computer Science, Academy of Sciences of the Czech Republic, Prague, 2000.
[18] M. Mäkelä, *Survey of bundle methods for nonsmooth optimization*, Optim. Methods Softw., 17 (2002), pp. 1–29.
[19] M. Mäkelä and P. Neittaanmäki, *Nonsmooth Optimization*, World Scientific, River Edge, NJ, 1992.
[20] R. Mifflin, *An algorithm for constrained optimization with semismooth functions*, Math. Oper. Res., 2 (1977), pp. 191–207.
[21] E. Polak, D. Mayne, and J. Higgins, *A superlinear convergent algorithm for min-max problems*, in Proceedings of the 28th IEEE Conference on Decision and Control, Tampa, FL, 1989, pp. 894–898.
[22] R. T. Rockafellar, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.
[23] H. Schramm and J. Zowe, *A version of the bundle idea for minimizing a nonsmooth function: Conceptual idea, convergence analysis, numerical results*, SIAM J. Optim., 2 (1992), pp. 121–152.
[24] N. Shor, *Minimizations Methods for Nondifferentiable Functions*, Springer-Verlag, Berlin, 1985.

SIAM J. OPTIM.
Vol. 14, No. 3, pp. 757–772

METRIC REGULARITY AND CONSTRAINT QUALIFICATIONS FOR CONVEX INEQUALITIES ON BANACH SPACES*

XI YIN ZHENG† AND KUNG FU NG‡

Abstract. We introduce new notions of the extended basic constraint qualification and the strong basic constraint qualification and discuss their relationship with other fundamental concepts such as the basic constraint qualification and the metric regularity; in particular we provide a solution to an open problem of Lewis and Pang on characterizing the metric regularity in terms of normal cones. We present a characterization of error bounds for convex inequalities in terms of the strong basic constraint qualification. As applications, we study the linear regularity for infinite collections of closed convex sets in a Banach space.

Key words. metric regularity, basic constraint qualification, strong basic constraint qualification, error bound, linear regularity, infinite system of convex inequalities

AMS subject classifications. 90C31, 90C25, 49J52

DOI. 10.1137/S1052623403423102

1. Introduction.

Let X be a Banach space and $\phi : X \to R\bigcup\{+\infty\}$ a proper lower semicontinuous convex function, and let us consider the convex inequality

$$\phi(x) \leq 0. \tag{1.1}$$

Let S denote the solution set of (1.1), that is,

$$S := \{x \in X : \phi(x) \leq 0\}.$$

We always assume $S \neq \emptyset$. Let $x_0 \in \partial S$, the topological boundary of S. Recall that (1.1) is said to be metrically regular at x_0 if there exist $\tau, \delta \in (0, +\infty)$ such that

$$dist(x, S) \leq \tau[\phi(x)]_+ \quad \forall x \in B(x_0, \delta), \tag{1.2}$$

where $B(x_0, \delta)$ denotes the open ball with center x_0 and radius δ. In this case, we also say that (1.1) is τ-metrically regular at x_0.

For a closed convex subset K of X and $a \in K$, let $N_K(a)$ denote the normal cone of K at a, that is,

$$N_K(a) := \{x^* \in X^* : \langle x^*, x - a\rangle \leq 0 \ \forall\, x \in K\}.$$

Let $\mathrm{dom}(\phi) := \{x \in X : \phi(x) < +\infty\}$ and $\mathrm{epi}(\phi) := \{(x, t) \in X \times R : \phi(x) \leq t\}$. Recall that the subdifferential and singular subdifferential of ϕ at $x \in \mathrm{dom}(\phi)$ are, respectively, the sets

$$\partial\phi(x) := \{x^* \in X^* : (x^*, -1) \in N_{\mathrm{epi}(\phi)}(x, \phi(x))\}$$

*Received by the editors February 14, 2003; accepted for publication (in revised form) August 12, 2003; published electronically January 30, 2004. This research was supported by an earmarked grant from the Research Grant Council of Hong Kong.

http://www.siam.org/journals/siopt/14-3/42310.html

†Department of Mathematics, Yunnan University, Kunming 650091, P. R. China (xyzheng@ynu.edu.cn). The research of this author was supported by PGS of the Chinese University of Hong Kong, the National Natural Science Foundation of P. R. China (grant 10361008), and the National Science Foundation of Yunnan Province, China (grant 2003A002M).

‡Department of Mathematics, The Chinese University of Hong Kong, Shatin, New Territory, Hong Kong (kfng@math.cuhk.edu.hk).

and

$$\partial^\infty \phi(x) := \{x^* \in X^* : \ (x^*, 0) \in N_{\mathrm{epi}(\phi)}(x, \phi(x))\}.$$

It is well known and easy to verify that

$$\partial \phi(x) = \{x^* \in X^* : \langle x^*, u - x\rangle \le \phi(u) - \phi(x) \ \ \forall\ u \in X\}$$

and

$$\partial^\infty \phi(x) = N_{\mathrm{dom}(\phi)}(x) \ \text{ and } \ \partial \phi(x) = \partial^\infty \phi(x) + \partial \phi(x). \tag{1.3}$$

Let R_+ denote the set of all nonnegative real numbers. In the case when ϕ is a continuous convex function on X, recall (cf. [11, 6]) that (1.1) satisfies the basic constraint qualification (BCQ) at x_0 if

$$N_S(x_0) = R_+ \partial \phi(x_0).$$

Incorporating the singular subdifferential, we can extend this property to the case when ϕ is a proper lower semicontinuous convex function.

For a nonempty interval P in R^+ and a subset A of X^*, we use PA to denote the set $\{ta^* : \ t \in P, a^* \in A\}$ if $A \neq \emptyset$ and adopt the convention that

$$PA = \{0\} \ \text{ if } A = \emptyset. \tag{1.4}$$

We say that (1.1) satisfies the extended BCQ at x_0 if

$$N_S(x_0) = \partial^\infty \phi(x_0) + R_+ \partial \phi(x_0).$$

Therefore, the extended BCQ is reduced to BCQ when ϕ is a continuous convex function.

We say that (1.1) satisfies the strong BCQ at x_0 if there exists $\tau \in (0, \ +\infty)$ such that

$$N_S(x_0) \cap B_{X^*} \subset \partial^\infty \phi(x_0) + [0, \ \tau] \partial \phi(x_0), \tag{1.5}$$

where B_{X^*} denotes the closed unit ball of X^*. When (1.5) holds, we also say that (1.1) satisfies the τ-strong BCQ.

By definition and (1.3) as well as recalling the convention (1.4), it is clear that, in the case when $\partial \phi(x_0) = \emptyset$,

$$\text{extended BCQ at } x_0 \Leftrightarrow \tau\text{-strong } BCQ \text{ at } x_0 \Leftrightarrow N_S(x_0) = \partial^\infty \phi(x_0). \tag{1.6}$$

When ϕ is a continuous convex function, by definition and noting that $\partial^\infty \phi(x_0) = \{0\}$, one has that

$$\text{strong BCQ at } x_0 \Longrightarrow \text{extended BCQ at } x_0 = \text{BCQ at } x_0,$$

but the converse of the first implication is not true (cf. Examples 1 and 2 in section 2). In the case when $X = R^n$ and $\phi(x) = \max\{\psi(x), dist(x, C)\}$, where ψ is a proper lower semicontinuous convex function and C is a closed convex subset of X, Lewis and Pang [9] proved that if $x_0 \in \partial S$ satisfies $\phi(x_0) = 0$ and $\partial \psi(x_0) \neq \emptyset$, then

$$(1.1) \text{ is metrically regular at } x_0 \Rightarrow N_S(x_0) = \overline{N_C(x_0) + R_+ \partial \phi(x_0)}, \tag{1.7}$$

where $\overline{A}$ denotes the closure of A. They raised an open problem: find a useful converse of (1.7) (characterize the metric regularity via the normal cone identity).

In section 2, we study the metric regularity, the BCQ, and the strong BCQ at a fixed point of ∂S; in particular we present an answer to the above problem of Lewis and Pang. Moreover, we give a characterization of the existence of error bounds for (1.1) in terms of the strong BCQ at each point of an appropriate subset of ∂S. For some other types of constraint qualifications, see [8].

Let I be an arbitrary (but nonempty) index set and $\{C_i\}_{i\in I}$ be a collection of closed convex subsets of X. Throughout we assume that $C := \bigcap_{i\in I} C_i$ is nonempty. Let $p \in [1,\ +\infty)$. We say that the collection $\{C_i\}_{i\in I}$ is p-linearly regular if there exists $\tau \in (0,\ +\infty)$ such that

$$dist(x, C) \le \tau \left(\sum_{i\in I} (dist(x, C_i))^p \right)^{\frac{1}{p}} \quad \forall x \in X. \tag{1.8}$$

We say that the collection $\{C_i\}_{i\in I}$ is boundedly p-linearly regular if there exist $\tau, \delta \in (0,\ +\infty)$ such that

$$dist(x, C) \le \tau \left(\sum_{i\in I} (dist(x, C_i))^p \right)^{\frac{1}{p}} \quad \forall x \in X \text{ with } \|x\| \le \delta. \tag{1.9}$$

The notions of the linear regularity and the bounded linear regularity for finite collections of closed convex sets have been studied by many authors (see [1, 2] and references therein). When $X = R^n$, the index set I is finite, and each C_i is a closed cone, in terms of Jameson's property (G), Bauschke, Borwein, and Li [1] presented a characterization of the linear regularity. Recently Ng and Yang [13] extended the result of Bauschke et al. [1, 2] to a finite collection of closed convex sets in a Banach space. In section 3, we consider infinite collections of closed convex sets on a Banach space. We introduce a kind of weak* p-sum for infinitely many closed convex sets in dual spaces. Using this new notion of sums, we generalize Jameson's property (G) to an infinite collection of closed convex cones of a Banach space. In terms of tangent cones and the property (G) we establish characterizations for the infinite collection $\{C_i\}_{i\in I}$ to be linearly regular. Moreover, we present some characterizations of the existence of error bounds for infinite systems of convex inequalities.

2. Metric regularity, extended BCQ and strong BCQ. Recalling [17], we say that a subset A of a closed convex subset K of X has property (R) if it recessionally generates K:

$$K = A + K^\infty,$$

where K^∞ denotes the recession cone of K, that is,

$$K^\infty := \{x \in X :\ K + R_+ x \subset K\}.$$

Trivially (but importantly), we have two examples of subsets of K having property (R): (a) K itself and (b) $\{0\}$, provided that K is a cone.

Throughout, X denotes a Banach space and ϕ denotes a proper lower semicontinuous convex function (unless stated otherwise).

PROPOSITION 2.1. *Let $\tau > 0$ and A be a subset of the solution set S of* (1.1) *with the property (R). Suppose that $\partial S \subset \phi^{-1}(0)$. Then* (1.1) *satisfies the extended*

BCQ (resp., the τ-strong BCQ) at each point of $A \cap \partial S$ if and only if (1.1) *satisfies the extended BCQ (resp., the τ-strong BCQ) at each point of ∂S.*

Proof. Suppose that (1.1) satisfies the extended BCQ (resp., the τ-strong BCQ) at each point of $A \cap \partial S$. Let $z \in \partial S$ and $x^* \in B_{X^*} \cap N_S(z)$ with $x^* \neq 0$. To prove the proposition, we need only show that

$$x^* \in \partial^\infty \phi(z) + R_+ \partial\phi(z) \ \ (\text{resp., } x^* \in \partial^\infty \phi(z) + [0,\ \tau]\partial\phi(z)). \tag{2.1}$$

Let $a \in A$ and $c \in S^\infty$ be such that $z = a + c$. Then $\langle x^*, a + c\rangle = \max_{u\in S}\langle x^*, u\rangle$. It follows from $a + R^+c \subset S$ that $\langle x^*, c\rangle = 0$. Thus $\langle x^*, a\rangle = \max_{u\in S}\langle x^*, u\rangle$. Hence $a \in \partial S$ and $x^* \in N_S(a)$. It follows that

$$x^* \in \partial^\infty \phi(a) + R_+ \partial\phi(a) \ (\text{resp., } x^* \in \partial\phi(a) + [0,\ \tau]\partial\phi(a)).$$

In the case when $\partial\phi(a) = \emptyset$, by (1.4) and (1.3), one has that $x^* \in \partial^\infty\phi(a) = N_{\mathrm{dom}(\phi)}(a)$; noting that $\langle x^*, a\rangle = \langle x^*, z\rangle$, it follows that $x^* \in N_{\mathrm{dom}(\phi)}(z) = \partial^\infty\phi(z)$. Hence (2.1) holds. It remains to consider the case that $\partial\phi(a) \neq \emptyset$. In this case, there exist $x_1^* \in \partial^\infty\phi(a)$, $x_2^* \in \partial\phi(a)$, and $t \in R+$ (resp., $t \in [0,\ \tau]$) such that $x^* = x_1^* + tx_2^*$. If $t = 0$, then $x^* \in \partial^\infty\phi(a) = N_{\mathrm{dom}(\phi)}(a)$ and hence $x^* \in N_{\mathrm{dom}(\phi)}(z) = \partial^\infty\phi(z)$, verifying (2.1). If $t > 0$, by (1.3) one has that $\frac{x^*}{t} \in \partial\phi(a)$, and so, for any $x \in X$,

$$\left\langle \frac{x^*}{t}, x - (a + c)\right\rangle = \left\langle \frac{x^*}{t}, x - a\right\rangle \leq \phi(x) - \phi(a) = \phi(x) - \phi(a + c),$$

thanks to the assumption $\partial S \subset \phi^{-1}(0)$. This shows that $x^* \in R^+\partial\phi(a + c)$ (resp., $x^* \in [0,\ \tau]\partial\phi(a + c)$). The proof is completed. □

Let $f_1, f_2 : X \to R \cup \{+\infty\}$ be proper lower semicontinuous convex functions and let $f_0(x) := \max\{f_1(x), f_2(x)\}$ for all $x \in X$. It is known (cf. [16, Theorem 2]) that

$$\partial f_0(a) = \mathrm{co}(\partial f_1(a) \cup \partial f_2(a)) + \partial^\infty f_2(a) \tag{2.2}$$

provided that f_1 is continuous at a (i.e., $a \in \mathrm{int}[\mathrm{dom}(f_1)]$) and $f_1(a) = f_2(a)$.

With the help of (2.2), we can prove the following characterization for the metric regularity in terms of normal cones.

THEOREM 2.2. *Let $z \in \partial S \subset \phi^{-1}(0)$ and $\tau > 0$. Then the following statements are equivalent.*

(i) (1.1) *is τ-metrically regular at z.*

(ii) *There exists $\delta > 0$ such that* (1.1) *satisfies the τ-strong BCQ at each point of $B(z, \delta) \cap \partial S$.*

Proof. (i)$\Longrightarrow$(ii). Suppose that there exists $r > 0$ such that

$$dist(x, S) \leq \tau[\phi(x)]_+ \quad \forall x \in B(z, r). \tag{2.3}$$

Take $\delta = \frac{r}{2}$, $a \in B(z, \delta) \cap \partial S$, and $x^* \in B_{X^*} \cap N_S(a)$. Noting that

$$B_{X^*} \cap N_S(a) = \partial dist(\cdot, S)(a) \ \text{ and } \ dist(a, S) = 0,$$

one has that

$$\langle x^*, x - a\rangle \leq dist(x, S) \ \ \forall\ x \in X.$$

It follows from (2.3) and $B(a, \delta) \subset B(z, r)$ that

$$\langle x^*, x - a\rangle \leq \tau[\phi(x)]_+ = \tau \max\{\phi(x), 0\} \quad \forall x \in B(a, \delta).$$

By $\phi(a) = 0$ (because $a \in \partial S \subset \phi^{-1}(0)$) and (2.2), one has that

$$x^* \in \tau[\mathrm{co}(\partial\phi(a) \cup \{0\}) + \partial^\infty\phi(a)] = [0,\ \tau]\partial\phi(a) + \partial^\infty\phi(a).$$

This shows that $B^*_X \cap N_S(a) \subset [0,\ \tau]\partial\phi(a) + \partial^\infty\phi(a)$ and so (1.1) satisfies the τ-strong BCQ at a.

(ii)$\Longrightarrow$(i). Suppose that there exists $\delta > 0$ such that

$$B_{X^*} \cap N_S(u) \subset \partial^\infty\phi(u) + [0,\ \tau]\partial\phi(u) \quad \forall u \in B(z,\delta) \cap \partial S. \tag{2.4}$$

Let $x \in B(z, \frac{\delta}{2}) \setminus S$ with $x \in \mathrm{dom}(\phi)$. It suffices to show that

$$dist(x, S) \le \tau\phi(x). \tag{2.5}$$

Noting that $dist(x, S) \le \|x - z\| < \frac{\delta}{2}$, pick $\gamma \in (0,\ 1)$ with $dist(x, S) < \frac{\gamma\delta}{2}$. By [12, Lemma 1.1 and Proposition 1.3], there exist $a \in \partial S$ and $x^* \in B_{X^*} \cap N_S(a)$ such that

$$\gamma\|x - a\| \le \langle x^*, x - a\rangle \tag{2.6}$$

and $\gamma\|x - a\| \le dist(x, S)$. It follows from $dist(x, S) < \frac{\gamma\delta}{2}$ that $\|x - a\| < \frac{\delta}{2}$. This and $x \in B(z, \frac{\delta}{2})$ imply that $a \in B(z,\delta)$. By (2.4) one has that

$$x^* \in \partial^\infty\phi(a) + [0,\ \tau]\partial\phi(a).$$

On the other hand, by (2.6) one has that $0 < \langle x^*, x - a\rangle$, and hence, by $x \in \mathrm{dom}(\phi)$, $x^* \notin N_{\mathrm{dom}(\phi)}(a) = \partial^\infty\phi(a)$. It follows that there exist $x^*_1 \in \partial^\infty\phi(a)$, $x^*_2 \in \partial\phi(a)$, and $t \in (0,\ \tau]$ such that $x^* = x^*_1 + tx^*_2$. This and (1.3) imply that $\frac{x^*}{t} \in \partial\phi(a)$. By the assumption $a \in \partial S \subset \phi^{-1}(0)$, it follows that $\langle \frac{x^*}{t}, x - a\rangle \le \phi(x)$. Hence $\langle x^*, x - a\rangle \le \tau\phi(x)$. This and (2.6) show that $\gamma\|x - a\| \le \tau\phi(x)$. By $dist(x, S) \le \|x - a\|$, letting $\gamma \to 1^-$, one has that (2.5) holds. □

Remark. When $X = R^n$, given a proper lower semicontinuous convex function ψ on R^n and a closed convex subset C of R^n, Lewis and Pang [9, Proposition 2] proved that for $z \in C$ with $\psi(z) = 0$ and $\partial\psi(z) \ne \emptyset$, the implication $(\alpha)\Longrightarrow(\beta)$ holds, where

(α) There exist $r, \tau \in (0,\ +\infty)$ such that

$$dist(x, C \cap \psi^{-1}(-\infty,\ 0]) \le \tau \max\{\psi(x), dist(x, C)\} \quad \forall x \in B(z, r).$$

(β) $N_{C\cap\psi^{-1}(-\infty,\ 0]}(z) = \overline{N_C(z) + R_+\partial\psi(z)}$.

Let $S := C \cap \psi^{-1}(-\infty,\ 0]$. Noting that the inclusion $N_S(z) \supset N_C(z) + R_+\partial\psi(z)$ is always true, (β) can be rewritten as

$$N_S(z) \subset \overline{N_C(z) + R_+\partial\psi(z)}.$$

Letting $\phi(\cdot) = \max\{\psi(\cdot), dist(\cdot, C)\}$, ($\alpha$) is equivalent to (i) of Theorem 2.2. Moreover, by (2.2) one has

$$\begin{aligned}\partial\phi(z) &= \mathrm{co}((B_{X^*} \cap N_C(z)) \cup \partial\psi(z)) + \partial^\infty\psi(z)\\ &\subset B_{X^*} \cap N_C(z) + [0,\ 1]\partial\psi(z) + \partial^\infty\psi(z).\end{aligned}$$

We claim that

$$[0,\ 1]\partial\psi(z) + \partial^\infty\psi(z) = \overline{[0,\ 1]\partial\psi(z)}. \tag{2.7}$$

Granting this and noting that $\partial^\infty\phi(z) = N_{\mathrm{dom}(\phi)}(z) = N_{\mathrm{dom}(\psi)}(z) = \partial^\infty\psi(z)$ (because of (1.3)),

$$[0,\ \tau]\partial\phi(z) + \partial^\infty\phi(z) \subset \overline{\tau B_{X^*} \cap N_C(z) + [0,\ \tau]\partial\psi(z)}.$$

Thus, our implication (i)$\Longrightarrow$(ii) provides a conclusion (stronger than (β)):

$$B_{X^*} \cap N_S(z) \subset \overline{\tau B_{X^*} \cap N_C(z) + [0,\ \tau]\partial\psi(z)},$$

if (α) holds. Next we prove (2.7). Let $x^* \in \overline{[0,\ 1]\partial\psi(z)}$ with $x^* \neq 0$. Then there exist $t_n \in (0,\ 1]$ and $x_n^* \in \partial\psi(z)$ such that $t_n x_n^* \to x^*$. Without loss of generality we can assume $t_n \to t_0 \in [0,\ 1]$. Noting that $(t_n x_n^*, -t_n) \in N_{\mathrm{epi}(\psi)}(z, \psi(z))$ (because $x_n^* \in \partial\psi(z)$), one has that $(x^*, -t_0) \in N_{\mathrm{epi}(\psi)}(z, \psi(z))$. It follows that $x^* \in [0,\ 1]\partial\psi(z) + \partial^\infty\psi(z)$. Hence

$$[0,\ 1]\partial\psi(z) + \partial^\infty\psi(z) \supset \overline{[0,\ 1]\partial\psi(z)}.$$

It remains to show that

$$[0,\ 1]\partial\psi(z) + \partial^\infty\psi(z) \subset \overline{[0,\ 1]\partial\psi(z)}. \tag{2.8}$$

Since $\partial^\infty\psi(z)$ is a cone, by (1.3) one has

$$(0,\ 1]\partial\psi(z) + \partial^\infty\psi(z) = (0,\ 1]\partial\psi(z).$$

Thus, to prove (2.8), we need only show that $\partial^\infty\psi(z) \subset \overline{[0,\ 1]\partial\psi(z)}$. Let $u^* \in \partial^\infty\psi(z)$, and take v^* in $\partial\psi(z)$. Then, by (1.3), $\frac{1}{n}v^* + u^* \in [0,\ 1]\partial\psi(z)$ for any natural number n. Letting $n \to \infty$, $u^* \in \overline{[0,\ 1]\partial\psi(z)}$. This shows that $\partial^\infty\psi(z) \subset \overline{[0,\ 1]\partial\psi(z)}$.

Let z be a fixed point of ∂S. Clearly the following implications hold: the metric regularity at $z\Rightarrow$the strong BCQ at $z\Rightarrow$the extended BCQ at z(= the BCQ at z if $\partial\phi(z) \neq \emptyset$). The following Examples 1 and 2 show that the converse of each of the implications is not valid. Moreover, Example 1 also shows that Theorem 2.3 is not true if the strong BCQ in (ii) is replaced by the BCQ. Thus, these and Theorem 2.3 present a complete answer to an open problem raised by Lewis and Pang in [9].

Let $\{C_1, \ldots, C_n\}$ be a collection of closed convex subsets of X such that $C = \bigcap_{i=1}^n C_i \neq \emptyset$. Recall (cf. [1, 2]) that $\{C_1, \ldots, C_n\}$ is said to have the strong conical hull intersection property (CHIP) if

$$N_C(x) = \sum_{i=1}^n N_{C_i}(x) \ \text{ for each } x \in C.$$

Following [2], suppose $p \in (1,\ +\infty)$ and let $\alpha_p := \frac{1}{p}$, $\beta_p = 1 - \alpha_p$, and ρ_p be the positive solution of $\frac{1}{\rho^2} = {\alpha_p}^{\alpha_p}{\beta_p}^{\beta_p}$. Let

$$S_3 := \{(x, y, z) \in R^3 :\ |y| \le \rho_3 x^{\alpha_3} z^{\beta_3}, x \ge 0, z \ge 0\}$$

and

$$\tilde{S}_2 := \{(x, y, z) \in R^3 :\ |x| \le \rho_2 y^{\alpha_2} z^{\beta_2}, y \ge 0, z \ge 0\}$$

(see [2, Definitions 2.18 and 2.22]). Let K and Y be subsets of R^4, respectively defined by

$$K := (\{0\} \times \tilde{S}_2) + (S_3 \times \{0\}) \ \text{ and } \ Y := \{0\} \times R^3.$$

Bauschke, Borwein, and Tseng [2] proved that the pair $\{K, Y\}$ has the strong CHIP but is not boundedly linearly regular (which corresponds to the case $p = 1$ of the "boundedly p-linearly regular" property defined in section 1); see [2, Theorem 3.1 and Corollary 3.2].

Example 1. Let K and Y be as in the above result proved by Bauschke, Borwein, and Tseng. Let $X = R^4$ and let $\phi(x) := \max\{dist(x, K), dist(x, Y)\}$ for all $x \in X$. Then (1.1) satisfies the BCQ at each point of ∂S but does not satisfy the τ-strong BCQ at 0 for any $\tau \in [0, +\infty)$.

Proof. Let $\tau \in [0, +\infty)$. Noting that K and Y are cones, by the above result of Bauschke, Borwein, and Tseng one has that (1.1) is not τ-metrically regular at 0. Since the solution set $S(= K \cap Y)$ is a cone, $\{0\}$ is a subset of S with the property (R). It follows from Proposition 2.1 and Theorem 2.2 that (1.1) does not satisfy the τ-strong BCQ at 0. On the other hand, since $\{K, Y\}$ has the strong CHIP,

$$N_S(a) = N_K(a) + N_Y(a) \quad \forall a \in S.$$

Fix an arbitrary a in ∂S. Noting that

$$\begin{aligned}\partial\phi(a) &= \mathrm{co}(\partial(dist(\cdot, K))(a) \cup \partial(dist(\cdot, Y))(a)) \\ &= \mathrm{co}((B_{X^*} \cap N_K(a)) \cup (B_{X^*} \cap N_Y(a))) \\ &\supset \frac{1}{2}B_{X^*} \cap N_K(a) + \frac{1}{2}B_{X^*} \cap N_Y(a),\end{aligned}$$

one has

$$R^+\partial\phi(a) \supset N_K(a) + N_Y(a) = N_S(a).$$

Therefore, $R^+\partial\phi(a) = N_S(a)$ (as the inclusion $R^+\partial\phi(a) \subset N_S(a)$ is trivial). This shows that (1.1) satisfies the BCQ at a. The proof is completed. □

Example 2. Take $X = R^2$, $C = \{(u, v) \in R^2 : v \le 1\}$, and $S = \{(u, v) \in R^2 : u^2 + v^2 \le 1\}$. Let

$$\phi(x) = \begin{cases} +\infty, & x \in X \setminus C, \\ dist^2(x, S), & x \in C. \end{cases}$$

Then $S = \{x \in X : \phi(x) \le 0\}$ and $z := (0, 1) \in \partial S$. It is clear that, for any $\tau \in (0, +\infty)$, (1.1) is not τ-metrically regular at z. On the other hand, noting that $\partial\phi(z) = N_S(z)$,

$$B_{X^*} \cap N_S(z) \subset N_S(z) = [0, \tau]N_S(z) = [0, \tau]\partial\phi(z) \text{ for any } \tau > 0$$

(because $N_S(z)$ is a cone). Hence (1.1) satisfies the τ-strong BCQ at z for any $\tau > 0$.

The following theorem provides characterizations for (1.1) to satisfy the τ-strong BCQ at a given point a in ∂S.

THEOREM 2.3. *Let $\tau > 0$ and $a \in \partial S$ with $a \in \mathrm{int}(\mathrm{dom}(\phi))$ (thus ϕ is continuous at a). Then the following statements are equivalent.*

(i) (1.1) *satisfies the τ-strong BCQ at a.*

(ii) $dist(h, T_S(a)) \le \tau[d^+\phi(a)(h)]_+$ *for all $h \in X$, where*

$$d^+\phi(a)(h) = \lim_{t \to 0^+} \frac{\phi(a + th) - \phi(a)}{t}.$$

(iii) $dist(x, a + T_S(a)) \le \tau[\phi(x)]_+$ *for all $x \in X$.*

In order to prove Theorem 2.3, we need the following theorem, which provides a characterization for (1.1) to have a global error bound.

THEOREM 2.4. *Let $\tau > 0$ and A be a convex subset of the solution set S with the property* (R). *Suppose that $\partial S \subset \phi^{-1}(0)$. Then*

$$dist(x, S) \leq \tau[\phi(x)]_+ \ \ \forall\, x \in X$$

if and only if (1.1) *satisfies the τ-strong BCQ at each point of $A \cap \partial S$.*

Proof. In view of Theorem 2.2, we need only to prove the sufficiency part. Let $x \in X \setminus S$ and $\gamma \in (0, 1)$. By [12, Lemma 1.1 and Proposition 1.3], there exists $z \in \partial S$ such that

$$dist(z + t(x - z), S) \geq \gamma t\|x - z\| \ \ \forall t \geq 0. \tag{2.9}$$

By Proposition 2.1 and Theorem 2.2 there exists $\delta \in (0,\ 1)$ such that

$$dist(z + \delta(x - z), S) \leq \tau[\phi(z + \delta(x - z))]_+ \leq \tau(\delta[\phi(x)]_+ + (1 - \delta)[\phi(z)]) = \tau\delta[\phi(x)]_+.$$

It follows from (2.9) and $dist(x, S) \leq \|x - z\|$ that

$$\gamma\delta dist(x, S) \leq \tau\delta[\phi(x)]_+.$$

Letting $\gamma \to 1$, one has that $dist(x, S) \leq \tau[\phi(x)]_+$. The proof is completed. □

Let C be a closed convex cone in X. We adopt the notation for (negative) polar set $C^\circ := \{x^* \in X^* :\ \langle x^*, x\rangle \leq 0 \ \text{ for all } x \in C\}$.

Proof of Theorem 2.3. Since $a \in \text{int}(\text{dom}(\phi))$, $\partial^\infty\phi(a) = \{0\}$ and

$$d^+\phi(a)(h) = \sup\{\langle x^*, h\rangle :\ x^* \in \partial\phi(a)\} \quad \forall h \in X,$$

it follows from $(T_S(a))^\circ = N_S(a) \supset R_+\partial\phi(a)$ that

$$T_S(a) \subset \{h \in X :\ d^+\phi(a)(h) \leq 0\}.$$

Therefore, if (i) or (ii) holds, one has

$$T_S(a) = \{h \in X :\ d^+\phi(a)(h) \leq 0\}.$$

Noting that $\{0\}$ is a subset of $T_S(a)$ with the property (R) as $T_S(a)$ is a cone, it follows from Theorem 2.4 that (ii) is equivalent to

$$B_{X^*} \cap N_{T_S(a)}(0) \subset [0,\ \tau]\partial(d^+\phi(a))(0)$$

(thanks to $\partial^\infty(d^+\phi(a))(0) = \{0\}$). Since $N_{T_S(a)}(0) = N_S(a)$ and $\partial(d^+\phi(a))(0) = \partial\phi(a)$, the equivalence (i)$\Longleftrightarrow$(ii) follows immediately.

(ii)$\Longrightarrow$(iii) is trivial as $d^+\phi(a)(x - a) \leq \phi(x)$ for each x.

(iii)$\Longrightarrow$(ii) Let $h \in X \setminus T_S(a)$. Then, for any $t > 0$, $\phi(a + th) > 0$ and it follows from (iii) that

$$dist(a + th, a + T_S(a)) \leq \tau\phi(a + th),$$

that is,

$$dist(h, T_S(a)) \leq \tau\frac{\phi(a + th) - \phi(a)}{t}.$$

Letting $t \to 0^+$, one has that $dist(h, T_S(a)) \le \tau d^+\phi(a)(h)$. Therefore, (ii) holds. □

We conclude this section with the case when ϕ is the maximum function of finitely many differentiable convex functions.

PROPOSITION 2.5. *Let $f_1, \dots, f_m : X \to R$ be differentiable convex functions. Let $\phi(x) = \max\{f_i(x) : 1 \le i \le n\}$ for all $x \in X$, and let $z \in \partial S$ be fixed. Then* (1.1) *satisfies the BCQ at z if and only if* (1.1) *satisfies the strong BCQ at z.*

Proof. The sufficiency part is trivial. Conversely we suppose that (1.1) satisfies the BCQ at z. Let $I(z) := \{1 \le i \le m : f_i(z) = \phi(z)\}$ and $D := \text{co}(\{f_i'(z) : i \in I(z)\})$. Then $\partial\phi(z) = D$, and hence, by the definition of BCQ,

$$N_S(z) = R_+ D. \tag{2.10}$$

Let P be a convex set generated by finitely many points $\{a_1, \dots, a_n\}$ of X^* (i.e., $P = \text{co}(a_1, \dots, a_n)$) and

$$E(P) := \{x^* \in [0,\ 1]P : tx^* \notin [0,\ 1]P \ \forall\, t > 1\}.$$

We claim that

$$\delta(P) := \inf\{\|x^*\| : x^* \in E(P)\} > 0. \tag{2.11}$$

Granting this, let $x^* \in N_S(z) \cap B_{X^*}$ with $x^* \ne 0$. By (2.10) let $r := \sup\{t > 0 : tx^* \in [0,\ 1]D\}$. Thus $rx^* \in E(D)$ and so $\delta(D) \le \|rx^*\| \le r$. Hence

$$x^* \in \frac{1}{r}E(D) \subset \left[0,\ \frac{1}{\delta(D)}\right] D = \left[0,\ \frac{1}{\delta(D)}\right] \partial\phi(z).$$

This shows that $N_S(z) \cap B_{X^*} \subset [0, \frac{1}{\delta(D)}]\partial\phi(z)$. Therefore (1.1) satisfies the strong BCQ at z. It remains to show that (2.11) holds. We will show this by induction.

(i) It is clear that (2.11) holds when $[0,\ 1]P$ is of dimension 1.

(ii) Suppose that (2.11) holds whenever $[0,\ 1]P$ is of dimension n.

(iii) We will show that (2.11) also holds when $[0,\ 1]P$ is of dimension $n + 1$. Suppose to the contrary that there exists a sequence $\{x_k^*\}$ in $E(P)$ such that $\|x_k^*\| \to 0$. Note that $E(P)$ is contained in the relative boundary of $[0,\ 1]P$. Since $[0,\ 1]P$, as a polyhedron of the finite dimensional space $\text{span}(P)$, has finitely many faces (cf. [15, Theorem 19.1]), by considering a subsequence if necessary we can assume that $\{x_k^*\} \subset \tilde{P}$ for some face $\tilde{P}$ of $[0,\ 1]P$. It follows from $\|x_k^*\| \to 0$ and the closedness of $\tilde{P}$ that $0 \in \tilde{P}$. Thus $[0,\ 1]\tilde{P} = \tilde{P}$ is of dimension n. By (ii), $\inf\{\|x_k^*\| : k = 1, 2, \dots\} > 0$, a contradiction. The proof is completed. □

In the case when $X = R^n$ and ϕ is as in Proposition 2.5, Li [10] proved an interesting result that (1.1) is metrically regular at each point of ∂S if and only if (1.1) satisfies the BCQ at each point of ∂S. In view of Proposition 2.5 and Li's result, it gives rise to a natural problem: under all conditions of Proposition 2.5, is it true that (1.1) is metrically regular at a fixed point z if (1.1) satisfies the BCQ at that point? We don't know the answer even for the case when $X = R^n$.

3. Infinite systems of convex inequalities. Let I be an arbitrary nonempty index set and $(f_i)_{i \in I}$ be a family of proper lower semicontinuous convex functions on a Banach space X. Consider the following infinite system of convex inequalities:

$$f_i(x) \le 0, \quad i \in I. \tag{3.1}$$

In what follows, we always denote by S the solution set of (3.1): $S = \{x \in X : f_i(x) \le 0,\ i \in I\}$. Let $p \in [1,\ +\infty)$. We say that a constant $\tau > 0$ is a p-error bound of (3.1) if

$$dist(x, S) \le \tau \left(\sum_{i \in I} [f_i(x)]_+^p \right)^{\frac{1}{p}} \quad \forall x \in X. \tag{3.2}$$

If $\sum_{i \in I} [f_i(x)]_+^p = +\infty$, then (3.2) holds trivially. This leads us to define the following concept.

We say that (3.1) is of type l^p if $(f_i(x))_{i \in I} \in l^p(I)$ for each $x \in \bigcap_{i \in I} \mathrm{dom}(f_i)$ (the basic properties of the classical Banach space $l^p(I)$ can be found in Day [5]). Let $\{C_i\}_{i \in I}$ be a collection of closed convex subsets of X. For each $i \in I$, define $f_i(x) = dist(x, C_i)$. Then (3.2) holds if and only if the collection $\{C_i\}_{i \in I}$ is p-linearly regular. Many authors (e.g., in [1, 2, 13]) have studied the linear regularity of $\{C_i\}_{i \in I}$ in the case when the index set I is finite. Using the results presented in section 2, we will establish, for general I, some characterizations of the existence of p-error bounds of (3.1). We first establish a few lemmas which are of some independent interest.

LEMMA 3.1. *Let $p \in [1,\ +\infty)$ and suppose that (3.1) is of type l^p. Let $a \in \partial S \cap \mathrm{int}(\bigcap_{i \in I} \mathrm{dom}(f_i))$, $I(a) := \{i \in I :\ f_i(a) = 0\}$, and let $\phi : X \to R$ be defined by*

$$\phi(x) = \left(\sum_{i \in I} [f_i(x)]_+^p \right)^{\frac{1}{p}} \ \forall\ x \in X.$$

Then, for each $h \in X$,

$$d^+\phi(a)(h) = \begin{cases} 0 & \text{if } I(a) = \emptyset, \\ \left(\sum\limits_{i \in I(a)} [d^+ f_i(a)(h)]_+^p \right)^{\frac{1}{p}} & \text{if } I(a) \ne \emptyset. \end{cases}$$

Proof. Let $h \in X$. Write $g_i(x)$ for $[f_i(x)]_+$. Since $a \in \mathrm{int}(\bigcap_{i \in I} \mathrm{dom}(f_i))$, $g_i(a) = 0$ and g_i is continuous at a. It is easy to verify that

$$d^+ g_i(a)(h) = \begin{cases} [d^+ f_i(a)(h)]_+ & \text{if } i \in I(a), \\ 0 & \text{if } i \in I \setminus I(a). \end{cases}$$

Take $\delta > 0$ such that $a + \delta h \in \bigcap_{i \in I} \mathrm{dom}(f_i)$ (because $a \in \mathrm{int}(\bigcap_{i \in I} \mathrm{dom}(f_i))$). By the convexity of g_i, one has that for each $i \in I$,

$$0 \le d^+ g_i(a)(h) \le \frac{g_i(a + th)}{t} \le \frac{g_i(a + \delta h)}{\delta} \le \frac{|f_i(a + \delta h)|}{\delta} \ \ \forall t \in (0,\ \delta].$$

Since $\left(\frac{|f_i(a+\delta h)|}{\delta} \right)_{i \in I} \in l^p(I)$, it follows that

$$\lim_{t \to 0^+} \sum_{i \in I} \left(\frac{g_i(a + th)}{t} \right)^p = \sum_{i \in I} (d^+ g_i(a)(h))^p,$$

and hence

$$\lim_{t \to 0^+} \frac{\phi(a + th)}{t} = \left(\sum_{i \in I} (d^+ g_i(a)(h))_+^p \right)^{\frac{1}{p}} = \begin{cases} 0 & \text{if } I(a) = \emptyset, \\ \left(\sum\limits_{i \in I(a)} [d^+ f_i(a)(h)]_+^p \right)^{\frac{1}{p}} & \text{if } I(a) \ne \emptyset. \end{cases}$$

From $\phi(a) = 0$, the lemma is proved. □

A family $\{x_i^*\}_{i\in I}$ of elements in X^* is said to be weak*-summable if $\sum_{i\in I}\langle x_i^*, h\rangle$ exists in R for each $h \in X$. In this case, one defines $x^* : X \to R$ by $\langle x^*, h\rangle = \sum_{i\in I}\langle x_i^*, h\rangle$. Then by a standard result in Banach space theory, one can show that $x^* \in X^*$, and this x^* will be denoted by

$$x^* = \sum_{i\in I}{}^* x_i^*. \tag{3.3}$$

A collection $(A_i)_{i\in I}$ of subsets of X^* is said to be weak*-summable if $\sum_{i\in I}{}^* x_i^*$ exists (in the sense that (3.3) holds for some $x^* \in X^*$) whenever $\{x_i^*\}_{i\in I} \subset X^*$ is such that $x_i^* \in A_i$ (for all $i \in I$). We write $\sum_{i\in I}{}^* A_i$ for the set $\{\sum_{i\in I}{}^* x_i^* : \ x_i^* \in A_i, \ i \in I\}$ when the collection $(A_i)_{i\in I}$ is weak*-summable. Let

$$l_+^p(I) := \{(t_i) \in l^p(I) : \ t_i \geq 0 \ \forall i \in I\}.$$

If $(t_iA_i)_{i\in I}$ is weak*-summable for each $(t_i)_{i\in I} \in l_+^p(I)$ with $\sum_{i\in I} t_i^p = 1$, we define $\mathrm{Co}_p(A_i)_{i\in I}$ as

$$\mathrm{Co}_p(A_i)_{i\in I} := \bigcup_{(t_i)_{i\in I}\in l_+^p(I),\, \sum_{i\in I} t_i^p = 1} \sum_{i\in I}{}^* t_iA_i.$$

LEMMA 3.2. *Let $p, q \in (1, +\infty)$ with $\frac{1}{p} + \frac{1}{q} = 1$; let ϕ and a be as in Lemma* 3.1. *Then*

$$\partial\phi(a) = \begin{cases} \{0\} & \text{if } I(a) = \emptyset, \\ \mathrm{Co}_q([0, \ 1]\partial f_i(a))_{i\in I(a)} & \text{if } I(a) \neq \emptyset. \end{cases}$$

Proof. If $I(a) = \emptyset$, then $d^+\phi(a)(h) = 0$ for all $h \in X$ by Lemma 3.1. This implies that $\partial\phi(a) = \{0\}$. In what follows, we suppose that $I(a) \neq \emptyset$. Let $g_i(x) = [f_i(x)]_+$ for all $x \in X$. Then, for each $i \in I(a)$,

$$\partial g_i(a) = [0, \ 1]\partial f_i(a) \text{ and } d^+g_i(a)(\cdot) = [d^+f_i(a)(\cdot)]_+$$

(see the proof of Lemma 3.1). Since $a \in \mathrm{int}(\bigcap_{i\in I} \mathrm{dom}(f_i))$, it follows that $[d^+f_i(a)(\cdot)]_+$ is the support functional of $[0, \ 1]\partial f_i(a)$ for each $i \in I(a)$; that is,

$$[d^+f_i(a)]_+(h) = \max\{\langle x^*, h\rangle : \ x^* \in [0, \ 1]\partial f_i(a)\}.$$

Therefore, for any subset I' of $I(a)$, any $(t_i)_{i\in I(a)} \in l_+^q(I(a))$ with $\sum_{i\in I(a)} t_i^q \leq 1$, any $x_i^* \in [0, \ 1]\partial f_i(a)$ $(i \in I(a))$, and any $h \in X$, one has

$$\sum_{i\in I'} t_i\langle x_i^*, h\rangle \leq \sum_{i\in I'} t_i[d^+f_i(a)(h)]_+ \leq \left(\sum_{i\in I'}[d^+f_i(a)(h)]_+^p\right)^{\frac{1}{p}} \leq d^+\phi(a)(h), \tag{3.4}$$

thanks to Lemma 3.1. It follows from a standard result in Banach space theory that $\sum_{i\in I}{}^* t_ix_i^*$ exists in X^*. Thus $\mathrm{Co}_q([0, \ 1]\partial f_i(a))_{i\in I(a)}$ is well defined. Indeed it is now easily verified that

$$\mathrm{Co}_q([0, \ 1]\partial f_i(a))_{i\in I(a)} = \left\{\sum_{i\in I}{}^* t_ia_i^* : \ a_i^* \in \partial f_i(a), \ (t_i)_{i\in I} \in l_+^q(I), \ \sum_{i\in I} t_i^q \leq 1\right\}.$$

In particular $\mathrm{Co}_q([0,\ 1]\partial f_i(a))_{i\in I(a)}$ is convex. Next we will show that it is weak* closed. Let $x^* \in \overline{\mathrm{Co}_q([0,\ 1]\partial f_i(a))_{i\in I(a)}}^{w*}$, where $\overline{A}^{w^*}$ denotes the weak* closure of A. Then there exists a directed set Λ and nets $(t_i(\alpha))_{\alpha\in\Lambda}$, $(x_i^*(\alpha))_{\alpha\in\Lambda}$ $(i \in I(a))$ such that $t_i(\alpha) \geq 0$, $x_i^*(\alpha) \in \partial f_i(a)$, $\sum_{i\in I(a)}(t_i(\alpha))^q \leq 1$, and

$$\lim_\alpha \sum_{i\in I(a)}{}^* t_i(\alpha)x_i^*(\alpha) = x^* \ \text{ with respect to the weak}^* \text{ topology.} \tag{3.5}$$

For each $\alpha \in \Lambda$, let $g_\alpha := (t_i(\alpha))_{i\in I(a)}$. Then $\{g_\alpha\}_{\alpha\in\Lambda}$ is a net in the unit ball of $l^q(I(a))$; hence without loss of generality we can assume that this net weak*-converges to $(\lambda_i)_{i\in I(a)} \in l^q_+(I(a))$ with $\sum_{i\in I(a)} \lambda_i^q \leq 1$. Let $I^+ = \{i \in I(a) :\ \lambda_i > 0\}$. Thus I^+ is at most countable. Noting that $\lim_\alpha t_i(\alpha) = \lambda_i = 0$ for each $i \in I(a) \setminus I^+$, by (3.4) one has that $\lim_\alpha \sum_{i\in I(a)\setminus I^+}{}^* t_i(\alpha)x_i^*(\alpha) = 0$ with respect to the weak* topology. It follows from (3.5) that

$$\lim_\alpha \sum_{i\in I^+}{}^* t_i(\alpha)x_i^*(\alpha) = x^* \ \text{ with respect to the weak}^* \text{ topology.} \tag{3.6}$$

Without loss of generality we can assume I^+ to be the set $\mathbf{N}$ of natural numbers. Since $\partial f_i(a)$ is weak* compact and $\{x_i^*(\alpha)\}_{\alpha\in\Lambda} \subset \partial f_i(a)$ for each i, there exists a subnet $\{x_1^*(\alpha)\}_{\alpha\in\Lambda_1}$ of $\{x_1^*(\alpha)\}_{\alpha\in\Lambda}$ weak*-convergent to $a_1^* \in \partial f_1(a)$, and hence there exists a subnet $\{x_2^*(\alpha)\}_{\alpha\in\Lambda_2}$ of $\{x_1^*(\alpha)\}_{\alpha\in\Lambda_1}$ weak*-convergent to $a_2^* \in \partial f_2(a), \ldots$. Continuing in this way, there exists a subnet $\{x_{i+1}^*(\alpha)\}_{\alpha\in\Lambda_{i+1}}$ of $\{x_i^*(\alpha)\}_{\alpha\in\Lambda_i}$ weak*-convergent to $a_{i+1}^* \in \partial f_{i+1}(a), \ldots$, and so on. Then

$$x^* = \sum_{i\in\mathbf{N}}{}^* \lambda_i a_i^*. \tag{3.7}$$

Indeed, let $h \in X$ and $\varepsilon > 0$. Take $n_0 \in \mathbf{N}$ such that

$$\left(\sum_{i=n_0}^\infty [d^+f_i(a)(\pm h)]^p\right)^{\frac{1}{p}} < \varepsilon.$$

Then, for any $n \geq n_0$, any $(t_i)_{i\in\mathbf{N}} \in l^q_+(\mathbf{N})$ with $\sum_{i\in\mathbf{N}} t_i^q \leq 1$, and any $x_i^* \in \partial f_i(a)$, one has from (3.4) that

$$\left|\sum_{i=n+1}^\infty t_i\langle x_i^*, h\rangle\right| \leq \max\left\{\left(\sum_{i=n+1}^\infty [d^+f_i(a)(h)]_+^p\right)^{\frac{1}{p}},\ \left(\sum_{i=n+1}^\infty [d^+f_i(a)(-h)]_+^p\right)^{\frac{1}{p}}\right\} < \varepsilon.$$

Since $\{t_i(\alpha)\}_{\alpha\in\Lambda_n}$ converges to λ_i and $\{x_i^*(\alpha)\}_{\alpha\in\Lambda_n}$ weak*-converges to a_i^* for $1 \leq i \leq n$, it follows from (3.6) that

$$\left|\sum_{i=1}^n \langle\lambda_i a_i^*, h\rangle - \langle x^*, h\rangle\right| \leq \varepsilon \quad \forall n \geq n_0.$$

This shows that (3.7) holds. For each i, let $r_i = \frac{\lambda_i}{\sum_{j\in\mathbf{N}}\lambda_j^q}$ and $z_i^* = (\sum_{j\in\mathbf{N}}\lambda_j^q)a_i^*$. Thus $\sum_{i\in\mathbf{N}} r_i^q = 1$, $z_i^* \in [0,\ 1]\partial f_i(a)$ and $x^* = \sum_{i\in\mathbf{N}}{}^* r_i z_i^* \in \mathrm{Co}_q([0,\ 1]\partial f_i(a))_{i\in I(a)}$. Thus, we have shown that $\mathrm{Co}_q([0,\ 1]\partial f_i(a))_{i\in I(a)}$ is a weak* closed convex set. In view of

Lemma 3.1 and since $d^+\phi(a)(\cdot)$ is the support functional of the weak* closed convex set $\partial\phi(a)$, to complete the proof it suffices to show that $(\sum_{i\in I(a)}[d^+f_i(a)(\cdot)]_+^p)^{\frac{1}{p}}$ is the support functional of the weak* closed convex set $\mathrm{Co}_q([0,\ 1]\partial f_i(a))_{i\in I(a)}$ (cf. [4, Proposition 2.1.4]). Let $h\in X$. By (3.4), one has that

$$\sup\{\langle x^*,h\rangle :\ x^*\in \mathrm{Co}_q([0,\ 1]\partial f_i(a))_{i\in I(a)}\}\le \left(\sum_{i\in I(a)}[d^+f_i(a)(h)]_+^p\right)^{\frac{1}{p}}. \tag{3.8}$$

On the other hand, since $a\in\bigcap_{i\in I}\mathrm{int}(\mathrm{dom}(f_i))$, for each $i\in I(a)$ there exists $z_i^*\in [0,\ 1]\partial f_i(a)$ such that $\langle z_i^*,h\rangle=[d^+f_i(a)(h)]_+$ (cf. [14, Proposition 2.24]). Noting that $([d^+f_i(a)(h)]_+)_{i\in I(a)}\in l_+^p(I(a))$, there exists $(t_i)_{i\in I(a)}\in l_+^q(I(a))$ with $\sum_{i\in I(a)}t_i^q=1$ such that

$$\left(\sum_{i\in I(a)}[d^+f_i(a)(h)]_+^p\right)^{\frac{1}{p}}=\sum_{i\in I(a)}t_i[d^+f_i(a)(h)]_+=\left\langle \sum_{i\in I(a)}{}^*t_iz_i^*,h\right\rangle.$$

This and (3.8) imply that $(\sum_{i\in I(a)}[d^+f_i(a)(\cdot)]_+^p)^{\frac{1}{p}}$ is the support functional of the set $\mathrm{Co}_q([0,\ 1]\partial f_i(a))_{i\in I(a)}$. The proof is completed. □

For convenience, we adopt the convention to define $\mathrm{Co}_q([0,\ \tau]\partial f_i(a))_{i\in I(a)}$ as $\{0\}$ if $I(a)=\emptyset$. Theorem 3.3 is immediate from Theorems 2.3 and 2.4 and Lemmas 3.1 and 3.2.

THEOREM 3.3. *Let $p,q\in(1,\ +\infty)$ with $\frac{1}{p}+\frac{1}{q}=1$ and suppose that the infinite system* (3.1) *is of type l^p. Let $\tau\ge 0$ and let A be a subset of the solution set S with the property* (R). *Suppose that $A\cap\partial S\subset\mathrm{int}(\bigcap_{i\in I}\mathrm{dom}(f_i))$. Then the following statements are equivalent.*

(i) $dist(x,S)\le\tau(\sum_{i\in I}[f_i(x)]_+^p)^{\frac{1}{p}}\quad\forall x\in X.$

(ii) $dist(h,T_S(a))\le\tau(\sum_{i\in I(a)}[d^+f_i(a)(h)]_+^p)^{\frac{1}{p}}\quad\forall a\in A\cap\partial S\ and\ \forall h\in X.$

(iii) $dist(x,a+T_S(a))\le\tau(\sum_{i\in I(a)}[f_i(x)]_+^p)^{\frac{1}{p}}\quad\forall a\in A\cap\partial S\ and\ \forall x\in X.$

(iv) $B_{X^*}\cap N_S(a)\subset\mathrm{Co}_q([0,\ \tau]\partial f_i(a))_{i\in I(a)}\quad\forall a\in A\cap\partial S.$

Finally, we consider the linear regularity of an infinite collection of closed convex subsets of a Banach space X. We first set some notations. Let $(K_i)_{i\in I}$ be an arbitrary collection of weak* closed subsets of X^* and $p\in[1,\ +\infty)$. If for any $(x_i^*)_{i\in I}$ with $x_i^*\in K_i$ (for all $i\in I$) and $\sum_{i\in I}\|x_i^*\|^p<+\infty$ there exists $x^*\in X^*$ such that $x^*=\sum_{i\in I}{}^*x_i^*$, we define the weak* p-sum of $(K_i)_{i\in I}$ as

$$p\text{-}\sum_{i\in I}{}^*K_i:=\left\{\sum_{i\in I}{}^*x_i^*:\ x_i^*\in K_i\ (\forall i\in I),\ \sum_{i\in I}\|x_i^*\|^p<+\infty\right\}.$$

If I is finite, then the weak* p-sum is the usual sum. Let $p\in(1,\ +\infty)$, $\tau>0$ and suppose that K_i is a weak* closed convex cone in X^* for each $i\in I$.
We say that $\{K_i\}_{i\in I}$ has property $(\mathrm{G},\tau)_p$ if

$$\left(p\text{-}\sum_{i\in I}{}^*K_i\right)\cap B_{X^*}\subset\tau\mathrm{Co}_p(K_i\cap B_{X^*})_{i\in I}.$$

Recall from Jameson [7] and Bauschke, Borwein, and Li [1] that a collection $\{C_1,\dots,C_m\}$ of closed convex cones of a Banach space Z is said to have property (G) if there exists

$\tau > 0$ such that

$$B_Z \cap \sum_{i=1}^{m} C_i \subset \tau \sum_{i=1}^{m} (C_i \cap B_Z).$$

Clearly in the case when the index set I is finite, it is easy to verify that $\{C_i\}_{i\in I}$ has property (G) if and only if $\{C_i\}_{i\in I}$ has property $(\mathrm{G},\tau)_p$ for some $\tau > 0$ and any $p \in (1,\ +\infty)$.

The following proposition provides a characterization of property $(\mathrm{G},\tau)_p$.

PROPOSITION 3.4. *Let X be a Banach space and $(K_i)_{i\in I}$ be an arbitrary collection of* weak* *closed convex cones of X^*. Let $\tau > 0$ and $p \in [1,\ +\infty)$. Suppose that p-$\sum_{i\in I}{}^* K_i$ exists. Then*

$$\left(p\text{-}\sum_{i\in I}{}^* K_i\right) \cap B_{X^*} \subset (0,\ \tau)\mathrm{Co}_p(K_i \cap B_{X^*})_{i\in I} \tag{3.9}$$

if and only if for each x^* in $(p$-$\sum_{i\in I}{}^* K_i) \setminus \{0\}$,

$$\inf\left\{\left(\sum_{i\in I}\|x_i^*\|^p\right)^{\frac{1}{p}} :\ x^* = \sum_{i\in I}{}^* x_i^*,\ x_i^* \in K_i\ (\forall i \in I)\right\} < \tau\|x^*\| \tag{3.10}$$

(by virtue of the Alaoglu theorem, every bounded weak* *closed subset of X^* is* weak* *compact. Hence, if I is finite, the infimum in* (3.10) *is attained).*

Proof. $\Rightarrow$ Let $x^* \in p$-$\sum_{i\in I}{}^* K_i \setminus \{0\}$. By (3.9) there exist $(t_i)_{i\in I} \in l_+^p(I)$ with $\sum_{i\in I} t_i^p \le 1$, $(z_i^*)_{i\in I}$ with $z_i^* \in K_i \cap B_{X^*}$ $(\forall i \in I)$, and $\alpha \in (0,\ \tau)$ such that

$$\frac{x^*}{\|x^*\|} = \alpha \sum_{i\in I}{}^* t_i z_i^*, \quad \text{that is,} \quad x^* = \sum_{i\in I}{}^* \alpha\|x^*\| t_i z_i^*.$$

Therefore,

$$\begin{aligned}\inf\left\{\left(\sum_{i\in I}\|x_i\|^p\right)^{\frac{1}{p}} :\ x^* = \sum_{i\in I}{}^* x_i^*,\ x_i^* \in K_i\ (\forall i \in I)\right\} &\le \left(\sum_{i\in I}(\|\alpha\|x^*\| t_i z_i^*\|)^p\right)^{\frac{1}{p}} \\ &\le \alpha\|x^*\| < \tau\|x^*\|.\end{aligned}$$

This shows that (3.10) holds.

$\Leftarrow$ Let $x^* \in (p$-$\sum_{i\in I}{}^* K_i) \cap B_{X^*}$ with $x^* \ne 0$. By (3.10) there exist $\alpha \in (0,\ \tau)$ and $x_i^* \in K_i$ (for all $i \in I$) such that

$$x^* = \sum_{i\in I}{}^* x_i^* \quad \text{and} \quad \left(\sum_{i\in I}\|x_i^*\|^p\right)^{\frac{1}{p}} < \alpha\|x^*\|.$$

It follows from $\|x^*\| \le 1$ that

$$x^* = \alpha \sum_{i\in I}{}^* \frac{\|x_i^*\|}{\alpha} \cdot \frac{x_i^*}{\|x_i^*\|} \in \alpha\mathrm{Co}_p(K_i \cap B_{X^*})_{i\in I} \subset (0,\ \tau)\mathrm{Co}_p(K_i \cap B_{X^*})_{i\in I}.$$

This shows that (3.9) holds. The proof is completed. ∎

Since each K_i is a cone, it is clear that

$$\text{the property } (\mathrm{G},\tau)_p \Leftrightarrow \left(p\text{-}\sum_{i\in I}{}^{*} K_i\right)\cap B_{X^*} \subset (0,\ \tau]\mathrm{Co}_p(K_i\cap B_{X^*})_{i\in I}.$$

Therefore, (3.9) means that $(K_i)_{i\in I}$ has property $(\mathrm{G},\tau')_p$ for any $\tau'\in(0,\ \tau)$; but we do not know whether or not it implies that $(K_i)_{i\in I}$ has property $(\mathrm{G},\tau)_p$.

THEOREM 3.5. *Let I be an arbitrary nonempty index set and $\{C_i\}_{i\in I}$ be a collection of closed convex subsets of a Banach space X such that $C:=\bigcap_{i\in I}C_i$ is nonempty. Let $\tau>0$, $p,q\in(1,\ +\infty)$ with $\frac{1}{p}+\frac{1}{q}=1$, and let A be a subset of C with property* (R). *Suppose that $(dist(\cdot,C_i))_{i\in I}$ is of type l^p. Then the following statements are equivalent.*

(i) $dist(x,C)\le\tau(\sum_{i\in I}(dist(x,C_i))^p)^{\frac{1}{p}}\quad\forall x\in X.$

(ii) $dist(x,T_C(a))\le\tau(\sum_{i\in I}(dist(x,T_{C_i}(a)))^p)^{\frac{1}{p}}\quad\forall x\in X$ *and* $\forall a\in A\cap\partial C.$

(iii) $dist(x,a+T_C(a))\le\tau(\sum_{i\in I}(dist(x,C_i))^p)^{\frac{1}{p}}\quad\forall x\in X$ *and* $\forall a\in A\cap\partial C.$

(iv) *For each $a\in A\cap\partial C$, $N_C(a)=q\text{-}\sum_{i\in I}{}^{*}N_{C_i}(a)$ and the collection $(N_{C_i}(a))_{i\in I}$ has property $(\mathrm{G},\tau)_q$.*

Proof. For each $i\in I$, let $f_i(x)=dist(x,C_i)$. We apply Theorem 3.3 and note that $S=C$. Recall from [3] that for each $a\in C_i$ and each $h\in X$,

$$d^+f_i(a)(h)=dist(h,T_{C_i}(a)) \text{ and } \partial f_i(a)=B_{X^*}\cap N_{C_i}(a). \tag{3.11}$$

Thus (i), (ii), (iii), and (iv) read as (i), (ii), (iii), and (iv) of Theorem 3.3, respectively (as the inclusion $N_C(a)\supset q\text{-}\sum_{i\in I}{}^{*}N_{C_i}(a)$ is trivial). ∎

COROLLARY 3.6. *Let I be an arbitrary index set and $\{C_i\}_{i\in I}$ be a collection of closed convex subsets of a Banach space X such that $C=\bigcap_{i\in I}C_i$ is a cone, and let $p,q\in(1,\ +\infty)$ with $\frac{1}{p}+\frac{1}{q}=1$ and $\tau>0$. Suppose that $(dist(\cdot,C_i))_{i\in I}$ is of type l^p. Then*

$$dist(x,C)\le\tau\left(\sum_{i\in I}(dist(x,C_i))^p\right)^{\frac{1}{p}}\quad\forall\, x\in X \tag{3.12}$$

if and only if $C^\circ=q\text{-}\sum_{i\in I}{}^{*}C_i^\circ$ and the collection $(C_i^\circ)_{i\in I}$ has property $(\mathrm{G},\tau)_q$.

Proof. Since C is a cone, $A:=\{0\}$ is a subset of C with property (R). Noting that $N_{C_i}(0)=C_i^\circ$ as $0\in C_i$ for each $i\in I$, Corollary 3.6 is immediate from the equivalence of (i) and (iv) in Theorem 3.5. ∎

Corollary 3.6 seems new even when I is finite. Note that when I is finite $(dist(\cdot,C_i))_{i\in I}$ is always of type l^p for each $p\in[1,\ +\infty)$, and Corollary 3.6 remains true if one of $p,\ q$ is $+\infty$ (with suitable interpretation for the right-hand side of (3.12) if $p=+\infty$).

Acknowledgment. We thank the referees for their helpful comments. One of them pointed out a gap in the proof of the early version of Theorem 2.3. His suggestion to incorporate the singular subdifferential helped us redefine strong BCQ.

REFERENCES

[1] H. H. Bauschke, J. M. Borwein, and W. Li, *Strong conical hull intersection property, bounded linear regularity, Jameson's property (G), and error bounds in convex optimization*, Math. Program. Ser. A, 86 (1999), pp. 135–160.

[2] H. H. Bauschke, J. M. Borwein, and P. Tseng, *Bounded linear regularity, strong CHIP, and CHIP are distinct properties*, J. Convex Anal., 7 (2000), pp. 395–412.

[3] J. V. Burke and M. C. Ferris, *On the Clarke subdifferential of distance function of a closed set*, J. Math. Anal. Appl., 166 (1992), pp. 199–213.

[4] F. H. Clarke, *Optimization and Nonsmooth Analysis*, CRM, Montreal, Canada, 1989.

[5] M. M. Day, *Normed Linear Spaces*, Springer-Verlag, Berlin, 1962.

[6] J.-B. Hiriart-Urruty and C. Lemarechal, *Convex Analysis and Minimization Algorithms* I, Springer-Verlag, Berlin, Heidelberg, 1993.

[7] G. J. O. Jameson, *The duality of pairs of wedges*, Proc. London Math. Soc., 24 (1972), pp. 531–547.

[8] D. Klatte and W. Li, *Asymptotic constraint qualifications and error bounds for convex inequalities*, Math. Program. Ser. A, 84 (1999), pp. 137–160.

[9] A. Lewis and J. S. Pang, *Error bounds for convex inequality systems*, in Generalized Convexity, Generalized Monotonicity: Recent Results, Proceedings of the Fifth Symposium on Generalized Convexity, Luminy, 1996, J.-P. Crouzeix, J.-E. Martinez-Legaz, and M. Volle, eds., Kluwer Academic Publishers, Dordrecht, The Netherlands, 1997, pp. 75–110.

[10] W. Li, *Abadie's constraint qualification, metric regularity, and error bounds for differentiable convex inequalities*, SIAM J. Optim., 7 (1997), pp. 966–978.

[11] W. Li, C. Nahak, and I. Singer, *Constraint qualifications for semi-infinite systems of convex inequalities*, SIAM J. Optim., 11 (2000), pp. 31–52.

[12] K. F. Ng and W. H. Yang, *Error bounds for abstract linear systems*, SIAM J. Optim., 13 (2002), pp. 24–43.

[13] K. F. Ng and W. H. Yang, *Regularities and their relations to error bounds*, Math. Program., to appear.

[14] R. R. Phelps, *Convex Functions, Monotone Operators and Differentiability*, Springer-Verlag, Berlin, 1993.

[15] R. T. Rockafellar, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.

[16] M. Volle, *Sous-différentiel d'une enveloppe supérieure de fonctions convexes*, C. R. Acad. Sci. Paris Sér. I Math., 317 (1993), pp. 845–849.

[17] X. Y. Zheng, *Error bounds for set inclusions*, Science in China, Series A, 46 (2003), pp. 750–763.

SIAM J. OPTIM.
Vol. 14, No. 3, pp. 773–782

WEAK CONVERGENCE OF A RELAXED AND INERTIAL HYBRID PROJECTION-PROXIMAL POINT ALGORITHM FOR MAXIMAL MONOTONE OPERATORS IN HILBERT SPACE*

FELIPE ALVAREZ†

Abstract. This paper introduces a general implicit iterative method for finding zeros of a maximal monotone operator in a Hilbert space which unifies three previously studied strategies: relaxation, inertial type extrapolation and projection step. The first two strategies are intended to speed up the convergence of the standard proximal point algorithm, while the third permits one to perform inexact proximal iterations with fixed relative error tolerance. The paper establishes the global convergence of the method for the weak topology under appropriate assumptions on the algorithm parameters.

Key words. Hilbert space, maximal monotone operator, proximal point, inexact iteration, relative error, separating hyperplane, orthogonal projection, relaxation, weak convergence

AMS subject classifications. 90C25, 65K05, 47J25

DOI. 10.1137/S1052623403427859

1. Introduction.

From now on, $(H, \langle\cdot,\cdot\rangle)$ is a real Hilbert space and the set-valued mapping $A : H \rightrightarrows H$ is a *maximal monotone operator*, that is, A is monotone, i.e., $\forall x, y \in H$, $\forall v \in A(x)$, $\forall w \in A(y)$, $\langle v - w, x - y\rangle \geq 0$, and the graph $GrA = \{(x, v) \in H \times H \mid v \in A(x)\}$ is not properly contained in the graph of any other monotone operator. We are interested in the resolution of the inclusion problem

$$\text{Find } x \in H \text{ such that } 0 \in A(x), \tag{1.1}$$

which appears in a wide variety of equilibrium problems such as convex programming and monotone variational inequalities. This article establishes the asymptotic convergence, for the weak topology, of some implicit iterative methods for solving (1.1) under some implementable inexact conditions. These algorithms, which generalize the classical *Proximal Point Algorithm* (PPA), are of inertial type in the sense that they are obtained by discretization of a second-order-in-time dissipative dynamical system.

Recall that PPA, which was proposed in [15, 16] (inspired by [18]), generates a sequence $(x^k) \subset H$ by the successive approximation scheme

$$x^{k+1} = x^k - \lambda_k v^k, \quad v^k \in A(x^{k+1}), \quad k = 0, 1, \ldots,$$

where $(\lambda_k) \subset \mathbb{R}_{++}$ is a sequence of positive regularization parameters. Equivalently,

$$x^{k+1} = J^A_{\lambda_k}(x^k), \tag{PPA}$$

where the single-valued (see [17]) function $J^A_\lambda := (I + \lambda A)^{-1} : H \to H$ is the *resolvent of A of parameter λ*. The resolvent is a nonexpansive mapping and, moreover,

$$J^A_\lambda(x) = x \text{ if and only if } 0 \in A(x). \tag{1.2}$$

*Received by the editors May 12, 2003; accepted for publication (in revised form) September 10, 2003; published electronically January 30, 2004. This work was partially supported by Fondecyt 1020610, ECOS-Conicyt C00E05, and Programa Iniciativa Científica Milenio P01-34.
http://www.siam.org/journals/siopt/14-3/42785.html

†Departamento de Ingeniería Matemática and Centro de Modelamiento Matemático (CNRS UMR 2071), Universidad de Chile, Casilla 170/3, Correo 3, Santiago, Chile (falvarez@dim.uchile.cl).

See [7] for further details. PPA may be viewed as an implicit one-step discretization method for the first-order-in-time differential inclusion $\dot{x}(t) + A(x(t)) \ni 0$, a.e. $t \geq 0$, λ_k being interpreted as a step size parameter. Set $S := A^{-1}(\{0\})$. When $S \neq \emptyset$ and A is demipositive, it is proved in [8] that every solution of this differential inclusion converges weakly in H to a point in S. Concerning PPA, similar convergence results are established in [15, 16] for variational inequalities on bounded sets. The general case is treated in [22], where the equation $x^{k+1} = J^A_{\lambda_k}(x^k)$ is replaced by some inexact criteria, permitting approximate computations of resolvents. See [5] for a counterexample to strong convergence in the continuous case with A being the gradient of a convex function; the same counterexample works for PPA as shown in [13].

To motivate the so-called *Inertial Proximal Point Algorithm* (IPPA), consider the equation for an oscillator with damping and conservative restoring force: $\ddot{x}(t)+\gamma\dot{x}(t)+\nabla f(x(t)) = 0$, where $\gamma > 0$ and $f : H \to \mathbb{R}$ is differentiable. This dynamical system is called *Heavy Ball with Friction* (HBF), and it seems to have been considered for the first time in [21] in the context of optimization problems. The inertial nature of HBF can be exploited in numerical computations in order to accelerate the trajectories and speed up convergence; see [3, 25] for discussions in this direction. Concerning asymptotic convergence, it is proved in [1] that if f is convex (i.e., ∇f is monotone) and $(\nabla f)^{-1}(\{0\}) \neq \emptyset$, then each trajectory of HBF converges weakly in H to some $\hat{x} \in H$ with $\nabla f(\hat{x}) = 0$; see [4] for additional convergence results. Consider the implicit discretization of HBF: $(x^{k+1}-2x^k+x^{k-1})/h^2+\gamma(x^{k+1}-x^k)/h+\nabla f(x^{k+1}) = 0$, which can be rewritten as $x^{k+1} = x^k + \alpha(x^k - x^{k-1}) - \lambda\nabla f(x^{k+1})$, with $\lambda = h^2/(1 + \gamma h)$ and $\alpha = 1/(1 + \gamma h)$. In terms of resolvents, $x^{k+1} = J^{\nabla f}_\lambda(x^k + \alpha(x^k - x^{k-1}))$. Note that λ is no longer a step size but is indeed a regularization parameter that combines the damping factor γ and the actual step size $h > 0$.

Replacing ∇f with a maximal monotone operator A, and considering possibly variable parameters $\lambda_k > 0$ and $\alpha_k \in [0, 1)$, the previous discussion motivates the introduction of the inertial type iteration

$$x^{k+1} = J^A_{\lambda_k}(x^k + \alpha_k(x^k - x^{k-1})), \tag{IPPA}$$

where the extrapolation term $\alpha_k(x^k - x^{k-1})$ is intended to speed up convergence. IPPA was first considered in [1] for a (nonsmooth) conservative operator $A = \partial f$, the subdifferential of a closed, proper, and convex function $f : H \to \mathbb{R} \cup \{\infty\}$; weak convergence toward a minimizer of f holds under suitable conditions (see [1, Thm. 3.1]). For the nonconservative case, a partial positive result for cocoercive operators is proved in [14], where comparisons with first-order-in-time methods are also given through some numerical tests, showing improvements in the speed of convergence.

The case of an arbitrary maximal monotone operator is treated in [2] under the conditions

$$\lambda := \inf_{k\geq 0} \lambda_k > 0, \tag{1.3}$$

$$\forall k \in \mathbb{N}, \alpha_k \in [0, 1) \quad \text{and} \quad \alpha := \sup_{k\geq 0} \alpha_k < 1, \tag{1.4}$$

$$\sum \alpha_k \|x^k - x^{k-1}\|^2 < \infty. \tag{1.5}$$

Since α_k may be chosen once x^{k-1} and x^k have been found, (1.5) is easy to implement in numerical computations. Furthermore, (1.5) holds automatically in some special

situations that can be checked a priori; see, for instance, [1, Thm. 3.1], [2, Prop. 2.1], and Proposition 2.5 below.

From a different point of view, in order to accelerate the standard PPA, the following *Relaxed Proximal Point Algorithm* is proposed in [9] (partially based on [12]):

$$x^{k+1} = [(1-\rho_k)I + \rho_k J^A_{\lambda_k}](x^k), \tag{RPPA}$$

where $\rho_k \in (0,2)$ is a *relaxation factor* which is supposed to satisfy

$$R_1 := \inf_{k\geq 0} \rho_k > 0 \text{ and } R_2 := \sup_{k\geq 0} \rho_k < 2. \tag{1.6}$$

The overrelaxation $\rho_k \in (1,2)$ may indeed speed up the convergence of the method; see, for instance, [6, pp. 129–131] and [10]. Weak convergence is proved in [9] for an inexact version of RPPA under a standard summable errors condition.

The first aim of this paper is to show that these two acceleration strategies may be coupled in an iteration of the type

$$x^{k+1} = [(1-\rho_k)I + \rho_k J^A_{\lambda_k}](x^k + \alpha_k(x^k - x^{k-1})), \tag{RIPPA}$$

keeping the weak convergence property of the iterates.

On the other hand, from a practical point of view, it is interesting to consider inexact versions of IPPA and RIPPA. Concerning IPPA, a first positive answer is given in [1, Thm. 3.1] for minimization problems, where at each iteration ∂f is replaced with the approximate subdifferential $\partial_{\varepsilon_k} f$, under the hypothesis $\sum \varepsilon_k < \infty$. In this direction, a straightforward adaptation (see, for instance, [19]) of the proof of [2, Thm. 2.1] allows one to deal with the ε_k-enlargement A^{ε_k} of the original operator A. On the other hand, inexact iterations of RPPA are considered in [9], permitting additive residuals in the approximate computation of resolvents under a summability condition analogous to that considered in [22] for PPA. Nevertheless, such inexact criteria requiring summable errors are rather restrictive.

The second goal of this article is to extend the hybrid projection-proximal algorithm introduced in [23] to cover relaxed proximal iterations as RPPA and more generally RIPPA. This hybrid algorithm combines an inexact iteration of PPA with a projection step. In fact, the inexact PPA is used to construct a hyperplane that strictly separates the current iterate x^k from the solution set S; next, x^k is projected onto this separating hyperplane. This method has the remarkable property of permitting a fixed relative error tolerance in the inexact PPA iteration, a less stringent condition, without affecting the global convergence of the algorithm.

This paper is organized as follows. Section 2 introduces an inexact *Relaxed and Inertial Hybrid Projection-Proximal Point Algorithm*, for which weak convergence is proved under conditions (1.3)–(1.6), and then additional conditions on α_k are given in order to ensure (1.5) a priori. Next, a more standard inexact version of RIPPA is considered in section 3, for which weak convergence holds under appropriate summability conditions on the errors.

2. Relaxed and inertial projection-proximal iteration with constant relative error. In what follows, $\sigma \in [0,1)$ is a fixed relative error tolerance. Consider the following iterative scheme:

($\mathcal{A}_1^\rho$) Given $x^k, x^{k-1} \in H, \lambda_k > 0$, $\alpha_k \in [0,1)$, and $\rho_k \in (0,2)$, find $z^k \in H$ such that

$$(z^k - y^k)/\lambda_k + v^k = \eta^k, \text{ for some } v^k \in \rho_k A(z^k/\rho_k + (1-1/\rho_k)y^k), \tag{2.1}$$

where $y^k := x^k + \alpha_k(x^k - x^{k-1})$ and the residual $\eta^k \in H$ satisfies

$$\|\eta^k\| \le \sigma \max\{\|z^k - y^k\|/\lambda_k, \|v^k\|\}. \tag{2.2}$$

$(\mathcal{A}_2^\rho)$ If $v^k = 0$ then set $x^n := y^k$ for all $n \ge k+1$ and stop.
Otherwise:

- Let $P_k : H \to H$ be the orthogonal projection operator onto the hyperplane

$$H_k = \{x \in H \mid \langle v^k, x - z^k\rangle = (1 - 1/\rho_k)\langle v^k, y^k - z^k\rangle\}. \tag{2.3}$$

- Set

$$x^{k+1} := y^k + \rho_k(P_k y^k - y^k) = y^k - \frac{\langle v^k, y^k - z^k\rangle}{\|v^k\|^2} v^k. \tag{2.4}$$

- Let $k \leftarrow k+1$ and return to $(\mathcal{A}_1^\rho)$.

First, note that (2.1) amounts to $z^k = (1-\rho_k)y^k + \rho_k J^A_{\lambda_k}(y^k + (\lambda_k/\rho_k)\eta^k)$. Indeed, the latter is equivalent to $y^k + (\lambda_k/\rho_k)\eta^k \in (I + \lambda_k A)(z^k/\rho_k + (1 - 1/\rho_k)y^k)$, which can be written as $(y^k - z^k)/\lambda_k + \eta^k \in \rho_k A(z^k/\rho_k + (1 - 1/\rho_k)y^k)$, and this is exactly (2.1). In particular, the algorithm described above is well defined.

Notice that if $\eta^k = 0$, then $x^{k+1} = y^k - \lambda_k v^k = y^k - (y^k - z^k) = (1 - \rho_k)y^k + \rho_k J^A_{\lambda_k}(y^k)$. Therefore, $(\mathcal{A}_1^\rho)$–$(\mathcal{A}_2^\rho)$ with $\eta^k = 0$ becomes an exact iteration of RIPPA.

Taking $\sigma > 0$, $\alpha_k \equiv 0$, and $\rho_k \equiv 1$, one recovers the *Hybrid Projection-Proximal Point Algorithm* introduced in [23] (see also [24]), whose main feature is the fixed relative error tolerance given by (2.2). Concerning the projection step given by (2.4), this is necessary in general to ensure the boundedness of the iterates (see [23, p. 62]), even for minimization problems (see [11]).

Some elementary, and key, properties of the relative error criterion are summarized in the following lemma.

LEMMA 2.1. *Let $\sigma \in [0, 1)$. If $v = u + \eta$ with $\|\eta\| \le \sigma \max\{\|u\|, \|v\|\}$, then*
(i) $\|v\| \le \|u\|/(1 - \sigma)$,
(ii) $\langle v, u\rangle \ge (1 - \sigma)\|u\|\|v\|$.

Proof. Suppose $\|v\| > \|u\|$ so that $\|\eta\| \le \sigma\|v\|$; then $\|v\| \le \|u\| + \sigma\|v\|$, or equivalently $\|v\| \le \|u\|/(1 - \sigma)$; otherwise, $\|v\| \le \|u\|$. In any case, (i) holds. For (ii), it suffices to consider the case $\|v\| \le \|u\|$, which implies $\langle v, u\rangle = \|u\|^2 + \langle \eta, u\rangle \ge (1 - \sigma)\|u\|^2 \ge (1 - \sigma)\|u\|\|v\|$. ∎

From (2.1), (2.2), and Lemma 2.1(i), it follows that $v^k = 0$ if and only if $z^k = y^k$. Then, if $v^{k_0} = 0$ for some k_0, then the algorithm ends with y^{k_0} satisfying $0 \in A(y^{k_0})$, a solution to (1.1).

THEOREM 2.2. *Let $(x^k) \subset H$ be a sequence generated by (2.1)–(2.4), where $A : H \rightrightarrows H$ is a maximal monotone operator with $S := A^{-1}(\{0\}) \ne \emptyset$, $\sigma \in [0, 1)$, and the parameters α_k and ρ_k satisfy (1.4) and (1.6), respectively. Under (1.5), the following hold:*

(i) *For all $\bar{x} \in S$, $\|x^k - \bar{x}\|$ is convergent, and*

$$\lim_{k\to\infty} \|x^{k+1} - z^k/\rho_k - (1 - 1/\rho_k)y^k\| = 0. \tag{2.5}$$

(ii) *If in addition λ_k satisfies (1.3), then $\lim_{k\to\infty} \|v^k\| = 0$ and there exists $x^* \in S$ such that $x^k \rightharpoonup x^*$ weakly in H as $k \to \infty$.*

Proof. From now on, assume that $v^k \neq 0$ for all $k \geq 1$; otherwise, the algorithm finishes in a finite number of iterations, providing a solution to (1.1).

Let $\bar{x} \in S = A^{-1}(\{0\})$ and define $\varphi_k := \frac{1}{2}\|x^k - \bar{x}\|^2$. It follows from (2.4) that

$$\begin{aligned}\varphi_{k+1} &= \tfrac{1}{2}\|y^k - \bar{x}\|^2 + \rho_k\langle P_k y^k - y^k, y^k - \bar{x}\rangle + \tfrac{\rho_k^2}{2}\|P_k y^k - y^k\|^2\\ &= \tfrac{1}{2}\|y^k - \bar{x}\|^2 - \rho_k\|P_k y^k - y^k\|^2 + \rho_k\langle P_k y^k - y^k, P_k y^k - \bar{x}\rangle + \tfrac{\rho_k^2}{2}\|P_k y^k - y^k\|^2\\ &= \tfrac{1}{2}\|y^k - \bar{x}\|^2 - \rho_k(1-\rho_k/2)\|P_k y^k - y^k\|^2 + \rho_k\langle P_k y^k - y^k, P_k y^k - \bar{x}\rangle.\end{aligned}$$

Next, notice that, by Lemma 2.1(i), $v^k \neq 0$ implies $(y^k - z^k)/\lambda_k \neq 0$ due to (2.1) and (2.2). Then, by virtue of Lemma 2.1(ii),

$$\ell_k(y^k) = \langle v^k, y^k - z^k\rangle \geq (1-\sigma)\|v^k\|\|y^k - z^k\| > 0, \tag{2.6}$$

where $\ell_k(x) = \langle v^k, x - z^k\rangle$. As $v^k \in \rho_k A(z^k/\rho_k + (1 - 1/\rho_k)y^k)$, the monotonicity of A gives $\langle v^k, \bar{x} - z^k/\rho_k - (1-1/\rho_k)y^k\rangle \leq 0$. Thus, $\bar{x}$ belongs to the half-space $H_k^{\leq} = \{x \in H \mid \ell_k(x) \leq (1-1/\rho_k)\ell_k(y^k)\}$. Therefore, since $\rho_k > 0$ and taking into account (2.6), the hyperplane H_k given by (2.3) strictly separates y^k from $\bar{x}$. Moreover, since the orthogonal projection of y^k onto H_k is also the orthogonal projection onto the half-space $H_k^{\leq}$, one gets $\langle P_k y^k - y^k, P_k y^k - \bar{x}\rangle \leq 0$. It follows that

$$\varphi_{k+1} \leq \tfrac{1}{2}\|y^k - \bar{x}\|^2 - \rho_k(1-\rho_k/2)\|P_k y^k - y^k\|^2. \tag{2.7}$$

But $\frac{1}{2}\|y^k - \bar{x}\|^2 = \varphi_k + \alpha_k\langle x^k - \bar{x}, x^k - x^{k-1}\rangle + \frac{\alpha_k^2}{2}\|x^k - x^{k-1}\|^2$. On the other hand, it is direct to verify that $\varphi_k = \varphi_{k-1} + \langle x^k - \bar{x}, x^k - x^{k-1}\rangle - \frac{1}{2}\|x^k - x^{k-1}\|^2$. Hence

$$\tfrac{1}{2}\|y^k - \bar{x}\|^2 = \varphi_k + \alpha_k(\varphi_k - \varphi_{k-1}) + \tfrac{\alpha_k + \alpha_k^2}{2}\|x^k - x^{k-1}\|^2. \tag{2.8}$$

Thus

$$\varphi_{k+1} \leq \varphi_k + \alpha_k(\varphi_k - \varphi_{k-1}) + \delta_k - \rho_k(1-\rho_k/2)\|P_k y^k - y^k\|^2, \tag{2.9}$$

where $\delta_k := \frac{\alpha_k + \alpha_k^2}{2}\|x^k - x^{k-1}\|^2$, which satisfies $\sum \delta_k < \infty$ thanks to (1.5) (recall that $\alpha_k \in [0, 1)$). The following elementary result is a useful tool for proving convergence for this type of recursive finite difference inequality (see [1, 2]).

LEMMA 2.3. *Let $\varphi_k \geq 0$ and $\delta_k \geq 0$ be such that $\varphi_{k+1} \leq \varphi_k + \alpha_k(\varphi_k - \varphi_{k-1}) + \delta_k$ with $\sum \delta_k < \infty$, and $0 \leq \alpha_k \leq \alpha < 1$. Then the following hold:*

(i) $\sum[\varphi_k - \varphi_{k-1}]_+ < \infty$, *where* $[t]_+ := \max\{t, 0\}$.

(ii) *There exists $\varphi^* \geq 0$ such that $\lim_{k\to\infty}\varphi_k = \varphi^*$.*

Proof. Set $\theta_k = \varphi_k - \varphi_{k-1}$. Then $[\theta_{k+1}]_+ \leq \alpha[\theta_k]_+ + \delta_k$. This yields $[\theta_{k+1}]_+ \leq \alpha^k[\theta_1]_+ + \sum_{j=0}^{k-1}\alpha^j\delta_{k-j}$, so that $\sum[\theta_{k+1}]_+ \leq 1/(1-\alpha)\,([\theta_1]_+ + \sum\delta_k) < \infty$. Set $w_k := \varphi_k - \sum_{j=1}^{k}[\theta_j]_+$, which is bounded from below and nonincreasing. It follows that (w_k) is convergent; hence $\lim_{k\to\infty}\varphi_k = \sum_{j\geq 1}[\theta_j]_+ + \lim_{k\to\infty} w_k$. ∎

By virtue of Lemma 2.3 applied to (2.9), the sequence (φ_k) is convergent under (1.4) and (1.5). Since $\bar{x} \in S$ is arbitrary, the latter proves the first assertion in Theorem 2.2(i).

On the other hand, by Lemma 2.3(i), it follows from (2.9) that

$$\sum \rho_k(1-\rho_k/2)\|P_k y^k - y^k\|^2 \leq \varphi_1 + \alpha\sum[\varphi_k - \varphi_{k-1}]_+ + \sum\delta_k < \infty,$$

which amounts to

$$(1/R_2 - 1/2)\sum(\langle v^k, y^k - z^k\rangle/\|v^k\|)^2 < \infty, \tag{2.10}$$

with $R_2 = \sup_{k\geq 1}\rho_k < 2$ thanks to (1.6). By Lemma 2.1, it may be concluded from (2.10) that

$$\sum \lambda_k^2\|v^k\|^2 \leq \sum \|y^k - z^k\|^2/(1-\sigma)^2 < \infty. \tag{2.11}$$

It follows that

$$\lim_{k\to\infty}\langle v^k, y^k - z^k\rangle/\|v^k\| = \lim_{k\to\infty}\|y^k - z^k\| = \lim_{k\to\infty}\lambda_k\|v^k\| = 0. \tag{2.12}$$

By (2.4), the first limit in (2.12) ensures that $\lim_{k\to\infty}\|x^{k+1} - y^k\| = 0$. From this fact, together with the second limit in (2.12), it follows that (2.5) holds because $R_1 = \inf\rho_k > 0$ due to (1.6). This completes the proof of Theorem 2.2(i).

In order to prove Theorem 2.2(ii), the idea is to apply the following well-known result on weak convergence in Hilbert spaces, whose proof is given here for the convenience of the reader.

LEMMA 2.4 (Opial). *Let H be a Hilbert space and (x^k) a sequence such that there exists a nonempty set $S \subset H$ verifying the following:*

(a) *For every $\bar{x} \in S$, $\lim_{k\to\infty}\|x^k - \bar{x}\|$ exists.*

(b) *If $x^{k_j} \rightharpoonup \hat{x}$ weakly in H for a subsequence $k_j \to \infty$, then $\hat{x} \in S$.*

Then, there exists $x^* \in S$ such that $x^k \rightharpoonup x^*$ weakly in H as $k \to \infty$.

Proof. It suffices to prove the uniqueness of the weak cluster point. The original proof in [20] requires S to be closed and convex. The following argument (see [15, 22]) does not need that hypothesis. Let $\widehat{x}_1, \widehat{x}_2 \in S$ be two cluster points of (x^k) for the weak topology of H. Set $l_i := \lim_{k\to\infty}\|x^k - \widehat{x}_i\|^2$ for each $i = 1, 2$. Take a sequence $k_j \to \infty$ such that $x^{k_j} \rightharpoonup \widehat{x}_1$ weakly in H. But $\|x^k - \widehat{x}_1\|^2 - \|x^k - \widehat{x}_2\|^2 = \|\widehat{x}_1 - \widehat{x}_2\|^2 + 2\langle\widehat{x}_1 - \widehat{x}_2, \widehat{x}_2 - x^k\rangle$, so that $l_1 - l_2 = -\|\widehat{x}_1 - \widehat{x}_2\|^2$. Similarly, taking $k_m \to \infty$ such that $x^{k_m} \rightharpoonup \widehat{x}_2$, $l_1 - l_2 = \|\widehat{x}_1 - \widehat{x}_2\|^2$. Consequently, $\|\widehat{x}_1 - \widehat{x}_2\| = 0$. ∎

By Theorem 2.2(i), condition (a) of Lemma 2.4 holds with $S = A^{-1}(\{0\})$. Next, suppose (1.3) and let $\hat{x}$ be a weak cluster point of (x^k). By (2.5), $z^k/\rho_k + (1 - 1/\rho_k)y^k \rightharpoonup \hat{x}$. But

$$v^k/\rho_k \in A(z^k/\rho_k + (1 - 1/\rho_k)y^k), \tag{2.13}$$

with $v^k/\rho_k \to 0$ strongly in H thanks to the last limit in (2.12) together with (1.3) and (1.6). Since the graph of the maximal monotone operator A is closed in $H \times H$ for the weak-strong topology (see [7]), it is possible to pass to the limit in (2.13) to deduce that $0 \in A(\hat{x})$, i.e., $\hat{x} \in S$. Thus, condition (b) of Lemma 2.4 is also satisfied, which proves the weak convergence of (x^k). ∎

Remark 1. If (1.3) is replaced with

$$\sum \lambda_k^2 = \infty, \tag{2.14}$$

then it may be deduced from (2.11) that there exists a subsequence of (v^k) that converges strongly to 0. In the finite dimensional case, this is sufficient for the convergence of (x^k) (see [23, Rem. 2.3]). Indeed, take $v^{k_i} \to 0$ and assume that $\dim H < \infty$. By

Theorem 2.2(i), (x^k) is bounded so that one may assume that, up to a subsequence, $x^{k_i+1} \to \hat{x}$ for some $\hat{x} \in H$. By virtue of (2.5), one may let $k_i \to \infty$ in (2.13) to deduce that $0 \in A(\hat{x})$. Hence $\hat{x} \in S$ and, by Theorem 2.2(i), $\|x^k - \hat{x}\|$ is convergent. Therefore $\lim_{k\to\infty} \|x^k - \hat{x}\| = \lim_{i\to\infty} \|x^{k_i+1} - \hat{x}\| = 0$.

In practical computations, it is easy to enforce (1.5) by means of a dynamic rule to update the inertial parameter α_k, taking into account the current value of $\|x^k - x^{k-1}\|$. Furthermore, (1.5) holds a priori in some special cases as the next result shows, extending [2, Prop. 2.1].

PROPOSITION 2.5. *Under the assumptions of Theorem* 2.2 *with, in addition,* (α_k) *being nondecreasing (i.e.,* $\alpha_{k+1} \geq \alpha_k$*) and satisfying* $0 \leq \alpha_k \leq \alpha$ *for some* $\alpha \in [0, 1)$ *such that*

$$0 < p(\alpha) := 1/R_2 - 1/2 - (2/R_1 - 1/2)\alpha - (1 - 1/R_2)\alpha^2, \tag{2.15}$$

then $\sum \|x^k - x^{k-1}\|^2 < \infty$. *In particular,* (1.5) *holds and thus there exists* $\widehat{x} \in S$ *such that* $x^k \rightharpoonup \widehat{x}$ *weakly in* H *as* $k \to \infty$.

Proof. Noticing that $\rho_k^2\|P_k y^k - y^k\|^2 = \|x^{k+1} - y^k\|^2 = \|x^{k+1} - x^k\|^2 - 2\alpha_k\langle x^{k+1} - x^k, x^k - x^{k-1}\rangle + \alpha_k^2\|x^k - x^{k-1}\|^2 \geq (1 - \alpha_k)\|x^{k+1} - x^k\|^2 - \alpha_k(1 - \alpha_k)\|x^k - x^{k-1}\|^2$, it follows from (2.9) that

$$\begin{aligned}\varphi_{k+1} \leq \varphi_k + \alpha_k(\varphi_k - \varphi_{k-1}) + [(\alpha_k + \alpha_k^2)/2 + (1/\rho_k - 1/2)\alpha_k(1 - \alpha_k)]\|x^k - x^{k-1}\|^2 \\ - (1/\rho_k - 1/2)(1 - \alpha_k)\|x^{k+1} - x^k\|^2,\end{aligned}$$

where $\varphi_k := \frac{1}{2}\|x^k - \bar{x}\|^2$. This yields

$$\begin{aligned}\varphi_{k+1} - \alpha_k\varphi_k \leq \varphi_k - \alpha_k\varphi_{k-1} + \alpha_k[1/\rho_k + (1 - 1/\rho_k)\alpha_k]\|x^k - x^{k-1}\|^2 \\ - (1/\rho_k - 1/2)(1 - \alpha_k)\|x^{k+1} - x^k\|^2.\end{aligned}$$

Setting $\mu_k := \varphi_k - \alpha_k\varphi_{k-1} + \alpha_k[1/\rho_k + (1 - 1/\rho_k)\alpha_k]\|x^k - x^{k-1}\|^2$, and since $\alpha_{k+1} \geq \alpha_k$, we obtain

$$\mu_{k+1} \leq \mu_k + [\alpha_{k+1}/\rho_{k+1} + (1 - 1/\rho_{k+1})\alpha_{k+1}^2 + (1/\rho_k - 1/2)(\alpha_k - 1)]\|x^{k+1} - x^k\|^2.$$

But $\alpha_{k+1}/\rho_{k+1} + (1 - 1/\rho_{k+1})\alpha_{k+1}^2 \leq \alpha/R_1 + (1 - 1/R_2)\alpha^2$ and $(1/\rho_k - 1/2)(\alpha_k - 1) \leq (1/R_1 - 1/2)\alpha - 1/R_2 + 1/2$. Therefore $\mu_{k+1} \leq \mu_k - p(\alpha)\|x^{k+1} - x^k\|^2$, where $p(\alpha)$ is given by (2.15). Since $p(\alpha) > 0$, (μ_k) is nonincreasing, which implies $\varphi_k \leq \alpha\varphi_{k-1} + \mu_k \leq \alpha\varphi_{k-1} + \mu_1$. This gives $\varphi_k \leq \alpha^k\varphi_0 + \mu_1\sum_{j=0}^{k-1}\alpha^j \leq \alpha^k\varphi_0 + \mu_1/(1 - \alpha)$. Furthermore, it follows that $p(\alpha)\sum_{j=0}^{k}\|x^{j+1} - x^j\|^2 \leq \mu_1 - \mu_{k+1} \leq \mu_1 + \alpha\varphi_k \leq \alpha^{k+1}\varphi_0 + \mu_1/(1 - \alpha)$. This shows that $\sum\|x^k - x^{k-1}\|^2 \leq 2\mu_1/((1 - \alpha)p(\alpha))$. The conclusion follows by Theorem 2.2. ∎

Remark 2. Suppose $R_2 \geq 1$. Since $p(0) = 1/R_2 - 1/2 > 0$ thanks to (1.6), there exists a unique positive root $\alpha^* > 0$ of the quadratic polynomial $p(\alpha)$, and for all $\alpha \in [0, \alpha^*)$, $p(\alpha) > 0$. For instance, when $\rho_k \equiv 1$, one gets $p(\alpha) = 1/2 - (3/2)\alpha$ and so $\alpha^* = 1/3$.

3. An alternative inexact scheme with summable residuals. It follows from (2.2) and (2.11) that the sequence of residuals (η^k) associated with the sequence (x^k) generated by (2.1)–(2.4) satisfies $\sum\lambda_k^2\|\eta^k\|^2 < \infty$, and hence $\sum\|\eta^k\|^2 < \infty$ in view of (1.3). However, it may occur that $\sum\|\eta^k\| = \infty$; see [11] for an example with $\rho_k \equiv 1$ and $\alpha_k \equiv 0$, which is based on [13]. The constant relative error criterion (2.2) is thus less stringent than

$$\sum\lambda_k\|\eta^k\| < \infty. \tag{3.1}$$

On the other hand, the next result, which extends [9, Thm. 3], shows that under such a summability condition the projection step is not necessary for convergence.

THEOREM 3.1. *Let $A : H \rightrightarrows H$ be a maximal monotone operator with $S := A^{-1}(\{0\}) \neq \emptyset$ and $(x^k) \subset H$ a sequence satisfying*

$$(x^{k+1} - y^k)/\lambda_k + v^k = \eta^k \quad \text{for some } v^k \in \rho_k A(x^{k+1}/\rho_k + (1 - 1/\rho_k)y^k), \tag{3.2}$$

where $y^k = x^k + \alpha_k(x^k - x^{k-1})$, and the parameters λ_k, α_k, and ρ_k satisfy (1.3), (1.4), *and* (1.6), *respectively. Suppose* (1.5), (3.1), *and*

$$\sum \lambda_k \|\eta^k\| \|y^k\| < \infty. \tag{3.3}$$

Then $v^k \to 0$ strongly in H and there exists $x^* \in S$ such that $x^k \rightharpoonup x^*$ weakly in H.

Proof. It is easy to see that (3.2) amounts to $x^{k+1} = (1 - \rho_k)y^k + \rho_k J^A_{\lambda_k}(y^k + (\lambda_k/\rho_k)\eta^k)$. Let (w^k) be the auxiliary sequence defined by

$$w^k := (1 - \rho_k)y^k + \rho_k J^A_{\lambda_k}(y^k). \tag{3.4}$$

Since J^A_λ is nonexpansive,

$$\|x^{k+1} - w^k\| \leq \lambda_k \|\eta^k\|. \tag{3.5}$$

On the other hand, (3.4) may be written as $w^k = y^k - \lambda_k \rho_k A_{\lambda_k}(y^k)$, where $A_\lambda : H \to H$ is given by $A_\lambda = \frac{1}{\lambda}(I - J^A_\lambda)$. Thanks to (1.2),

$$0 \in A(x) \text{ if and only if } A_\lambda(x) = 0. \tag{3.6}$$

Moreover, as J^A_λ is nonexpansive, A_λ is a cocoercive maximal monotone operator of parameter λ; that is,

$$\forall x_1, x_2 \in H, \langle A_\lambda(x_1) - A_\lambda(x_2), x_1 - x_2\rangle \geq \lambda \|A_\lambda(x_1) - A_\lambda(x_2)\|^2. \tag{3.7}$$

Let $\bar{x} \in S$. By (3.6), $A_\lambda(\bar{x}) = 0$ and since $\frac{1}{2}\|w^k - \bar{x}\|^2 = \frac{1}{2}\|y^k - \bar{x}\|^2 - \rho_k\lambda_k\langle y^k - \bar{x}, A_{\lambda_k}(y^k)\rangle + \frac{(\rho_k\lambda_k)^2}{2}\|A_{\lambda_k}(y^k)\|^2$, the cocoercivity property (3.7) yields

$$\tfrac{1}{2}\|w^k - \bar{x}\|^2 \leq \tfrac{1}{2}\|y^k - \bar{x}\|^2 - \lambda_k^2\rho_k(1 - \rho_k/2)\|A_{\lambda_k}(y^k)\|^2. \tag{3.8}$$

Define $\varphi_k := \frac{1}{2}\|x^k - \bar{x}\|^2$. Then $\varphi_{k+1} \leq \frac{1}{2}\|w^k - \bar{x}\|^2 + \|x^{k+1} - w^k\|\|w^k - \bar{x}\| + \frac{1}{2}\|x^{k+1} - w^k\|^2$. By (3.5) and (3.8),

$$\varphi_{k+1} \leq \tfrac{1}{2}\|y^k - \bar{x}\|^2 - \lambda_k^2\rho_k(1 - \rho_k/2)\|A_{\lambda_k}(y^k)\|^2 + \lambda_k\|\eta^k\|\|y^k - \bar{x}\| + \tfrac{\lambda_k^2}{2}\|\eta^k\|^2. \tag{3.9}$$

Recalling (2.8), it follows that

$$\varphi_{k+1} \leq \varphi_k + \alpha_k(\varphi_k - \varphi_{k-1}) + \delta_k - \lambda_k^2\rho_k(1 - \rho_k/2)\|A_{\lambda_k}(y^k)\|^2, \tag{3.10}$$

where $\delta_k := \frac{\alpha_k + \alpha_k^2}{2}\|x^k - x^{k-1}\|^2 + \lambda_k\|\eta^k\|\|y^k - \bar{x}\| + \frac{\lambda_k^2}{2}\|\eta^k\|^2$. Under (1.5), (3.1), if (3.3) holds, then $\sum \delta_k < \infty$. Thus $\sum \delta_k < \infty$ and, by virtue of Lemma 2.3, (φ_k) is convergent. Moreover, we deduce that $\sum \lambda_k^2\|A_{\lambda_k}(y^k)\|^2 < \infty$. Set $\xi^k := y^k - J^A_{\lambda_k}(y^k)$, which amounts to

$$\xi^k/\lambda_k \in A(y^k - \xi^k). \tag{3.11}$$

Since $\sum \|\xi^k\|^2 < \infty$, in particular $\lim_{k\to\infty} \xi^k = 0$. Let $\hat{x}$ be a weak cluster point of (x^k). Since $\lim_{k\to\infty} \alpha_k \|x^k - x^{k-1}\| = 0$, $y^k \rightharpoonup \hat{x}$ and consequently $y^k - \xi^k \rightharpoonup \hat{x}$. By the weak-strong closedness of the graph of A, letting $k \to \infty$ in (3.11) gives $0 \in A(\hat{x})$. Therefore, condition (b) of Lemma 2.4 holds, which finishes the proof. □

Remark 3. Under (1.5) and (3.1), assume

$$\sum \alpha_k \|x^k - x^{k-1}\| < \infty. \tag{3.12}$$

From (3.9), it follows that $\|x^{k+1} - \bar{x}\| \le \|y^k - \bar{x}\| + \lambda_k \|\eta^k\| \le \|x^k - \bar{x}\| + \alpha_k \|x^k - x^{k-1}\| + \lambda_k \|\eta^k\|$. Using (3.1) and (3.12), $\|x^k - \bar{x}\|$ is convergent; in particular (y^k) is bounded. Hence in view of (3.1), condition (3.3) is realized.

Acknowledgments. The author thanks the hospitality of the Sydoco research team of INRIA-Rocquencourt (France) where part of this work was carried out. The author also wishes to thank the anonymous referee whose remarks helped him to improve the presentation of this paper.

REFERENCES

[1] F. Alvarez, *On the minimizing property of a second order dissipative system in Hilbert spaces*, SIAM J. Control Optim., 38 (2000), pp. 1102–1119.

[2] F. Alvarez and H. Attouch, *An inertial proximal method for maximal monotone operators via discretization of a nonlinear oscillator with damping*, in Wellposedness in Optimization and Related Topics (Gargnano, 1999), Set-Valued Anal., 9 (2001), pp. 3–11.

[3] A. S. Antipin, *Minimization of convex functions on convex sets by means of differential equations*, Differential Equations, 30 (1994), pp. 1365–1375.

[4] H. Attouch, X. Goudou, and P. Redont, *The heavy ball with friction method.* I. *The continuous dynamical system*, Commun. Contemp. Math., 2 (2000), pp. 1–34.

[5] J. B. Baillon, *Un exemple concernant le comportement asymptotique de la solution du problème* $du/dt + \partial\varphi(u) \ni 0$, J. Funct. Anal., 28 (1978), pp. 369–376.

[6] D. P. Bertsekas, *Constrained Optimization and Lagrange Multiplier Methods*, Academic Press, New York, 1982.

[7] H. Brezis, *Opérateurs maximaux monotones et semi-groupes de contractions dans les espaces de Hilbert*, Mathematics Studies 5, North–Holland, Amsterdam, 1973.

[8] R. E. Bruck, *Asymptotic convergence of nonlinear contraction semigroups in Hilbert space*, J. Funct. Anal., 18 (1975), pp. 15–26.

[9] J. Eckstein and D. P. Bertsekas, *On the Douglas-Rachford splitting method and the proximal point algorithm for maximal monotone operators*, Math. Program., 55 (1992), pp. 293–318.

[10] J. Eckstein and M. C. Ferris, *Operator-splitting methods for monotone affine variational inequalities, with a parallel application to optimal control*, INFORMS J. Comput., 10 (1998), pp. 218–235.

[11] O. R. Gárciga, A. Iusem, and B. F. Svaiter, *On the need for hybrid steps in hybrid proximal point methods*, Oper. Res. Lett., 29 (2001), pp. 217–220.

[12] E. G. Gol'shtein and N. V. Tret'yakov, *Modified Lagrangians in convex programming and their generalizations*, Math. Program. Stud., 10 (1979), pp. 86–97.

[13] O. Güler, *On the convergence of the proximal point algorithm for convex minimization*, SIAM J. Control Optim., 29 (1991), pp. 403–419.

[14] F. Jules and P. E. Maingé, *Numerical approach to a stationary solution of a second order dissipative dynamical system*, Optimization, 51 (2002), pp. 235–255.

[15] B. Martinet, *Régularisation d'inéquations variationnelles par approximations successives*, Rev. Française Informat. et Recherche Opérationnelle, 4 (1970), pp. 154–158.

[16] B. Martinet, *Determination approchée d'un point fixe d'une application pseudo-contractante*, C. R. Acad. Sci. Paris Ser. A-B, 274 (1972), pp. 163–165.

[17] G. Minty, *Monotone (nonlinear) operators in Hilbert space*, Duke Math. J., 29 (1962), pp. 341–346.

[18] J. J. Moreau, *Proximité et dualité dans un espace hilbertien*, Bull. Soc. Math. France, 93 (1965), pp. 273–299.

[19] A. Moudafi, *Second-order differential proximal methods for equilibrium problems*, JIPAM J. Inequal. Pure Appl. Math., 14 (2003).

[20] Z. Opial, *Weak convergence of the sequence of successive approximations for nonexpansive mappings*, Bull. Amer. Math. Soc., 73 (1967), pp. 591–597.
[21] B. T. Polyak, *Some methods of speeding up the convergence of iterative methods*, Zh. Vychisl. Mat. Mat. Fiz., 4 (1964), pp. 1–17.
[22] R. T. Rockafellar, *Monotone operators and the proximal point algorithm*, SIAM J. Control Optim., 14 (1976), pp. 877–898.
[23] M. V. Solodov and B. F. Svaiter, *A hybrid projection-proximal point algorithm*, J. Convex Anal., 6 (1999), pp. 59–70.
[24] M. V. Solodov and B. F. Svaiter, *Forcing strong convergence of proximal point iterations in a Hilbert space*, Math. Program., 87 (2000), pp. 189–202.
[25] F. Zirilli, F. Aluffi and V. Parisi, *DAFNE: A Differential Equations Algorithm for Nonlinear Equations*, ACM Trans. Math. Software, 10 (1984), pp. 317–324.

SIAM J. OPTIM.
Vol. 14, No. 3, pp. 783–806

A SQUARED SMOOTHING NEWTON METHOD FOR NONSMOOTH MATRIX EQUATIONS AND ITS APPLICATIONS IN SEMIDEFINITE OPTIMIZATION PROBLEMS*

JIE SUN†, DEFENG SUN‡, AND LIQUN QI§

Abstract. We study a smoothing Newton method for solving a nonsmooth matrix equation that includes semidefinite programming and the semidefinite complementarity problem as special cases. This method, if specialized for solving semidefinite programs, needs to solve only one linear system per iteration and achieves quadratic convergence under strict complementarity and nondegeneracy. We also establish quadratic convergence of this method applied to the semidefinite complementarity problem under the assumption that the Jacobian of the problem is positive definite on the affine hull of the critical cone at the solution. These results are based on the strong semismoothness and complete characterization of the B-subdifferential of a corresponding squared smoothing matrix function, which are of general theoretical interest.

Key words. matrix equations, Newton's method, nonsmooth optimization, semidefinite complementarity problem, semidefinite programming

AMS subject classifications. 65K05, 90C25, 90C33

DOI. 10.1137/S1052623400379620

1. Introduction.

1.1. Motivation. Let $\mathcal{S}(n_1,\ldots,n_m)$ be the linear space of symmetric block-diagonal matrices with m blocks of sizes $n_k \times n_k,\ k=1,\ldots,m$, respectively, and let Ψ be a mapping from $\mathcal{S}(n_1,\ldots,n_m)$ to $\mathcal{S}(n_1,\ldots,n_m)$ itself. We consider the problem of finding a root of $\Psi(X)=0$. This symmetric block-diagonal-matrix-valued equation problem (*matrix equation problem* for short) has many applications in optimization. For example, arising from Lyapunov stability analysis of systems under uncertainty [4, 23], we desire to know whether there exists an $n \times n$ symmetric matrix X such that the following system is feasible:

$$\begin{cases} \lambda X - (L_i X + X L_i) \succeq 0, & i=1,\ldots,k, \\ X - I \succeq 0, \end{cases} \tag{1.1}$$

where λ is a given constant, $I, L_i,\ i=1,\ldots,k$ are given $n \times n$ symmetric matrices, and for an arbitrary symmetric matrix Y we write $Y \succ 0$ and $Y \succeq 0$ if Y is positive definite and positive semidefinite, respectively. It is easy to convert (1.1) into a matrix equation problem. For $X \succeq 0$ we denote its symmetric square root by $X^{1/2}$. Let $|X| := (X^2)^{1/2}$ and $X_+ := (X+|X|)/2$ for any $X \in \mathcal{S}(n_1,\ldots,n_m)$. Note that

*Received by the editors October 17, 2000; accepted for publication (in revised form) October 15, 2003; published electronically March 5, 2004.
http://www.siam.org/journals/siopt/14-3/37962.html

†School of Business and Singapore-MIT Alliance, National University of Singapore, Republic of Singapore (jsun@nus.edu.sg). The research of this author was partially supported by grant R314-000-028/042-112 of the National University of Singapore and a grant from Singapore-MIT Alliance.

‡Department of Mathematics, National University of Singapore, Republic of Singapore (matsundf@nus.edu.sg). The research of this author was partially supported by the Australian Research Council and grant R146-000-035-101 of the National University of Singapore.

§Department of Applied Mathematics, the Hong Kong Polytechnic University, Hong Kong, China (maqilq@polyu.edu.hk). The research of this author was supported by the Research Grant Council of Hong Kong.

$|X| - X = 0$ if and only if X is positive semidefinite. Let

$$\Psi(X) := \sum_{i=1}^{k} [\,|\,\lambda X - L_i X - X L_i\,|\,-\lambda X + L_i X + X L_i\,] + [\,|\,X - I\,|\,- X + I\,].$$

Then solving problem (1.1) is equivalent to solving the matrix equation $\Psi(X) = 0$. Note that this equation is not differentiable (in the sense of Fréchet), but is strongly semismooth [5, 32]. For the definition of semismooth matrix functions and some related topics see section 2 or references [5, 32] for more details.

Another application of matrix equations refers to semidefinite programming (SDP). As a modeling tool of optimization and a powerful relaxation form of some combinatorial optimization problems, SDP has received much attention in the research community in recent years. The website of semidefinite programming[1] contains a nice categorized list of papers in this area. Assuming strict feasibility of both primal and dual problems, a semidefinite program is equivalent to finding $X \succeq 0$, $S \succeq 0$, and $y \in \mathbb{R}^m$ such that

$$A_i \bullet X = b_i, \quad i = 1, \ldots, m, \quad \sum_{i=1}^{m} y_i A_i + S = C, \quad X \bullet S = 0, \tag{1.2}$$

where $\bullet$ denotes the matrix Frobenius inner product. It is shown by Tseng [35] that

$$X \succeq 0, \quad S \succeq 0, \quad X \bullet S = 0 \quad \Longleftrightarrow \quad X - [X - S]_+ = 0. \tag{1.3}$$

Thus, system (1.2) can be rewritten as

$$A_i \bullet X = b_i, \quad i = 1, \ldots, m, \quad \sum_{i=1}^{m} y_i A_i + S = C,\ X - [X - S]_+ = 0, \tag{1.4}$$

which has the form of $\Psi(W) = 0$ with $W := \mathrm{diag}\,(y_1, \ldots, y_m, S, X)$ being a block-diagonal matrix.

A generalization of SDP—the semidefinite complementarity problem (SDCP)—can also be reformulated as a matrix equation. The SDCP is to find, for a given continuously differentiable mapping $F : \mathcal{S}(n_1, \ldots, n_m) \to \mathcal{S}(n_1, \ldots, n_m)$, an $X \in \mathcal{S}(n_1, \ldots, n_m)$ such that

$$X \succeq 0, \quad F(X) \succeq 0, \quad X \bullet F(X) = 0. \tag{1.5}$$

By (1.3) this problem is equivalent to

$$X - [X - F(X)]_+ = 0. \tag{1.6}$$

A special case of the SDCP, where F is linear, was introduced by Kojima, Shindo, and Hara [19] and further studied in, e.g., [12, 13, 17, 18]. For the general (nonlinear) SDCP, Monteiro and Pang [21, 22] treated it as a constrained equation and introduced interior-point methods for solving the constrained equation. Tseng [35] introduced merit functions to reformulate the SDCP as an optimization problem. Chen and Tseng [6] studied noninterior continuation methods for solving the SDCP. Kanzow and Nagel [15] analyzed smoothing paths for the Karush–Kuhn–Tucker (KKT) system of

[1] http://www.zib.de/helmberg/semidef.html

the SDP and proposed smoothing-type methods for solving the KKT system. Pang, Sun, and Sun [24] studied semismooth homeomorphisms and strong stability of the SDCP.

The interest in the nonlinear SDCP stems from the research on nonlinear semidefinite optimization problems. Shapiro [29] studied first- and second-order perturbation analysis of nonlinear semidefinite optimization problems. Jarre [14] gave an interior-point method for solving nonconvex semidefinite programs. Fares, Noll, and Apkarian [7] investigated a sequential SDP approach for a variety of problems in optimal control, which can be cast as minimizing a linear objective function subject to linear matrix inequality constraints and nonlinear matrix equality constraints. Leibfritz and Mostafa [20] proposed an interior-point constrained trust-region method for a special class of nonlinear SDP problems. Tseng [36] conducted a convergence analysis for an infeasible interior-point trust-region method for nonlinear semidefinite programs.

In this paper we study a smoothing Newton method for solving a nonsmooth matrix equation that includes the SDP and the SDCP as special cases. In particular, for the SDP, this method achieves quadratic convergence under strict complementarity and nondegeneracy. For the SDCP, quadratic convergence is proved under the condition that the Jacobian of the problem is positive definite on the affine hull of the critical cone at the solution. The strict complementarity condition is not assumed here. To establish these results, we investigate the strong semismoothness and the Bouligand-subdifferential (B-subdifferential) of the so-called squared smoothing matrix function, which are of their own theoretical interest.

The study on smoothing Newton methods can be traced back to a nonsmooth version of Newton's method by Qi and Sun [27] for solving nonsmooth vector valued equations. It was later found that smoothing techniques could be applied to the nonsmooth Newton method to improve its computational performance. Many researchers have contributed to this area, see, for example, [11] and the references therein. The basic idea of the smoothing Newton method is to replace the nonsmooth equation $\Psi(X) = 0$ by a smoothing equation $G(\varepsilon, X) = 0$, where $G : \mathbb{R} \times \mathcal{S}(n_1, \ldots, n_m) \to \mathcal{S}(n_1, \ldots, n_m)$, such that

$$G(\varepsilon, Y) \to \Psi(X) \quad \text{as } (\varepsilon, Y) \to (0, X).$$

Here the function G is required to be continuously differentiable at (ε, X) unless $\varepsilon = 0$. The classical damped Newton method can then be used to solve $G(\varepsilon, X) = 0$ as $\varepsilon \downarrow 0$ to get a solution of $\Psi(X) = 0$. Computational results show that this type of method is quite efficient in solving vector complementarity problems [37].

For $\varepsilon \in \mathbb{R}$ and $X \in \mathcal{S}(n_1, \ldots, n_m)$, the *squared smoothing function* $\Phi : \mathbb{R} \times \mathcal{S}(n_1, \ldots, n_m) \to \mathcal{S}(n_1, \ldots, n_m)$ is defined by

$$\Phi(\varepsilon, X) := (\varepsilon^2 I + X^2)^{1/2}, \quad (\varepsilon, X) \in \mathbb{R} \times \mathcal{S}(n_1, \ldots, n_m). \tag{1.7}$$

Then, Φ is continuously differentiable at (ε, X) unless $\varepsilon = 0$, and for any $X \in \mathcal{S}(n_1, \ldots, n_m)$,

$$[Y + \Phi(\varepsilon, Y)]/2 \to X_+ \quad \text{as } (\varepsilon, Y) \to (0, X).$$

Thus we can use Φ to construct smoothing functions for nonsmooth systems (1.4) and (1.6). We show that the smoothing function

$$G(\varepsilon, X) := X - [X - F(X) + \Phi(\varepsilon, X - F(X))]/2 \tag{1.8}$$

can be used to design a quadratically convergent algorithm for (1.4) and (1.6). We note that Chen and Tseng [6] have developed a nice smoothing Newton method for the SDCP and reported promising computational results. The difference between our paper and theirs is that we show the strong semismoothness of the smoothing function, which can be utilized to establish quadratic convergence, whereas paper [6] did not prove the strong semismoothness of the smoothing function. As a result, paper [6] needs the strict complementarity assumption and the convergence rate proved there is only superlinear, whereas we obtain quadratic rate of convergence without this assumption for the SDCP.

1.2. Notation and organization of the paper. The notation used is fairly standard. Generally, we use calligraphic letters for sets, capital letters for matrices and matrix functions, lowercase letters for vectors, and Greek letters for scalars and index sets, respectively. A diagonal matrix is denoted by $\mathrm{diag}\,(\lambda_1, \ldots, \lambda_n)$, where $\lambda_1, \ldots, \lambda_n$ are the diagonal entries. Similarly, a block-diagonal matrix is written as $\mathrm{diag}\,(B_1, \ldots, B_m)$ with $B_1, \ldots, B_m$ being the block matrices.

Let α and β be two sets of indices. We designate by $A_{\alpha\beta}$ the submatrix of A whose row indices belong to α and whose column indices belong to β. In particular, A_{ij} stands for the (i,j)th entry of A. For matrices $A, B \in \mathcal{S}(n_1, \ldots, n_m)$, the Frobenius inner product is defined as

$$A \bullet B := \mathrm{Trace}\,(A^T B) = \mathrm{Trace}\,(AB).$$

Consequently, the Frobenius norm of $A \in \mathcal{S}(n_1, \ldots, n_m)$ is

$$\|A\| := (A \bullet A)^{1/2}.$$

The Hadamard product of A and B is denoted by $A \circ B$, namely, $(A \circ B)_{ij} := A_{ij} B_{ij}$ for all i and j. The 2-norm of a vector x is denoted by $\|x\|$. Let I be the identity matrix of appropriate dimension.

This paper is organized as follows. In section 2 we review some results on nonsmooth matrix functions and prove the strong semismoothness of Φ defined in (1.7). Section 3 is devoted to characterizing the B-subdifferential of Φ, which will be used in the sequel. We describe the squared smoothing Newton method in section 4. Applications of the smoothing Newton method to the SDP and SDCP are discussed in sections 5 and 6, respectively. Some final remarks are given in section 7.

2. Strong semismoothness of $\Phi(\varepsilon, X)$. This section is devoted to proving the strong semismoothness of the squared smoothing function Φ defined by (1.7). As a preparation we introduce some basic definitions and results on a general matrix function $\Psi : \mathcal{S}(n_1, \ldots, n_m) \to \mathcal{S}_1$, where $\mathcal{S}_1$ is also a symmetric block-diagonal matrix space, but could be of different shape and size from $\mathcal{S}(n_1, \ldots, n_m)$.

Suppose that $\Psi : \mathcal{S}(n_1, \ldots, n_m) \to \mathcal{S}_1$ is a locally Lipschitz matrix function. According to [32], Ψ is differentiable almost everywhere. Denote the set of points at which Ψ is differentiable by D_Ψ and for any $X \in D_\Psi$, let $J\Psi(X)$ denote the Jacobian of Ψ at X. Let $\partial_B \Psi(X)$ be the B-subdifferential of Ψ at X defined by

$$\partial_B \Psi(X) = \left\{ \lim_{\substack{X^k \to X \\ X^k \in D_\Psi}} J\Psi(X^k) \right\}, \tag{2.1}$$

and let $\partial \Psi(X)$ denote the convex hull of $\partial_B \Psi(X)$.

DEFINITION 2.1. *Suppose that* $\Psi : \mathcal{S}(n_1, \ldots, n_m) \to \mathcal{S}_1$ *is a locally Lipschitz matrix function.* Ψ *is said to be semismooth at* $X \in \mathcal{S}(n_1, \ldots, n_m)$ *if* Ψ *is directionally differentiable at* X *and for any* $V \in \partial\Psi(X + H)$ *and* $H \in \mathcal{S}(n_1, \ldots, n_m)$,

$$\Psi(X + H) - \Psi(X) - V(H) \,=\, o(\| H \|)\,.$$

Ψ *is said to be strongly semismooth at* X *if* Ψ *is semismooth at* X *and*

$$\Psi(X + H) - \Psi(X) - V(H) \,=\, O(\| H \|^2)\,. \tag{2.2}$$

Instead of showing the strong semismoothness by definition, we will use the following result [32, Theorem 3.6].

THEOREM 2.2. *Suppose that* $\Psi : \mathcal{S}(n_1, \ldots, n_m) \to \mathcal{S}_1$ *is locally Lipschitz and directionally differentiable in a neighborhood of* X. *Then* Ψ *is strongly semismooth at* X *if and only if for any* $X + H \in D_\Psi$,

$$\Psi(X + H) - \Psi(X) - J\Psi(X + H)(H) \,=\, O(\| H \|^2). \tag{2.3}$$

In order to show that $\Phi(\varepsilon, X)$ satisfies (2.3), we will first identify the differentiable points of Φ. We shall show that Φ is differentiable at (ε, X) if and only if $\varepsilon^2 I + X^2$ is nonsingular. Here we view Φ as a function from $\mathcal{S}(1, n)$ to $\mathcal{S} \equiv \mathcal{S}(n)$. This result easily can be extended to the general block-diagonal case. Unless stated otherwise, $\mathcal{S}$ is assumed to be of this simple structure here and below.

For any $X \in \mathcal{S}$, let L_X be the Lyapunov operator

$$L_X(Y) \,:=\, XY + YX \quad \forall\, Y \,\in\, \mathcal{S}$$

with L_X^{-1} being its inverse (if it exists at all).

For $X \in \mathcal{S}$, there exist an orthogonal matrix P and a diagonal matrix $\Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$ of eigenvalues of X such that

$$X \,=\, P\Lambda P^T. \tag{2.4}$$

Define three index sets associated with the eigenvalues of matrix X:

$$\alpha \,:=\, \{\, i \,:\, \lambda_i \,>\, 0\,\}, \quad \beta \,:=\, \{\, i \,:\, \lambda_i \,=\, 0\,\}, \quad \text{and} \quad \gamma \,:=\, \{\, i \,:\, \lambda_i \,<\, 0\,\}.$$

By permuting the rows and columns of X if necessary, we assume that Λ can be written as

$$\Lambda \,=\, \begin{bmatrix} \Lambda_\alpha & 0 & 0 \\ 0 & \Lambda_\gamma & 0 \\ 0 & 0 & 0 \end{bmatrix},$$

where Λ_α and Λ_γ are diagonal matrices with diagonal elements $\lambda_i, i \in \alpha$ and $\lambda_i, i \in \gamma$, respectively. Let $\kappa := \alpha \cup \gamma$. Define two diagonal matrices of order $|\,\kappa\,|$:

$$D \,:=\, \begin{bmatrix} \Lambda_\alpha & 0 \\ 0 & \Lambda_\gamma \end{bmatrix}$$

and $|D| = (\,D^2\,)^{1/2}$, i.e.,

$$|\,D\,| \,=\, \begin{bmatrix} \Lambda_\alpha & 0 \\ 0 & |\,\Lambda_\gamma\,| \end{bmatrix}.$$

LEMMA 2.3. *For* $(\varepsilon, X) \in \mathbb{R} \times \mathcal{S}$, *the following statements hold.*

(a) *If $\varepsilon^2 I + X^2$ is nonsingular, then Φ is continuously differentiable at (ε, X) and $J\Phi(\varepsilon, X)$ satisfies the following equation:*

$$J\Phi(\varepsilon, X)(\tau, H) = L^{-1}_{\Phi(\varepsilon,X)}(L_X(H) + 2\varepsilon\tau I) \quad \forall\, (\tau, H) \in \mathbb{R} \times \mathcal{S}. \tag{2.5}$$

In particular, in this case,

$$\| J\Phi(\varepsilon, X)(\tau, H) \| \le \sqrt{n}\, |\, \tau\, | + \| H \|. \tag{2.6}$$

(b) *Φ is globally Lipschitz continuous and for any (ε, X), $(\tau, Y) \in \mathbb{R} \times \mathcal{S}$,*

$$\| \Phi(\varepsilon, X) - \Phi(\tau, Y) \| \le \sqrt{n}\, |\, \varepsilon - \tau\, | + \| X - Y \|. \tag{2.7}$$

(c) *Φ is directionally differentiable at $(0, X)$ and for $(\tau, H) \in \mathbb{R} \times \mathcal{S}$,*

$$\Phi'((0, X); (\tau, H)) = P \begin{bmatrix} L^{-1}_{|D|}[D\widetilde{H}_{\kappa\kappa} + \widetilde{H}_{\kappa\kappa} D] & |D|^{-1} D\widetilde{H}_{\kappa\beta} \\ \widetilde{H}^T_{\kappa\beta} D|D|^{-1} & (\, \tau^2 I + \widetilde{H}^2_{\beta\beta}\,)^{1/2} \end{bmatrix} P^T,$$

where $\widetilde{H} := P^T H P$.

(d) *Φ is differentiable at (ε, X) if and only if $\varepsilon^2 I + X^2$ is nonsingular.*

Proof. (a) For any $C \succ 0$, we have, by applying [35, Lemma 6.2] or direct calculation, that $(\, C^2 + W\,)^{1/2} - C = L_C^{-1}(W) + o(\|W\|)$ for all $W \in \mathcal{S}$ sufficiently small. Then, for $\varepsilon^2 I + X^2$ nonsingular (and hence positive definite), we have that

$$\begin{aligned} \Phi(\varepsilon + \tau, X + H) - \Phi(\varepsilon, H) &= (C^2 + W)^{1/2} - C \\ &= L_C^{-1}\left(L_X(H) + 2\varepsilon\tau I\right) + O\left(\tau^2 + \|H\|^2\right) + o(\|W\|), \end{aligned}$$

where $(\tau, H) \in \mathbb{R} \times \mathcal{S}$, $C := \Phi(\varepsilon, X)$, and $W := L_X(H) + 2\varepsilon\tau I + \tau^2 I + H^2$. Thus, Φ is differentiable at (ε, X) and

$$J\Phi(\varepsilon, X)(\tau, H) = L_C^{-1}\left(L_X(H) + 2\varepsilon\tau I\right).$$

By noting the fact that for all $(\varepsilon+\tau, X+H)$ sufficiently close to (ε, X), $\Phi(\varepsilon+\tau, X+H)$ is positive definite, from the definition of L_Φ^{-1} we know that L_Φ^{-1} is continuous at (ε, X). Hence, Φ is continuously differentiable at (ε, X).

Let P and Λ be defined as in (2.4). To prove (2.6), we first note that

$$L_X(H) + 2\varepsilon\tau I = P\left(L_\Lambda(P^T H P) + 2\varepsilon\tau I\right) P^T,$$

and for any $Y \in \mathcal{S}$,

$$L_C^{-1}(Y) = P L^{-1}_{\Phi(\varepsilon,\Lambda)}(P^T Y P) P^T.$$

Thus, we have

$$P^T J\Phi(\varepsilon, X)(\tau, H) P = L^{-1}_{\Phi(\varepsilon,\Lambda)}\left(L_\Lambda(P^T H P) + 2\varepsilon\tau I\right).$$

Hence, by direct calculation, for $i, j = 1, \ldots, n$,

$$(P^T J\Phi(\varepsilon, X)(\tau, H) P)_{ij} = \begin{cases} (P^T H P)_{ij}(\lambda_i + \lambda_j)\left(\sqrt{\varepsilon^2 + \lambda_i^2} + \sqrt{\varepsilon^2 + \lambda_j^2}\right)^{-1} & \text{if } i \ne j, \\ \left(\lambda_i (P^T H P)_{ii} + \varepsilon\tau\right)\left(\varepsilon^2 + \lambda_i^2\right)^{-1/2} & \text{otherwise,} \end{cases}$$

which implies that

$$\sum_{i,j=1}^{n} \left(\left(P^T J\Phi(\varepsilon, X)(\tau, H)P \right)_{ij} \right)^2 \leq n\tau^2 + \sum_{i,j=1}^{n} \left((P^T HP)_{ij} \right)^2 .$$

Hence,

$$\begin{aligned} \| J\Phi(\varepsilon, X)(\tau, H) \|^2 &= \| P^T J\Phi(\varepsilon, X)(\tau, H)P \|^2 \\ &\leq n\tau^2 + \| P^T HP \|^2 = n\tau^2 + \| H \|^2 . \end{aligned}$$

This completes the proof of part (a).

(b) By part (a) of this lemma, for $\varepsilon \neq 0$ and $\tau \neq 0$ we have that

$$\begin{aligned} & \| \Phi(\varepsilon, X) - \Phi(\tau, Y) \| \\ &= \| \Phi(|\varepsilon|, X) - \Phi(|\tau|, Y) \| \\ &= \left\| \int_0^1 J\Phi(|\tau| + t(|\varepsilon| - |\tau|), Y + t(X - Y))(|\varepsilon| - |\tau|, X - Y)dt \right\| \\ &\leq \sqrt{n}\, |\, (|\varepsilon| - |\tau|)\, | + \| X - Y \| \\ &\leq \sqrt{n}\, |\varepsilon - \tau| + \| X - Y \| . \end{aligned}$$

By a limiting process the above inequality is also true for $\varepsilon\tau = 0$. Hence, (2.7) holds.

(c) Let P and Λ be defined as in (2.4). For any $\tau \in \mathbb{R}$, $H \in \mathcal{S}$, and $t \in [0, \infty)$, let

$$\Delta(t) := \Phi(t\tau, X + tH) - \Phi(0, X)$$

and

$$\widetilde{\Delta}(t) := P^T \Delta(t) P .$$

Then,

$$\begin{aligned} \widetilde{\Delta}(t) &= P^T \Phi(t\tau, X + tH)P - P^T \Phi(0, X)P \\ &= \left(t^2\tau^2 I + (P^T (X + tH)P)^2 \right)^{1/2} - |\, P^T XP \,| \\ &= \left(t^2\tau^2 I + (\, P^T XP + tP^T HP\,)^2 \right)^{1/2} - |\, P^T XP \,| \\ &= \left(t^2\tau^2 I + (\Lambda + t\widetilde{H})^2 \right)^{1/2} - |\, \Lambda \,| , \end{aligned}$$

where $\widetilde{H} := P^T HP$. Thus,

$$\widetilde{\Delta}(t) = \left(|\Lambda|^2 + \widetilde{W} \right)^{1/2} - |\, \Lambda \,| ,$$

where

$$\widetilde{W} := t^2\tau^2 I + t\Lambda\widetilde{H} + t\widetilde{H}\Lambda + t^2\widetilde{H}^2$$

and

$$|\Lambda| = \begin{bmatrix} |D| & 0 \\ 0 & 0 \end{bmatrix} .$$

After simple computations we have that

$$\widetilde{W} = t\begin{bmatrix} D\widetilde{H}_{\kappa\kappa} + \widetilde{H}_{\kappa\kappa}D & D\widetilde{H}_{\kappa\beta} \\ \widetilde{H}_{\kappa\beta}^T D & 0 \end{bmatrix} + \begin{bmatrix} O(t^2) & O(t^2) \\ O(t^2) & t^2\tau^2 I + t^2[\widetilde{H}_{\kappa\beta}^T\widetilde{H}_{\kappa\beta} + \widetilde{H}_{\beta\beta}^2] \end{bmatrix}. \tag{2.8}$$

By Lemma 6.2 in Tseng [35], we have that

$$\widetilde{\Delta}(t)_{\kappa\kappa} = L_{|D|}^{-1}(\widetilde{W}_{\kappa\kappa}) + o(\|\widetilde{W}\|), \tag{2.9}$$

$$\widetilde{\Delta}(t)_{\kappa\beta} = |D|^{-1}\widetilde{W}_{\kappa\beta} + o(\|\widetilde{W}\|), \tag{2.10}$$

and

$$\widetilde{W}_{\beta\beta} = \widetilde{\Delta}(t)_{\kappa\beta}^T\widetilde{\Delta}(t)_{\kappa\beta} + \widetilde{\Delta}(t)_{\beta\beta}^2\,. \tag{2.11}$$

Hence,

$$\widetilde{\Delta}(t)_{\kappa\beta} = t|D|^{-1}D\widetilde{H}_{\kappa\beta} + o(t)\,, \tag{2.12}$$

which implies that

$$\widetilde{\Delta}(t)_{\kappa\beta}^T\widetilde{\Delta}(t)_{\kappa\beta} = t^2\widetilde{H}_{\kappa\beta}^T(|D|^{-1}D)^2\widetilde{H}_{\kappa\beta} + o(t^2) = t^2\widetilde{H}_{\kappa\beta}^T\widetilde{H}_{\kappa\beta} + o(t^2)\,. \tag{2.13}$$

According to (2.9) and (2.8),

$$\widetilde{\Delta}(t)_{\kappa\kappa} = tL_{|D|}^{-1}(D\widetilde{H}_{\kappa\kappa} + \widetilde{H}_{\kappa\kappa}D) + o(t)\,. \tag{2.14}$$

Since

$$\widetilde{W}_{\beta\beta} = t^2\tau^2 I + t^2[\widetilde{H}_{\kappa\beta}^T\widetilde{H}_{\kappa\beta} + \widetilde{H}_{\beta\beta}^2]\,,$$

from (2.11) and (2.13), we obtain that

$$\widetilde{\Delta}(t)_{\beta\beta}^2 = t^2\tau^2 I + t^2\widetilde{H}_{\beta\beta}^2 + o(t^2)\,. \tag{2.15}$$

Furthermore, since $\widetilde{\Delta}(\tau)_{\beta\beta}$ is positive semidefinite (see the definition of $\widetilde{\Delta}(t)$), we know from (2.15) that $\widetilde{\Delta}(t)_{\beta\beta}$ is well defined and

$$\widetilde{\Delta}(t)_{\beta\beta} = t\left(\tau^2 I + \widetilde{H}_{\beta\beta}^2 + o(1)\right)^{1/2}. \tag{2.16}$$

Hence, from (2.14), (2.12), (2.16), and the continuity of $(\cdot)^{1/2}$,

$$\lim_{t\downarrow 0}\frac{\widetilde{\Delta}(t)}{t} = \begin{bmatrix} L_{|D|}^{-1}[D\widetilde{H}_{\kappa\kappa} + \widetilde{H}_{\kappa\kappa}D] & |D|^{-1}D\widetilde{H}_{\kappa\beta} \\ \widetilde{H}_{\kappa\beta}^T D|D|^{-1} & (\,\tau^2 I + \widetilde{H}_{\beta\beta}^2\,)^{1/2} \end{bmatrix},$$

which completes the proof of part(c).

(d) Only the "only if" part needs a proof. Obviously $\varepsilon^2 I + X^2$ is nonsingular at $\varepsilon \neq 0$. If Φ is differentiable at $(0, X)$, then part (c) of this lemma shows that $\Phi'((0,X);(\tau,H))$ is a linear function of (τ, H) only if $\beta = \emptyset$; i.e., only if X is nonsingular. □

Lemma 2.3 shows that the squared smoothing matrix function Φ is directionally differentiable everywhere and globally Lipschitz continuous. It also shows that it is differentiable at $(\varepsilon, X) \in \mathbb{R} \times \mathcal{S}$ if and only if $\varepsilon^2 I + X^2$ is nonsingular.

The next result is vital in order to prove the strong semismoothness of Φ. By noting the fact that I and X can be simultaneously diagonalized, we may extend the proof used in [32, Lemma 4.12] from $|X|$ to Φ. Here we follow the outline of a simpler proof given in [5, Proposition 4.10].

LEMMA 2.4. *Let $X \in \mathcal{S}$. Then, for any $\tau \in \mathbb{R}$ and $H \in \mathcal{S}$ such that $\tau^2 I + (X+H)^2$ is nonsingular, Φ is differentiable at $(\tau, X+H)$ and*

$$\Phi(\tau, X+H) - \Phi(0,X) - J\Phi(\tau, X+H)(\tau, H) \;=\; O(\|\,\Delta Z\,\|^2)\,, \tag{2.17}$$

where $\Delta Z := (\tau, H)$.

Proof. Let $\mathcal{D}$ denote the space of $n \times n$ real diagonal matrices with nonincreasing diagonal entries. For each $Y \in \mathcal{S}$, define

$$\mathcal{O}_Y \;:=\; \{\, P \in \mathcal{O} \;:\; P^T Y P \;\in\; \mathcal{D}\,\},$$

where $\mathcal{O} := \{\, P \in \mathbb{R}^{n\times n} \;:\; P^T P = I\,\}$.

Let $\lambda_1 \geq \cdots \geq \lambda_n$ denote the eigenvalues of X. By [6, Lemma 3] or [33, Proposition 4.4], there exist scalars $\eta > 0$ and $\rho > 0$ such that

$$\min_{P \in \mathcal{O}_X} \|\,P - Q\,\| \;\leq\; \eta\,\|\,Y - X\,\| \quad \text{whenever} \;\; Y \;\in\; \mathcal{S},\; \|\,Y - X\,\| \;\leq\; \rho,\; Q \;\in\; \mathcal{O}_Y\,.$$

If $\tau = 0$, then the left-hand side of (2.17) reduces to $\Psi(X+H) - \Psi(X) - J\Psi(X+H)(H)$, where for each $Y \in \mathcal{S}$, $\Psi(Y) := |Y|$. Then, it follows from [32, Lemma 4.12] that (2.17) holds.

Suppose $\tau \neq 0$. Let $\mu_1 \geq \cdots \geq \mu_n$ denote the eigenvalues of $X + H$, and choose any $Q \in \mathcal{O}_{X+H}$. Then, by (2.18), there exists $P \in \mathcal{O}_X$ satisfying

$$\|\,P - Q\,\| \;\leq\; \eta \|\,H\,\|\,.$$

For simplicity, let R denote the left-hand side of (2.17), i.e.,

$$R \;:=\; \Phi(\tau, X+H) - \Phi(0,X) - J\Phi(\tau, X+H)(\tau, H)\,.$$

Letting $C := \Phi(\tau, X+H) = (\,\tau^2 I + (X+H)^2\,)^{1/2}$ and noting that $Q \in \mathcal{O}_C$, we obtain from Lemma 2.3 and the formula for L_C^{-1} given in [35, Page 171] that

$$\begin{aligned} J\Phi(\tau, X+H)(\tau, H) &= L_C^{-1}[(X+H)H + H(X+H) + 2\tau^2 I] \\ &= Q[\Xi \circ (Q^T((X+H)H + H(X+H))Q + 2\tau^2 I)]Q^T, \end{aligned}$$

where the matrix $\Xi \in \mathcal{S}$ has entries

$$\Xi_{ij} \;=\; 1/(\theta_i + \theta_j)$$

and $\theta_i = \sqrt{\tau^2 + \mu_i^2}$ is the ith eigenvalue of C. Then, letting $\widetilde{R} := Q^T R Q$ and $\widetilde{H} := Q^T H Q$, we have that

$$\widetilde{R} \;=\; \Sigma - S^T \Lambda S - \Xi \circ (U + 2\tau^2 I)\,, \tag{2.18}$$

where $\Sigma := \mathrm{diag}\,(\sqrt{\tau^2+\mu_1^2},\ldots,\sqrt{\tau^2+\mu_n^2})$, $\Lambda := \mathrm{diag}\,(\lambda_1,\ldots,\lambda_n)$, $S := P^TQ$, and $U_{ij} := (\mu_i+\mu_j)\widetilde{H}_{ij}$ for all i,j.

Since $\mathrm{diag}\,(\mu_1,\ldots,\mu_n) = Q^T(X+H)Q = S^T\,\mathrm{diag}\,(\lambda_1,\ldots,\lambda_n)S+\widetilde{H}$, we have that

$$\sum_{k=1}^{n} S_{ki}S_{kj}\lambda_k + \widetilde{H}_{ij} \;=\; \begin{cases} \mu_i & \text{if } i=j \\ 0 & \text{else,} \end{cases} \qquad i,j=1,\ldots,n\,. \tag{2.19}$$

Since $S = P^TQ = (P-Q)^TQ + I$ and $\|P-Q\| \le \eta\|H\|$, it follows that

$$S_{ij} \;=\; O(\|\,H\,\|) \quad \forall\; i \neq j\,. \tag{2.20}$$

Since $P, Q \in \mathcal{O}$, we have $S \in \mathcal{O}$ so that $S^TS = I$. This implies

$$1 \;=\; S_{ii}^2 \;+\; \sum_{k\neq i} S_{ki}^2 \;=\; S_{ii}^2 \;+\; O(\|\,H\,\|^2), \quad i=1,\ldots,n, \tag{2.21}$$

and

$$\begin{aligned} 0 \quad &= \quad S_{ii}S_{ij} \;+\; S_{ij}S_{jj} + \sum_{k\neq i,j} S_{ki}S_{kj} \\ &= \quad S_{ii}S_{ij} \;+\; S_{ji}S_{jj} + O(\|\,H\,\|^2) \quad \forall\; i \neq j\,. \end{aligned} \tag{2.22}$$

We now show that $\widetilde{R} = O(\|\,\Delta Z\,\|^2)$, which, by $\|R\| = \|\widetilde{R}\|$, would prove (2.17). For any $i \in \{1,\ldots,n\}$, we have from (2.18) and (2.19) that

$$\begin{aligned} \widetilde{R}_{ii} &= \sqrt{\tau^2+\mu_i^2} - \sum_{k=1}^{n} S_{ki}^2|\lambda_k| - \frac{1}{2\theta_i}(2\tau^2 + 2\mu_i\widetilde{H}_{ii}) \\ &= \sqrt{\tau^2+\mu_i^2} - \sum_{k=1}^{n} S_{ki}^2|\lambda_k| - \frac{\tau^2}{\theta_i} - \frac{\mu_i}{\theta_i}\left(\mu_i - \sum_{k=1}^{n} S_{ki}^2\lambda_k\right) \\ &= \sqrt{\tau^2+\mu_i^2} - S_{ii}^2|\lambda_i| - \frac{\tau^2}{\theta_i} - \frac{\mu_i}{\theta_i}(\mu_i - S_{ii}^2\lambda_i) + O(\|H\|^2) \\ &= \sqrt{\tau^2+\mu_i^2} - (1+O(\|H\|^2))|\lambda_i| - \frac{\tau^2}{\theta_i} - \frac{\mu_i}{\theta_i}(\mu_i - (1+O(\|H\|^2))\lambda_i) + O(\|H\|^2) \\ &= \sqrt{\tau^2+\mu_i^2} - |\lambda_i| - \frac{\tau^2}{\theta_i} - \frac{\mu_i}{\theta_i}(\mu_i - \lambda_i) + O(\|H\|^2) \\ &= f(\tau,\mu_i) - f(0,\lambda_i) - Jf(\tau,\mu_i)(\tau,\mu_i-\lambda_i) + O(\|H\|^2), \end{aligned} \tag{2.23}$$

where the third and fifth equalities use (2.20), (2.21), and the fact that $|\mu_i/\theta_i| \le 1$. The last equality follows by defining $f(\tau,\mu) := \sqrt{\tau^2+\mu^2}$. Since f is known to be strongly semismooth and, by a result of Weyl [2, page 63],

$$|\,\mu_i - \lambda_i\,| \;\le\; \|\,H\,\| \quad \forall\; i, \tag{2.24}$$

the right-hand side of (2.23) is $O(\|\Delta Z\|)^2$. For any $i,j \in \{1,\ldots,n\}$ with $i \neq j$, we have from (2.18) and (2.19) that

$$\widetilde{R}_{ij} \;= -\sum_{k=1}^{n} S_{ki}S_{kj}|\lambda_k| - \Xi_{ij}(\mu_i+\mu_j)\widetilde{H}_{ij}$$

$$
\begin{aligned}
&= -\sum_{k=1}^{n} S_{ki}S_{kj}|\lambda_k| + \Xi_{ij}(\mu_i+\mu_j)\sum_{k=1}^{n} S_{ki}S_{kj}\lambda_k \\
&= -(S_{ii}S_{ij}|\lambda_i| + S_{ji}S_{jj}|\lambda_j|) + \Xi_{ij}(\mu_i+\mu_j)(S_{ii}S_{ij}\lambda_i + S_{ji}S_{jj}\lambda_j) + O(\|H\|^2) \\
&= -\left((S_{ii}S_{ij}+S_{ji}S_{jj}|\lambda_i|) + S_{ji}S_{jj}(|\lambda_j|-|\lambda_i|)\right) \\
&\quad + \Xi_{ij}(\mu_i+\mu_j)\left((S_{ii}S_{ij}+S_{ji}S_{jj})\lambda_i + S_{ji}S_{jj}(\lambda_j-\lambda_i)\right) + O(\|H\|^2) \\
&= -S_{ji}S_{jj}\left(|\lambda_j|-|\lambda_i| - \Xi_{ij}(\mu_i+\mu_j)(\lambda_j-\lambda_i)\right) + O(\|H\|^2) \\
&= -S_{ji}S_{jj}\left(|\lambda_j|-|\lambda_i| - \frac{\mu_j+\mu_i}{\theta_j+\theta_i}(\lambda_j-\lambda_i)\right) + O(\|H\|^2), \qquad (2.25)
\end{aligned}
$$

where the third and fifth equalities use (2.20), (2.22), and $\Xi_{ij}|\mu_i+\mu_j| \le 1$. We have that

$$
\begin{aligned}
&|\,\lambda_j\,| - |\,\lambda_i\,| - \frac{\mu_j+\mu_i}{\theta_j+\theta_i}(\lambda_j-\lambda_i) \\
&= |\,\lambda_j\,| - |\,\lambda_i\,| - \frac{\mu_j+\mu_i}{\theta_j+\theta_i}(\mu_j-\mu_i) - \frac{\mu_j+\mu_i}{\theta_j+\theta_i}(\lambda_j-\mu_j+\mu_i-\lambda_i) \\
&= |\,\lambda_j\,| - |\,\lambda_i\,| - \frac{(\tau^2+\mu_j^2)-(\tau^2+\mu_i^2)}{\sqrt{\tau^2+\mu_j^2}+\sqrt{\tau^2+\mu_i^2}} - \frac{\mu_j+\mu_i}{\theta_j+\theta_i}(\lambda_j-\mu_j+\mu_i-\lambda_i) \\
&= |\,\lambda_j\,| - |\,\lambda_i\,| - \left(\sqrt{\tau^2+\mu_j^2}-\sqrt{\tau^2+\mu_i^2}\right) - \frac{\mu_j+\mu_i}{\theta_j+\theta_i}(\lambda_j-\mu_j+\mu_i-\lambda_i). \qquad (2.26)
\end{aligned}
$$

Since $|\mu_j+\mu_i|/(\theta_j+\theta_i) \le 1$ and $|\,|\,\lambda_k\,| - \sqrt{\tau^2+\mu_k^2}\,| = |\,\|(0,\lambda_k)\| - \|(\tau,\mu_k)\|\,| \le |\,(0,\lambda_k)-(\tau,\mu_k)\,| \le |\tau| + |\lambda_k-\mu_k|$ for $k \in \{i,j\}$, we see from (2.24) that the right-hand side of (2.26) is $O(|\tau|+\|H\|)$. This, together with (2.20), implies the right-hand side of (2.25) is $O(\|H\|(|\tau|+\|H\|))$. The proof is completed. □

According to Theorem 2.2 and Lemmas 2.3 and 2.4, we obtain the following main result of this section.

THEOREM 2.5. *The squared smoothing matrix function* Φ *is strongly semismooth at* $(0,X) \in \mathbb{R}\times\mathcal{S}$.

The theorem above provides a basis for quadratic convergence of the squared smoothing Newton method for the SDCP, which is to be discussed in section 5.

3. Properties of the B-subdifferential of Φ. In this section, we shall discuss some properties of the B-subdifferential of the squared smoothing function Φ at $(0,X) \in \mathbb{R}\times\mathcal{S}$. These properties play a key role in the proof of nonsingularity of the Jacobians arising from the SDP and the SDCP. Assume that X has the eigen-decomposition as in (2.4), i.e.,

$$X \;=\; P\Lambda P^T,$$

where P is an orthogonal matrix and Λ is the diagonal matrix of eigenvalues of X and has the form

$$\Lambda \;=\; \begin{bmatrix} \Lambda_\alpha & 0 & 0 \\ 0 & \Lambda_\gamma & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$

Partition the orthogonal matrix P according to

$$P = [\, W_\alpha \; W_\gamma \; W_\beta \,],$$

with $W_\alpha \in \mathbb{R}^{n\times|\alpha|}$, $W_\gamma \in \mathbb{R}^{n\times|\gamma|}$, and $W_\beta \in \mathbb{R}^{n\times|\beta|}$.

Recall that the critical cone of $\mathcal{S}_+ := \{\, X \succeq 0 : \; X \in \mathcal{S} \,\}$ at $X \in \mathcal{S}$ is defined as

$$\mathcal{C}(X;\mathcal{S}_+) \; := \mathcal{T}(X_+;\mathcal{S}_+) \cap (\, X_+ - X \,)^\perp,$$

where $\mathcal{T}(X_+;\mathcal{S}_+)$ is the tangent cone of $\mathcal{S}_+$ at X_+ and $(X_+ - X)^\perp$ is the subset of matrices in $\mathcal{S}$ that are orthogonal to $(X_+ - X)$ under the matrix Frobenius inner product. The critical cone can be completely described [3, 9] by

$$\mathcal{C}(X;\mathcal{S}_+) = \{\, Y \in \mathcal{S} : \, W_\gamma^T Y W_\gamma = 0,\, W_\gamma^T Y W_\beta = 0,\, W_\beta^T Y W_\beta \succeq 0 \,\}. \tag{3.1}$$

Consequently, the affine hull of $\mathcal{C}(X;\mathcal{S}_+)$, which we denote by $\mathcal{L}(X;\mathcal{S}_+)$, is the linear subspace

$$\{\, Y \in \mathcal{S} : \, W_\gamma^T Y W_\gamma = 0,\, W_\gamma^T Y W_\beta = 0 \,\}.$$

PROPOSITION 3.1. *For any $(0,H) \in \mathbb{R}\times\mathcal{S}$ and $V \in \partial_B\Phi(0,X)$, it holds that*

$$V(0,H) = P\,(\Omega \circ P^T H P)P^T, \tag{3.2}$$

$$H + V(0,H) \in \mathcal{L}(X;\mathcal{S}_+), \tag{3.3}$$

and

$$[H - V(0,H)] \bullet [H + V(0,H)] \geq 0\,, \tag{3.4}$$

where the matrix $\Omega \in \mathcal{S}$ has entries

$$\Omega_{ij} = \begin{cases} t \in [-1,1] & \text{if } (i,j) \in \beta\times\beta, \\ \dfrac{\lambda_i + \lambda_j}{|\lambda_i| + |\lambda_j|} & \text{otherwise.} \end{cases}$$

Proof. Let $V \in \partial_B\Phi(0,X)$. By Lemma 2.3 and the definition of the elements in $\partial_B\Phi(0,X)$, it follows that there exists a sequence $\{(\varepsilon^k, X^k)\}$ converging to $(0,X)$ with $(\varepsilon^k)^2 I + (X^k)^2$ being nonsingular such that

$$V(0,H) = \lim_{k\to\infty} J\Phi(\varepsilon^k, X^k)(0,H) = \lim_{k\to\infty} L_{C^k}^{-1}\,(L_{X^k}(H))\,,$$

where $C^k := \Phi(\varepsilon^k, X^k)$. Let $X^k = P^k\,\Lambda^k\,(P^k)^T$ be the orthogonal decomposition of X^k, where Λ^k is the diagonal matrix of eigenvalues of X^k and P^k is a corresponding orthogonal matrix. Without loss of generality, by taking subsequences if necessary, we may assume that $\{P^k\}$ is a convergent sequence with limit $P = \lim_{k\to\infty} P^k$ and $\Lambda = \lim_{k\to\infty}\Lambda^k$ (clearly $X = P\Lambda P^T$). Then,

$$\lim_{k\to\infty} \Lambda_\beta^k = 0\,.$$

For any $H \in \mathcal{S}$ with $\widetilde{H}^k := (P^k)^T H P^k$, we have that

$$L_{C^k}\left(J\Phi(\varepsilon^k, X^k)(0,H)\right) = L_{X^k}(H)\,;$$

i.e.,

$$\left((\varepsilon^k)^2 I + (\Lambda^k)^2 \right)^{1/2} \widetilde{U}^k + \widetilde{U}^k \left((\varepsilon^k)^2 I + (\Lambda^k)^2 \right)^{1/2} = \Lambda^k \widetilde{H}^k + \widetilde{H}^k \Lambda^k ,$$

where $\widetilde{U}^k := (P^k)^T \left[J\Phi(\varepsilon^k, X^k)(0, H) \right] P^k$. By denoting $\widetilde{C}^k := ((\varepsilon^k)^2 I + (\Lambda^k)^2)^{1/2}$, we have that

$$\begin{bmatrix} \widetilde{C}^k_{\alpha\alpha}\widetilde{U}^k_{\alpha\alpha} + \widetilde{U}^k_{\alpha\alpha}\widetilde{C}^k_{\alpha\alpha} & \widetilde{C}^k_{\alpha\alpha}\widetilde{U}^k_{\alpha\gamma} + \widetilde{U}^k_{\alpha\gamma}\widetilde{C}^k_{\gamma\gamma} & \widetilde{C}^k_{\alpha\alpha}\widetilde{U}^k_{\alpha\beta} + \widetilde{U}^k_{\alpha\beta}\widetilde{C}^k_{\beta\beta} \\ \widetilde{C}^k_{\gamma\gamma}\widetilde{U}^k_{\gamma\alpha} + \widetilde{U}^k_{\gamma\alpha}\widetilde{C}^k_{\alpha\alpha} & \widetilde{C}^k_{\gamma\gamma}\widetilde{U}^k_{\gamma\gamma} + \widetilde{U}^k_{\gamma\gamma}\widetilde{C}^k_{\gamma\gamma} & \widetilde{C}^k_{\gamma\gamma}\widetilde{U}^k_{\gamma\beta} + \widetilde{U}^k_{\gamma\beta}\widetilde{C}^k_{\beta\beta} \\ \widetilde{C}^k_{\beta\beta}\widetilde{U}^k_{\beta\alpha} + \widetilde{U}^k_{\beta\alpha}\widetilde{C}^k_{\alpha\alpha} & \widetilde{C}^k_{\beta\beta}\widetilde{U}^k_{\beta\gamma} + \widetilde{U}^k_{\beta\gamma}\widetilde{C}^k_{\gamma\gamma} & \widetilde{C}^k_{\beta\beta}\widetilde{U}^k_{\beta\beta} + \widetilde{U}^k_{\beta\beta}\widetilde{C}^k_{\beta\beta} \end{bmatrix}$$

$$= \begin{bmatrix} \Lambda^k_\alpha \widetilde{H}^k_{\alpha\alpha} + \widetilde{H}^k_{\alpha\alpha}\Lambda^k_\alpha & \Lambda^k_\alpha \widetilde{H}^k_{\alpha\gamma} + \widetilde{H}^k_{\alpha\gamma}\Lambda^k_\gamma & \Lambda^k_\alpha \widetilde{H}^k_{\alpha\beta} + \widetilde{H}^k_{\alpha\beta}\Lambda^k_\beta \\ \Lambda^k_\gamma \widetilde{H}^k_{\gamma\alpha} + \widetilde{H}^k_{\gamma\alpha}\Lambda^k_\alpha & \Lambda^k_\gamma \widetilde{H}^k_{\gamma\gamma} + \widetilde{H}^k_{\gamma\gamma}\Lambda^k_\gamma & \Lambda^k_\gamma \widetilde{H}^k_{\gamma\beta} + \widetilde{H}^k_{\gamma\beta}\Lambda^k_\beta \\ \Lambda^k_\beta \widetilde{H}^k_{\beta\alpha} + \widetilde{H}^k_{\beta\alpha}\Lambda^k_\alpha & \Lambda^k_\beta \widetilde{H}^k_{\beta\gamma} + \widetilde{H}^k_{\beta\gamma}\Lambda^k_\gamma & \Lambda^k_\beta \widetilde{H}^k_{\beta\beta} + \widetilde{H}^k_{\beta\beta}\Lambda^k_\beta \end{bmatrix}.$$

For each k, define the matrix $\Omega^k \in \mathcal{S}$ with entries

$$\Omega^k_{ij} = \left(\sqrt{(\varepsilon^k)^2 + (\lambda^k_i)^2} + \sqrt{(\varepsilon^k)^2 + (\lambda^k_j)^2} \right)^{-1} (\lambda^k_i + \lambda^k_j), \quad i, j = 1, \ldots, n .$$

Since $\{\Omega^k\}$ is bounded, by taking a subsequence if necessary, we assume that $\{\Omega^k\}$ is a convergent sequence and that

$$\lim_{k\to\infty} \Omega^k = \Omega .$$

Hence, it follows that

$$\lim_{k\to\infty} \widetilde{U}^k = \lim_{k\to\infty} \Omega^k \circ \widetilde{H}^k = \Omega \circ P^T H P ,$$

which proves (3.2). Let $\widetilde{H} := P^T H P$. Then, we obtain that

$$P^T V(0, H) P = \begin{bmatrix} \widetilde{H}_{\alpha\alpha} & \Omega_{\alpha\gamma} \circ \widetilde{H}_{\alpha\gamma} & \widetilde{H}_{\alpha\beta} \\ \widetilde{H}^T_{\alpha\gamma} \circ \Omega^T_{\alpha\gamma} & -\widetilde{H}_{\gamma\gamma} & -\widetilde{H}_{\gamma\beta} \\ \widetilde{H}^T_{\alpha\beta} & -\widetilde{H}^T_{\gamma\beta} & \Omega_{\beta\beta} \circ \widetilde{H}_{\beta\beta} \end{bmatrix}.$$

Let $\mathbf{E} \in \mathcal{S}$ be the matrix whose entries are all ones. Thus,

$$P^T[H + V(0, H)]P$$

$$(3.5) = \begin{bmatrix} 2\widetilde{H}_{\alpha\alpha} & (\Omega_{\alpha\gamma} + \mathbf{E}_{\alpha\gamma}) \circ \widetilde{H}_{\alpha\gamma} & 2\widetilde{H}_{\alpha\beta} \\ \widetilde{H}^T_{\alpha\gamma} \circ (\Omega_{\alpha\gamma} + \mathbf{E}_{\alpha\gamma})^T & 0 & 0 \\ 2\widetilde{H}^T_{\alpha\beta} & 0 & (\Omega_{\beta\beta} + \mathbf{E}_{\beta\beta}) \circ \widetilde{H}_{\beta\beta} \end{bmatrix}$$

and

$$P^T[H - V(0,H)]P$$

$$(3.6) \;=\; \begin{bmatrix} 0 & (\mathbf{E}_{\alpha\gamma} - \Omega_{\alpha\gamma}) \circ \widetilde{H}_{\alpha\gamma} & 0 \\ \widetilde{H}^T_{\alpha\gamma} \circ (\mathbf{E}_{\alpha\gamma} - \Omega_{\alpha\gamma})^T & 2\widetilde{H}_{\gamma\gamma} & 2\widetilde{H}_{\gamma\beta} \\ 0 & 2\widetilde{H}^T_{\gamma\beta} & (\mathbf{E}_{\beta\beta} - \Omega_{\beta\beta}) \circ \widetilde{H}_{\beta\beta} \end{bmatrix}.$$

Hence, from (3.5), we get that

$$W^T_\gamma[H + V(0,H)]W_\gamma = 0 \quad \text{and} \quad W^T_\gamma[H + V(0,H)]W_\beta = 0\,,$$

which proves (3.3).

By noting the fact that $\Omega_{ij} \in [-1,1]$ for all $i,j = 1,\ldots,n$, from (3.5) and (3.6) we obtain that

$$\begin{aligned} & [H - V(0,H)] \bullet [H + V(0,H)] \\ &= \big(P^T[H - V(0,H)]P\big) \bullet \big(P^T[H + V(0,H)]P\big) \\ &= \sum_{i\in\alpha,\, j\in\gamma} 2(1-\Omega_{ij})(1+\Omega_{ij})\widetilde{H}^2_{ij} + \sum_{i\in\beta,\, j\in\beta} (1-\Omega_{ij})(1+\Omega_{ij})\widetilde{H}^2_{ij} \\ &\geq 0\,, \end{aligned}$$

which proves (3.4). This completes the proof. □

4. The squared smoothing Newton method. Let $\Psi : \mathcal{S}(n_1,\ldots,n_m) \to \mathcal{S}(n_1,\ldots,n_m)$ be locally Lipschitz continuous. Let $G : \mathbb{R}\times\mathcal{S}(n_1,\ldots,n_m) \to \mathcal{S}(n_1,\ldots,n_m)$ be an approximate function of Ψ such that G is continuously differentiable at $(\varepsilon, X) \in \mathbb{R}\times\mathcal{S}(n_1,\ldots,n_m)$ unless $\varepsilon = 0$ and

$$\lim_{(\varepsilon,Y)\to(0,X)} G(\varepsilon,Y) \;=\; \Psi(X)\,.$$

The existence of such a G was proved in [31] for vector-valued functions. It can be easily extended to matrix-valued functions by making use of the isometry between $\mathbb{R}^n$ and $\mathcal{S}(n_1,\ldots,n_m)$. For the SDP and the SDCP, there are many choices for G. In particular, a computationally efficient form for the SDCP is

$$(4.1) \qquad G(\varepsilon,X) := \; X \;-\; [X - F(X) \;+\; \Phi(\varepsilon, X - F(X))]\,/2\,.$$

The *squared smoothing Newton method*, in particular, solves the auxiliary equation

$$(4.2) \qquad E(\varepsilon,X) := \begin{bmatrix} \varepsilon \\ G(\varepsilon,X) \end{bmatrix} = 0$$

and uses the merit function $\phi(Z) := \varepsilon^2 + \|G(Z)\|^2$ for the line search, where $Z := (\varepsilon, X)$.

Let $\bar\varepsilon \in \mathbb{R}_{++}$ and $\eta \in (0,1)$ be such that $\eta\bar\varepsilon < 1$. Define an auxiliary point $\bar Z$ by

$$\bar Z := \; (\bar\varepsilon, 0) \;\in\; \mathbb{R}\times\mathcal{S}(n_1,\ldots,n_m)$$

and $\theta : \mathbb{R}\times\mathcal{S}(n_1,\ldots,n_m) \mapsto \mathbb{R}_+$ by

$$\theta(Z) := \; \eta\min\{1, \phi(Z)\}\,.$$

Let

$$\mathcal{N} := \{\, Z = (\varepsilon, X) \in \mathbb{R} \times \mathcal{S}(n_1, \ldots, n_m) \,:\, \varepsilon \geq \theta(Z)\bar{\varepsilon} \,\}\,.$$

ALGORITHM 4.1.

Step 0. Select constants $\delta \in (0,1)$ and $\sigma \in (0, 1/2)$. Let $\varepsilon^0 := \bar{\varepsilon}$, $X^0 \in \mathcal{S}(n_1, \ldots, n_m)$ be an arbitrary point and $k := 0$.

Step 1. If $E(Z^k) = 0$, then stop. Otherwise, let $\theta_k := \theta(Z^k)$.

Step 2. Compute $\Delta Z^k := (\Delta\varepsilon^k, \Delta X^k) \in \mathbb{R} \times \mathcal{S}(n_1, \ldots, n_m)$ by

$$E(Z^k) \,+\, JE(Z^k)(\Delta Z^k) \,=\, \theta_k \bar{Z}\,. \tag{4.3}$$

Step 3. Let l_k be the smallest nonnegative integer l satisfying

$$\phi(Z^k + \delta^l \Delta Z^k) \,\leq\, [\,1 - 2\sigma(1 - \eta\bar{\varepsilon})\delta^l\,]\phi(Z^k)\,. \tag{4.4}$$

Define $Z^{k+1} := Z^k + \delta^{l_k}\Delta Z^k$.

Step 4. Replace k by $k+1$ and go to Step 1.

THEOREM 4.2. *Assume that*

(i) *for every $k \geq 0$, if $\varepsilon^k \in \mathbb{R}_{++}$ and $Z^k \in \mathcal{N}$, then $JE(Z^k)$ is nonsingular; and*

(ii) *for any accumulation point $Z^* = (\varepsilon^*, X^*)$ of $\{Z^k\}$, if $\varepsilon^* > 0$ and $Z^* \in \mathcal{N}$, then $JE(Z^*)$ is nonsingular.*

Then an infinite sequence $\{Z^k\} \subset \mathcal{N}$ is generated by Algorithm 4.1 *and each accumulation point Z^* of $\{Z^k\}$ is a solution of $E(Z) = 0$. Moreover, if E is strongly semismooth at Z^* and if all $V \in \partial_B E(Z^*)$ are nonsingular, then the whole sequence $\{Z^k\}$ converges to Z^*,*

$$\|\, Z^{k+1} - Z^* \,\| \,=\, O(\|\, Z^k - Z^* \,\|^2), \tag{4.5}$$

and

$$\varepsilon^{k+1} = O((\varepsilon^k)^2)\,. \tag{4.6}$$

The vector version of the above convergence result is proved in [26], where the smoothing parameter is a vector rather than a scalar. However, the proof was independent of the dimension of the parameter vector. Therefore, with a slight revision if necessary, its matrix version can be established similarly. For brevity we omit the proof.

The key conditions for quadratic convergence of Algorithm 4.1 are: (a) the strong semismoothness of the smoothing function E and (b) the nonsingularity of all $V \in \partial_B E(Z^*)$. (In [26], $\partial E(Z^*)$, rather than $\partial_B E(Z^*)$, was used. However, it is easy to check whether the convergence properties are still valid if we replace $\partial E(Z^*)$ by $\partial_B E(Z^*)$ in the analysis.) In the subsequent sections we will provide sufficient conditions for (b) to hold in the cases of SDP and SDCP where (a) is naturally implied by the strong semismoothness of Φ.

5. Application to the SDP. In this section we shall show how to use Algorithm 4.1 to solve (1.4), which constitutes the optimality conditions of the SDP. For this purpose, we assume that $\{A_i\}_{i=1}^m$ are linearly independent, i.e., any $\alpha \in \mathbb{R}^m$ satisfying $\sum_{i=1}^m \alpha_i A_i = 0$ implies $\alpha_i = 0$, $i = 1, \ldots, m$.

Define $\mathcal{A} : \mathcal{S} \to \mathbb{R}^m$ as

$$\mathcal{A}(X) := \begin{bmatrix} A_1 \bullet X \\ \vdots \\ A_m \bullet X \end{bmatrix}, \quad X \in \mathcal{S}\,.$$

Then solving (1.4) is equivalent to finding a solution to

$$\Psi(X,y,S) := \begin{bmatrix} \mathcal{A}(X) - b \\ \sum_{i=1}^{m} y_i A_i + S - C \\ X - [X-S]_+ \end{bmatrix} = 0, \quad (X,y,S) \in \mathcal{S} \times \mathbb{R}^m \times \mathcal{S}\,. \tag{5.1}$$

Define $G : \mathbb{R} \times \mathcal{S} \times \mathbb{R}^m \times \mathcal{S} \to \mathbb{R}^m \times \mathcal{S} \times \mathcal{S}$ as

$$G(\varepsilon,X,y,S) := \begin{bmatrix} \mathcal{A}(X) - b \\ \sum_{i=1}^{m} y_i A_i + S - C \\ X - [X - S + \Phi(\varepsilon, X-S)]/2 \end{bmatrix}. \tag{5.2}$$

Then G is continuously differentiable at (ε,X,y,S) with $\varepsilon \neq 0$. Let

$$E(\varepsilon,X,y,S) := \begin{bmatrix} \varepsilon \\ G(\varepsilon,X,y,S) \end{bmatrix}. \tag{5.3}$$

Hence, finding a solution of $\Psi(X,y,S) = 0$ is equivalent to finding a solution of $E(\varepsilon,X,y,S) = 0$.

Similar smoothing functions for the SDP were first used in [6] and very recently in [15]. Based on these smoothing functions, smoothing Newton methods were also designed in [6, 15]. The major differences between our method and those in [6, 15] in the context of SDP are (i) our algorithm needs to solve only one linear system per iteration while the methods in [6, 15] need to solve two; (ii) quadratic convergence has been established for our algorithm while only superlinear convergence has been established for methods in [6, 15]; and (iii) numerical results are reported in [6, 15] while our paper is focused on theoretical analysis.

The next result shows that $JE(\varepsilon,X,Y,S)$ is nonsingular at $(\varepsilon,X,y,S) \in \mathbb{R} \times \mathcal{S} \times \mathbb{R}^m \times \mathcal{S}$ with $\varepsilon \neq 0$. Similar proofs can be found in [6, 15, 34].

PROPOSITION 5.1. *For any $(\varepsilon,X,y,S) \in \mathbb{R} \times \mathcal{S} \times \mathbb{R}^m \times \mathcal{S}$ with $\varepsilon \neq 0$, $JE(\varepsilon,X,Y,S)$ is nonsingular.*

Proof. By Lemma 2.3, we know that $JE(\varepsilon,X,Y,S)$ exists. Suppose that there exists $(\tau,H,z,T) \in \mathbb{R} \times \mathcal{S} \times \mathbb{R}^m \times \mathcal{S}$ such that

$$JE(\varepsilon,X,Y,S)(\tau,H,z,T) = 0\,;$$

i.e.,

$$\begin{bmatrix} \tau \\ \mathcal{A}(H) \\ \sum_{i=1}^{m} z_i A_i + T \\ H - [H - T + J\Phi(\varepsilon, X-S)(\tau, H-T)]/2 \end{bmatrix} = 0, \tag{5.4}$$

which implies that

$$\tau = 0 \quad \text{and} \quad 2H - [H - T + J\Phi(\varepsilon, X-S)(0, H-T)] = 0\,.$$

Hence, by Lemma 2.3,

$$2H - \left[H - T + L^{-1}_{\Phi(\varepsilon, X-S)} L_{(X-S)}(H-T)\right] \,=\, 0\,,$$

which implies that

$$L_{\Phi(\varepsilon, X-S)}(H+T) \,=\, L_{(X-S)}(H-T)\,;$$

i.e.,

$$\begin{aligned}&\left(\varepsilon^2 I + (X-S)^2\right)^{1/2}(H+T) \,+\, (H+T)\left(\varepsilon^2 I + (X-S)^2\right)^{1/2}\\ &=\, (X-S)(H-T) \,+\, (H-T)(X-S)\,.\end{aligned}$$

Since $X - S \in \mathcal{S}$, there exist an orthogonal matrix P and a diagonal matrix Λ of eigenvalues of $X - S$ such that

$$X - S \,=\, P\Lambda P^T.$$

By denoting $\widetilde{H} := P^T H P$ and $\widetilde{T} := P^T T P$, we have that

$$(\,\varepsilon^2 I + \Lambda^2\,)^{1/2}\,(\widetilde{H} + \widetilde{T}) \,+\, (\widetilde{H} \,+\, \widetilde{T})(\,\varepsilon^2 I + \Lambda^2\,)^{1/2} \,=\, \Lambda(\widetilde{H} - \widetilde{T}) \,+\, (\widetilde{H} - \widetilde{T})\Lambda\,.$$

Hence,

$$\widetilde{H} + \widetilde{T} \,=\, \Omega \,\circ\, (\widetilde{H} - \widetilde{T})\,,$$

where the matrix $\Omega \in \mathcal{S}$ has entries

$$\Omega_{ij} \,=\, \left(\sqrt{\varepsilon^2 + \lambda_i^2} \,+\, \sqrt{\varepsilon^2 + \lambda_j^2}\right)^{-1}(\lambda_i + \lambda_j), \quad i, j = 1, \ldots, n.$$

Thus,

$$\widetilde{H} \,=\, \widetilde{\Omega} \circ \widetilde{T}\,,$$

where the matrix $\widetilde{\Omega} \in \mathcal{S}$ has entries

$$\widetilde{\Omega}_{ij} \,=\, \left(\lambda_i + \lambda_j - \sqrt{\varepsilon^2 + \lambda_i^2} \,-\, \sqrt{\varepsilon^2 + \lambda_j^2}\right)^{-1}\left(\lambda_i + \lambda_j + \sqrt{\varepsilon^2 + \lambda_i^2} \,+\, \sqrt{\varepsilon^2 + \lambda_j^2}\right),$$

where $i, j = 1, \ldots, n$. From (5.4), we know that

$$A_i \,\bullet\, H \,=\, 0,\ i = 1, \ldots, m, \quad \text{and} \quad \sum_{i=1}^{m} z_i A_i + T \,=\, 0\,,$$

which implies that

$$T \,\bullet\, H \,=\, \sum_{i=1}^{m} z_i A_i \,\bullet\, H + T \,\bullet\, H \,=\, \left(\sum_{i=1}^{m} z_i A_i + T\right) \,\bullet\, H \,=\, 0\,.$$

Hence,

$$0 \,=\, T \,\bullet\, H \,=\, \widetilde{T} \,\bullet\, \widetilde{H} \,=\, \widetilde{T} \,\bullet\, (\widetilde{\Omega} \,\circ\, \widetilde{T})\,,$$

which, together with the fact that $\widetilde{\Omega}_{ij} < 0$ for all i and j, implies that $\widetilde{T} = 0$. Thus,

$$\widetilde{H} \,=\, \widetilde{\Omega} \bullet \widetilde{T} \,=\, 0 \quad \text{and} \quad T \,=\, H \,=\, 0\,.$$

From the linear independence of $\{A_i\}_{i=1}^m$ and that fact $\sum_{i=1}^m z_i A_i + T \,=\, 0$, we can conclude that $z = 0$. This shows that $JE(\varepsilon, X, y, S)$ is nonsingular. □

Proposition 5.1 shows that Algorithm 4.1 is well defined when it is applied to the SDP. We state it formally in the following theorem. Its proof is a direct application of Theorem 4.2 and Proposition 5.1.

THEOREM 5.2. *If Algorithm* 4.1 *is applied to the SDP, then an infinite sequence* $\{Z^k\}$ *is generated and each accumulation point* Z^* *of* $\{Z^k\}$ *is a solution of* $E(Z) = 0$.

For local convergence analysis of Algorithm 4.1 for the SDP, we need the nonsingularity of $\partial_B E(Z^*)$ at a solution Z^* of $E(Z) = 0$. Next, we discuss a sufficient condition to guarantee the nonsingularity of $\partial_B E(Z^*)$ at a *strict complementary and nondegenerate* solution $Z^* = (0, X^*, y^*, S^*)$ of $E(Z) = 0$; i.e., Z^* satisfies the following two conditions: (a) $X^* + S^* \succ 0$ and (b) for any $(H, z, T) \in \mathcal{S} \times \mathbb{R}^m \times \mathcal{S}$ satisfying

$$\mathcal{A}(H) \,=\, 0, \quad \sum_{i=1}^m z_i A_i + T \,=\, 0, \quad \text{and} \quad X^*T \,+\, HS^* \,=\, 0\,,$$

it holds that $H = T = 0$. Condition (a) is called the strict complementarity, under which E is continuously differentiable at Z^*. Condition (b) was first introduced by Kojima, Shida, and Shindoh [16] for local analysis of interior-point methods. Conditions (a) and (b) are also used in noninterior-point methods for solving the SDP [6, 15]. See [1] for a discussion on strict complementarity and nondegeneracy conditions in the SDP.

PROPOSITION 5.3. *Let* $Z^* = (0, X^*, y^*, S^*) \in \mathbb{R} \times \mathcal{S} \times \mathbb{R}^m \times \mathcal{S}$ *be a strict complementary and nondegenerate solution of* $E(Z) = 0$. *Then* $JE(Z^*)$ *is nonsingular.*

Proof. Since (X^*, y^*, S^*) is a solution to the SDP, we have that

$$X^* \succeq 0, \quad S^* \,\succeq\, 0, \quad X^*S^* \,=\, S^*X^* \,=\, 0\,,$$

which implies that there exists an orthogonal matrix P such that

$$X^* \,=\, P\,\Delta\,P^T \quad \text{and} \quad S^* \,=\, P\,\Sigma\,P^T\,,$$

where $\Delta = \,\mathrm{diag}\,(\delta_1, \ldots, \delta_n)$ and $\Sigma = \,\mathrm{diag}\,(\sigma_1, \ldots, \sigma_n)$ are two positive semidefinite diagonal matrices and $\delta_i \sigma_i \,=\, 0$, $i \,=\, 1, \ldots, n$, where $\delta_1, \ldots, \delta_n$ and $\sigma_1, \ldots, \sigma_n$ are eigenvalues of X^* and S^*, respectively. By using the fact that $X^* + S^* \succ 0$, we also have that

$$\delta_i + \sigma_i > 0, \quad i = 1, \ldots, n\,.$$

Denote $\Lambda := \Delta - \Sigma$. Then, $\Lambda = \,\mathrm{diag}\,(\lambda_1, \ldots, \lambda_n)$ is nonsingular and

$$X^* - S^* \,=\, P\,\Lambda\,P^T\,,$$

where $\lambda_i = \delta_i - \sigma_i$, $i = 1, \ldots, n$.

Suppose that there exists $(\tau, H, z, T) \in \mathbb{R} \times \mathcal{S} \times \mathbb{R}^m \times \mathcal{S}$ such that

$$JE(0, X^*, y^*, S^*)(\tau, H, z, T) \,=\, 0\,.$$

We have that $\tau = 0$ and

$$\begin{bmatrix} \mathcal{A}(H) \\ \sum_{i=1}^{m} z_i A_i + T \\ H + T - J\Phi(0, X^* - S^*)(0, H - T) \end{bmatrix} = 0\,. \tag{5.5}$$

In particular, from the third equality of (5.5), we obtain that

$$P^T(H + T)P - P^T J\Phi(0, X^* - S^*)(0, H - T)\, P = 0\,,$$

which, together with Proposition 3.1, implies that

$$\widetilde{H} + \widetilde{T} = P^T J\Phi(0, X^* - S^*)(0, H - T)\, P = \Omega \circ (\widetilde{H} - \widetilde{T})\,,$$

where $\widetilde{H} := P^T H P$, $\widetilde{T} = P^T T P$, and $\Omega \in \mathcal{S}$ has entries

$$\Omega_{ij} = \frac{\lambda_i + \lambda_j}{|\,\lambda_i\,| + |\,\lambda_j\,|}, \quad i, j = 1, \ldots, n\,.$$

Hence,

$$(E - \Omega) \circ \widetilde{H} + \widetilde{T} \circ (E + \Omega) = 0\,, \tag{5.6}$$

where $E \in \mathcal{S}$ denotes the matrix whose entries are all ones. Denote two index sets

$$\alpha := \{\, \lambda_i : \lambda_i > 0 \,\} \quad \text{and} \quad \gamma := \{\, \lambda_i : \lambda_i < 0 \,\}\,.$$

By noting the fact that $\lambda_i = \delta_i$ if $\lambda_i > 0$ and $\lambda_i = -\sigma_i$ if $\lambda_i < 0$ and $\alpha \cup \gamma = \{1, \ldots, n\}$, from (5.6) we have that

$$\widetilde{T}_{ij} = 0 \quad \forall\, (i, j) \in \alpha \times \alpha;$$

$$\widetilde{H}_{ij}\sigma_j + \widetilde{T}_{ij}\delta_i = 0 \quad \forall\, (i, j) \in \alpha \times \gamma$$

and

$$\widetilde{H}_{ij} = 0 \quad \forall\, (i, j) \in \gamma \times \gamma\,.$$

Thus,

$$\Delta\widetilde{T} + \widetilde{H}\Sigma = 0\,;$$

i.e.,

$$X^*T + HS^* = 0,$$

which, together with the first and second equalities of (5.5) and the nondegeneracy assumption at Z^*, shows that

$$H = T = 0.$$

The linear independence of $\{A_i\}_{i=1}^m$ and the fact that $T = 0$ imply $z = 0$. Hence, $JE(Z^*)$ is nonsingular. □

We can now state quadratic convergence of Algorithm 4.1 for solving the SDP, which does not require a proof.

THEOREM 5.4. *If an accumulation point Z^* of $\{Z^k\}$ generated by Algorithm* 4.1 *for solving the SDP is a strict complementary and nondegenerate solution of $E(Z) = 0$, then the whole sequence $\{Z^k\}$ converges to Z^* with*

$$\| Z^{k+1} - Z^* \| = O(\| Z^k - Z^* \|^2) \tag{5.7}$$

and

$$\varepsilon^{k+1} = O((\varepsilon^k)^2)\,. \tag{5.8}$$

In the above theorem for the SDP, we need the nondegeneracy to prove quadratic convergence of Algorithm 4.1. In the next section, we shall show that, for the SDCP, this assumption can be replaced by the positive definiteness of the Jacobian of the problem on a certain subspace.

6. Application to the SDCP. In this section, we shall deduce quadratic convergence of the squared smoothing Newton method in solving the SDCP. We first prove a result on the generalized Jacobian for a composite function.

PROPOSITION 6.1. *Let $\mathcal{S}, \mathcal{S}_1$, and $\mathcal{S}_2$ be symmetric block-diagonal matrix spaces. Let $F : \mathcal{S} \to \mathcal{S}_1$ be continuously differentiable on an open neighborhood $\mathcal{N}$ of $\overline{X}$ and $\Psi : \mathcal{S}_1 \to \mathcal{S}_2$ be locally Lipschitz continuous and semismooth on an open neighborhood of $F(\overline{X})$. Then, for any $H \in \mathcal{S}$, it holds that*

$$\partial_B \Upsilon(\overline{X})(H) \subseteq \partial_B \Psi(F(\overline{X}))\, JF(\overline{X})(H), \tag{6.1}$$

where for any $X \in \mathcal{N}$, $\Upsilon(X) := \Psi(F(X))$.

Proof. Since Υ is locally Lipschitz continuous, by Rademacher's theorem (see [28, page 403]), Υ is differentiable almost everywhere in $\mathcal{N}$. For any $V \in \partial_B \Upsilon(\overline{X})$, there exists a sequence of differentiable points $\{X^k\} \subset \mathcal{N}$ of Υ converging to $\overline{X}$ such that

$$V = \lim_{k\to\infty} J\Upsilon(X^k)\,.$$

Since Ψ is directionally differentiable on an open neighborhood of $F(\overline{X})$, for any $H \in \mathcal{S}$,

$$J\Upsilon(X^k)(H) = \Psi'(F(X^k); JF(X^k)(H))\,.$$

Since Ψ is semismooth at $F(X^k)$, there exists a $W \in \partial_B \Psi(F(X^k))$ such that [25]

$$\Psi'(F(X^k); JF(X^k)(H)) = W\, JF(X^k)(H).$$

Thus,

$$J\Upsilon(X^k)(H) \in \partial_B \Psi(F(X^k))\, JF(X^k)(H)\,,$$

which, together with the upper semicontinuity of ∂_B (see [25]), implies that

$$\lim_{k\to\infty} J\Upsilon(X^k)(H) \in \partial_B \Psi(F(\overline{X}))\, JF(\overline{X})(H)\,.$$

This proves (6.1). ∎

In the following analysis, we assume that $F : \mathcal{S} \to \mathcal{S}$ is continuously differentiable and $E : \mathbb{R} \times \mathcal{S} \to \mathbb{R} \times \mathcal{S}$ is defined as

$$E(\varepsilon, X) \;=\; \begin{bmatrix} \varepsilon \\ G(\varepsilon, X) \end{bmatrix}, \quad (\varepsilon, X) \;\in\; \mathbb{R} \times \mathcal{S}\,, \tag{6.2}$$

where $G : \mathbb{R} \times \mathcal{S} \to \mathcal{S}$ is defined by (4.1); i.e.,

$$G(\varepsilon, X) \;=\; X - [X - F(X) + \Phi(\varepsilon, X - F(X))]\,/2$$

and for any $Y \in \mathcal{S}$,

$$\Phi(\varepsilon, Y) \;=\; (\,\varepsilon^2 I + Y^2\,)^{1/2}\,.$$

Then solving the SDCP is equivalent to solving the following equation:

$$E(\varepsilon, X) \;=\; 0\,. \tag{6.3}$$

The next result is on the nonsingularity of the B-subdifferential of E at $(0, X) \in \mathbb{R} \times S$.

PROPOSITION 6.2. *Suppose that for a given $X \in \mathcal{S}$, the Jacobian $JF(X)$ of F at X is positive definite on the linear subspace $\mathcal{L}(X - F(X); S_+)$, the affine hull of $\mathcal{C}(X - F(X); S_+)$. Then all $U \in \partial_B E(0, X)$ are nonsingular.*

Proof. Let U be an element of $\partial_B E(0, X)$. Assume that $(\tau, H) \in \mathbb{R} \times \mathcal{S}$ is such that $U(\tau, H) = 0$. Then, from the definition of the B-subdifferential of E, we know that $\tau = 0$ and there exists a $W \in \partial_B G(0, X)$ such that $W(0, H) = 0$. By Proposition 6.1, there exists a $V \in \partial_B \Phi(0, X - F(X))$ such that

$$W(0, H) \;=\; H - [H - JF(X)(H) + V(0, H - JF(X)(H))]/2\,,$$

which, together with the fact that $W(0, H) = 0$, implies that

$$2H - [H - JF(X)(H)] - V(0, H - JF(X)(H)) \;=\; 0\,.$$

Let $\overline{H} := H - JF(X)(H)$. We have that

$$2H \;=\; \overline{H} + V(0, \overline{H}) \tag{6.4}$$

and that

$$2\left[\,\overline{H} + JF(X)\left((\overline{H} + V(0, \overline{H}))/2\right)\right] - \overline{H} - V(0, \overline{H}) \;=\; 0\,;$$

i.e.,

$$\overline{H} - V(0, \overline{H}) + JF(X)(\overline{H} + V(0, \overline{H})) \;=\; 0\,,$$

which implies that

$$\begin{aligned} &[\overline{H} + V(0, \overline{H})] \bullet [\overline{H} - V(0, \overline{H})] \\ &\quad + [\overline{H} + V(0, \overline{H})] \bullet [JF(X)(\overline{H} + V(0, \overline{H}))] \;=\; 0\,. \end{aligned} \tag{6.5}$$

By Proposition 3.1, (6.5), and the assumption that $JF(X)$ is positive definite on $\mathcal{L}(X - F(X); S_+)$, we conclude that

$$\overline{H} + V(0, \overline{H}) \;=\; 0\,,$$

which, together with (6.4), implies that $H = 0$. This shows that for any $(\tau, H) \in \mathbb{R} \times \mathcal{S}$

satisfying $U(\tau, H) = 0$, one has $(\tau, H) = 0$. Hence, U is nonsingular. The proof is completed. □

Finally, we can state quadratic convergence of the squared smoothing Newton method for solving the SDCP.

THEOREM 6.3. *Suppose that $F : \mathcal{S} \to \mathcal{S}$ is continuously differentiable on $\mathcal{S}$. Suppose that for each $X \in \mathcal{S}$, $JF(X)$ is positive semidefinite. Then an infinite sequence $\{Z^k\}$ is generated by Algorithm* 4.1 *for solving* (6.3) *and each accumulation point Z^* of $\{Z^k\}$ is a solution of $E(Z) = 0$. Moreover, if $JF(\cdot)$ is Lipschitz continuous around X^* and $JF(X^*)$ is positive definite on the linear subspace $\mathcal{L}(X^* - F(X^*); S_+)$, the affine hull of $\mathcal{C}(X^* - F(X^*); S_+)$, then the whole sequence $\{Z^k\}$ converges to Z^*,*

$$\|Z^{k+1} - Z^*\| = O(\|Z^k - Z^*\|^2), \tag{6.6}$$

and

$$\varepsilon^{k+1} = O((\varepsilon^k)^2). \tag{6.7}$$

Proof. For any $\varepsilon \neq 0$ and $X \in \mathcal{S}$, by Lemma 2.3, E is continuously differentiable at (ε, X). It is easy to check that $JE(\varepsilon, X)$ is nonsingular if and only if $JG(\varepsilon, X)(0, H) = 0$ implies $H = 0$. It has been shown by Chen and Tseng [6] that the latter is true. Thus, for any $\varepsilon \neq 0$ and $X \in \mathcal{S}$, $JE(\varepsilon, X)$ is nonsingular. By Theorem 4.2, an infinite sequence $\{Z^k\}$ is generated by Algorithm 4.1 and each accumulation point Z^* of $\{Z^k\}$ is a solution of $E(Z) = 0$.

If $JF(\cdot)$ is Lipschitz continuous around X^*, then by Theorem 2.5 and a property on the strong semismoothness of a composite function (originally due to Fischer [10]; for the matrix version, see [32, Theorem 3.10]), we know that E is strongly semismooth at $(0, X^*)$. Furthermore, by Proposition 6.2, all $U \in \partial_B E(0, X^*)$ are nonsingular. Thus, by Theorem 4.2, the whole sequence $\{Z^k\}$ converges to Z^*, and (6.6) and (6.7) hold. □

7. Conclusions. We have studied quadratic convergence of a squared smoothing Newton method for nonsmooth matrix equations. For the SDCP, the strong semismoothness of G, together with the positive definiteness of $JF(X^*)$ on the affine hull of $\mathcal{C}(X^* - F(X^*); S_+)$, implies that the proposed algorithm has quadratic rate of convergence without requiring the strict complementarity.

There are several possible directions to extend our work. One direction is to study the strong semismoothness of other smoothing functions used in [6] and then to improve the local analysis in [6]; another direction is to relax the nonsingularity condition on the Jacobians. It is also possible to use some regularization techniques, for example, the Tikhonov-type regularization, to get stronger global convergence results as has been done for vector-valued complementarity problems [8, 30].

Acknowledgments. The authors are grateful to the referees for their very constructive comments. In particular, the present proof of Lemma 2.4 was suggested by a referee.

REFERENCES

[1] F. ALIZADEH, J. -P. HAEBERLY, AND M. L. OVERTON, *Complementarity and nondegeneracy in semidefinite programming*, Math. Programming, Ser. B, 77 (1997), pp. 111–128.
[2] R. BHATIA, *Matrix Analysis*, Springer-Verlag, New York, 1997.
[3] J. F. BONNANS AND A. SHAPIRO, *Perturbation Analysis of Optimization Problems*, Springer-Verlag, New York, 2000.

[4] S. Boyd, L. El Ghaoui, E. Feron, and V. Balakrishnan, *Linear Matrix Inequalities in System and Control Theory*, SIAM Studies in Applied Mathematics 15, Philadelphia, 1994.
[5] X. Chen, H. Qi, and P. Tseng, *Analysis of nonsmooth symmetric-matrix-valued functions with applications to semidefinite complementarity problems*, SIAM J. Optim., 13 (2003), pp. 960–985.
[6] X. Chen and P. Tseng, *Non-interior continuation methods for solving semidefinite complementarity problems*, Math. Program., 95 (2003), pp. 431–474.
[7] B. Fares, D. Noll, and P. Apkarian, *Robust control via sequential semidefinite programming*, SIAM J. Control Optim., 40 (2002), pp. 1791–1820.
[8] F. Facchinei and C. Kanzow, *Beyond monotonicity in regularization methods for nonlinear complementarity problems*, SIAM J. Control Optim., 37 (1999), pp. 1150–1161.
[9] F. Facchinei and J. S. Pang, *Finite-Dimensional Variational Inequalities and Complementarity Problems*, Vols. I and II, Springer-Verlag, New York, 2003.
[10] A. Fischer, *Solution of monotone complementarity problems with locally Lipschitzian functions*, Math. Programming, 76 (1997), pp. 513–532.
[11] M. Fukushima and L. Qi, eds., *Reformulation: Nonsmooth, Piecewise Smooth, Semismooth and Smoothing Methods*, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1999.
[12] M. S. Gowda and T. Parthasarathy, *Complementarity forms of theorems of Lyapunov and Stein, and related results*, Linear Algebra Appl., 320 (2000), pp. 131–144.
[13] M. S. Gowda and Y. Song, *On semidefinite linear complementarity problems*, Math. Program., 88 (2000), pp. 575–587.
[14] F. Jarre, *An interior method for nonconvex semidefinite programs*, Optim. Eng., 1 (2000), pp. 347–372.
[15] C. Kanzow and C. Nagel, *Semidefinite programs: New search directions, smoothing-type methods, and numerical results*, SIAM J. Optim., 13 (2002), pp. 1–23.
[16] M. Kojima, M. Shida, and S. Shindoh, *Local convergence of predictor-corrector infeasible-interior-point algorithm for SDPs and SDLCPs*, Math. Programming, 80 (1998), pp. 129–160.
[17] M. Kojima, M. Shida, and S. Shindoh, *A predictor-corrector interior-point algorithm for the semidefinite linear complementarity problem using the Alizadeh–Haeberly–Overton search direction*, SIAM J. Optim., 9 (1999), pp. 444–465.
[18] M. Kojima, M. Shida, and S. Shindoh, *Search directions in the SDP and the monotone SDLCP: Generalization and inexact computation*, Math. Program., 85 (1999), pp. 51–80.
[19] M. Kojima, S. Shindo, and S. Hara, *Interior-point methods for the monotone semidefinite linear complementarity problem in symmetric matrices*, SIAM J. Optim., 7 (1997), pp. 86–125.
[20] F. Leibfritz and E. M. E. Mostafa, *An interior point constrained trust region method for a special class of nonlinear semidefinite programming problems*, SIAM J. Optim., 12 (2002), pp. 1048–1074.
[21] R. D. C. Monteiro and J. S. Pang, *On two interior-point mappings for nonlinear semidefinite complementarity problems*, Math. Oper. Res., 23 (1998), pp. 39–60.
[22] R. D. C. Monteiro and J.-S. Pang, *A potential reduction Newton method for constrained equations*, SIAM J. Optim., 9 (1999), pp. 729–754.
[23] Y. Nesterov and A. Nemirovskii, *Interior Point Polynomial Algorithms in Convex Programming*, SIAM Studies in Applied Mathematics 13, Philadelphia, 1994.
[24] J. S. Pang, D. Sun, and J. Sun, *Semismooth homeomorphisms and strong stability of semidefinite and Lorentz complementarity problems*, Math. Oper. Res., 28 (2003), pp. 39–63.
[25] L. Qi, *Convergence analysis of some algorithms for solving nonsmooth equations*, Math. Oper. Res., 18 (1993), pp. 227–244.
[26] L. Qi, D. Sun, and G. Zhou, *A new look at smoothing Newton methods for nonlinear complementarity problems and box constrained variational inequalities*, Math. Program., 87 (2000), pp. 1–35.
[27] L. Qi and J. Sun, *A nonsmooth version of Newton's method*, Math. Programming, 58 (1993), pp. 353–367.
[28] R. T. Rockafellar and R. J.-B. Wets, *Variational Analysis*, Springer-Verlag, Berlin, 1998.
[29] A. Shapiro, *First and second order analysis of nonlinear semidefinite programs*, Math. Programming, Ser. B, 77 (1997) pp. 301–320.
[30] D. Sun, *A regularization Newton method for solving nonlinear complementarity problems*, Appl. Math. Optim., 40 (1999), pp. 315–339.
[31] D. Sun and L. Qi, *Solving variational inequality problems via smoothing-nonsmooth reformulations*, J. Comput. Appl. Math., 129 (2001), pp. 37–62.

[32] D. Sun and J. Sun, *Semismooth matrix valued functions*, Math. Oper. Res., 27 (2002), pp. 150–169.

[33] D. Sun and J. Sun, *Strong semismoothness of eigenvalues of symmetric matrices and its application to inverse eigenvalue problems*, SIAM J. Numer. Anal., 40 (2003), pp. 2352–2367.

[34] M. J. Todd, K. C. Toh, and R. H. Tütüncü, *On the Nesterov–Todd direction in semidefinite programming*, SIAM J. Optim., 8 (1998), pp. 769–796.

[35] P. Tseng, *Merit functions for semidefinite complementarity problems*, Math. Programming, 83 (1998), pp. 159–185.

[36] P. Tseng, *Convergent infeasible interior-point trust-region methods for constrained minimization*, SIAM J. Optim., 13 (2002), pp. 432–469.

[37] G. Zhou, D. Sun, and L. Qi, *Numerical experiments for a class of squared smoothing Newton methods for box constrained variational inequality problems*, in Reformulation: Nonsmooth, Piecewise Smooth, Semismooth and Smoothing Methods, M. Fukushima and L. Qi, eds., Kluwer Academic Publishers, Dordrecht, The Netherlands, 1999, pp. 421–441.

SIAM J. OPTIM.
Vol. 14, No. 3, pp. 807–840

CONVERGENCE OF APPROXIMATE AND INCREMENTAL SUBGRADIENT METHODS FOR CONVEX OPTIMIZATION*

KRZYSZTOF C. KIWIEL†

Abstract. We present a unified convergence framework for approximate subgradient methods that covers various stepsize rules (including both diminishing and nonvanishing stepsizes), convergence in objective values, and convergence to a neighborhood of the optimal set. We discuss ways of ensuring the boundedness of the iterates and give efficiency estimates. Our results are extended to incremental subgradient methods for minimizing a sum of convex functions, which have recently been shown to be promising for various large-scale problems, including those arising from Lagrangian relaxation.

Key words. nondifferentiable optimization, convex programming, subgradient optimization, approximate subgradients, efficiency

AMS subject classifications. 65K05, 90C25

DOI. 10.1137/S1052623400376366

1. Introduction.

We are interested in the convex constrained minimization problem

$$f_* := \inf\{\, f(x) : x \in S \,\} \quad \text{with} \quad f := \sum_{i=1}^{m} f_i, \tag{1.1}$$

where $S \neq \emptyset$ is a closed convex set in the Euclidean space $\mathbb{R}^n$ with inner product $\langle \cdot, \cdot \rangle$ and norm $|\cdot|$, and each $f_i : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ is a closed proper convex function finite on S. Let $S_* := \operatorname{Arg\,min}_S f$ denote the *optimal set* of problem (1.1) and $f_S := f + \mathrm{I}_S$ its *extended objective*, where I_S is the *indicator function* of S ($\mathrm{I}_S(x) = 0$ if $x \in S$, ∞ if $x \notin S$). Then $f_* = \inf f_S$ and $S_* = \operatorname{Arg\,min} f_S$; note that f_S is a closed proper convex function.

The *approximate subgradient projection method* generates a sequence $\{x^k\}_{k=1}^{\infty} \subset S$ via

$$x^{k+1} := P_S(x^k - \nu_k g^k), \quad g^k \in \partial_{\epsilon_k} f_S(x^k), \quad k = 1, 2, \dots, \quad x^1 \in S, \tag{1.2}$$

where $P_S x := \arg\min_S |x - \cdot|$ is the *projector* on S, $\nu_k > 0$ is a *stepsize*, and $\epsilon_k \geq 0$ is an error tolerance of an approximate subgradient g^k that belongs to the ϵ_k-subdifferential of f_S at x^k:

$$\partial_{\epsilon_k} f_S(x^k) := \left\{\, g : f_S(x) \geq f_S(x^k) + \left\langle g, x - x^k \right\rangle - \epsilon_k \quad \forall x \,\right\}. \tag{1.3}$$

This method, introduced by Shor [Sho62] and first analyzed in [Erm66, Pol67] has extensive literature; see, e.g., the books [Ber99, BSS93, DeV81, Min86, Nes89, Pol83, Sho79] (and, e.g., [Erm76, MGN87, Nur79] for extensions to stochastic and nonconvex problems). However, most authors tailor their analyses to particular stepsizes, such as $\nu_k := \lambda_k |g^k|^{-1}$ with $\sum_k \lambda_k = \infty$.

*Received by the editors August 4, 2000; accepted for publication (in revised form) July 2, 2002; published electronically March 5, 2004. This research was supported by the State Committee for Scientific Research under grant 8T11A00622.

http://www.siam.org/journals/siopt/14-3/37636.html

†Systems Research Institute, Polish Academy of Sciences, Newelska 6, 01–447 Warsaw, Poland (kiwiel@ibspan.waw.pl).

This paper presents a unified convergence framework for the method (1.2) that covers various stepsize rules (including both diminishing and nonvanishing stepsizes), convergence in the objective values $f(x^k)$, and convergence of $\{x^k\}$ to the optimal set S_* or its neighborhood for nonvanishing stepsizes. We discuss ways of ensuring boundedness of the iterates and give efficiency estimates. Our results subsume those in the literature.

Our analysis extends to the *incremental subgradient projection method* given by

$$x_1^k := x^k, \quad x_{i+1}^k := P_S(x_i^k - \nu_k g_i^k), \quad g_i^k \in \partial_{\epsilon_i^k} f_i^S(x_i^k), \quad i = 1\colon m, \tag{1.4a}$$

$$x^{k+1} := x_{m+1}^k, \tag{1.4b}$$

where $f_i^S := f_i + \mathrm{I}_S$. In other words, subgradient steps are taken for successive objectives f_i of (1.1), hoping that one iteration with m steps should be almost as effective as m ordinary iterations (1.2), although it is much cheaper. This hope is supported by the recent analysis and numerical results of [BTMN01, NeB01], where this version is shown to be promising for certain large-scale problems, including those arising from Lagrangian relaxation. The incremental version stems from [Kib79], but for differentiable problems it is related to backpropagation methods in neural networks; see, e.g., [Ber97, BeT00, Gai94, Gri94, Luo91, LuT94, MaS94].

The paper is organized as follows. In section 2 we recall some elementary results on ergodic convergence and coercivity. General convergence results are given in section 3, and the cases where f_S is coercive or $\{x^k\}$ is bounded are studied in sections 4 and 5. In section 6 we discuss techniques that ensure boundedness of $\{x^k\}$, whereas in section 7 we analyze stepsize rules that do not need such techniques. (Unfortunately, they do not extend to the incremental case.) Efficiency estimates for various stepsizes are given in section 8. Finally, section 9 extends the preceding convergence and efficiency results to the incremental case.

Our notation is fairly standard. $B_\rho := \{x : |x| \le \rho\}$ is the ball with center 0, and radius ρ. $d_C(\cdot) := \inf_{y\in C} |\cdot - y|$ is the distance function of a set $C \subset \mathbb{R}^n$.

2. Technical preliminaries. We present the following three lemmas in order to make the paper more self-contained.

LEMMA 2.1. *Suppose $\nu_k > 0$ and $\nu_{\mathrm{sum}}^k := \sum_{j=1}^k \nu_j \to \infty$ as $k \to \infty$. Given a scalar sequence $\{a_k\}$, let $\bar a_k := \sum_{j=1}^k \nu_j a_j / \nu_{\mathrm{sum}}^k$ for all k. Then $\underline{\lim}_{k\to\infty} a_k \le \underline{\lim}_{k\to\infty} \bar a_k \le \overline{\lim}_{k\to\infty} \bar a_k \le \overline{\lim}_{k\to\infty} a_k$. In particular, if $\lim_{k\to\infty} a_k$ exists, then $\lim_{k\to\infty} \bar a_k = \lim_{k\to\infty} a_k$.*

Proof. To show that $a := \underline{\lim}_k a_k \le \underline{\lim}_k \bar a_k$, suppose $a > -\infty$. For any $\epsilon > 0$, pick $\bar\jmath$ such that $a_j \ge a - \epsilon$ for all $j \ge \bar\jmath$ and $\sum_{j=1}^{\bar\jmath} \nu_j (a_j - a)/\nu_{\mathrm{sum}}^k \ge -\epsilon$ for all $k \ge \bar\jmath$; then

$$\bar a_k - a = \sum_{j=1}^{\bar\jmath} \nu_j (a_j - a)/\nu_{\mathrm{sum}}^k + \sum_{j=\bar\jmath+1}^{k} \nu_j (a_j - a)/\nu_{\mathrm{sum}}^k \ge -\epsilon - \epsilon \sum_{j=\bar\jmath+1}^{k} \nu_j/\nu_{\mathrm{sum}}^k \ge -2\epsilon$$

for all $k \ge \bar\jmath$. Applying this to $b_k := -a_k$, $\bar b_k := -\bar a_k$ gives $-\overline{\lim}_k a_k \le -\overline{\lim}_k \bar a_k$. ∎

LEMMA 2.2 (Silverman–Toeplitz's theorem [DuS88, p. 75]). *Let $a_{kj} \in \mathbb{R}_+$, $j = 1\colon k$, $k = 1, 2, \ldots$, be such that $\sum_{j=1}^k a_{kj} = 1$ for all k, $\lim_{k\to\infty} a_{kj} = 0$ for all*

j (*e.g.*, $a_{kj} = \nu_j/\nu_{\text{sum}}^k$ *as in Lemma* 2.1). *If* $\{u^j\} \subset \mathbb{R}^n$ *is a sequence such that* $\lim_{j\to\infty} u^j = u$, *then* $\lim_{k\to\infty} \sum_{j=1}^k a_{kj} u^j = u$.

LEMMA 2.3. *Let* $\{a_k\}$, $\{b_k\}$, *and* $\{c_k\}$ *be sequences in* $\mathbb{R}_+$ *such that* $a_{k+1} \le a_k(1+b_k) + c_k$ *for* $k = 1, 2, \ldots$, $\sum_{k=1}^\infty b_k < \infty$, $\sum_{k=1}^\infty c_k < \infty$. *Then* $\{a_k\}$ *converges to some* $a_\infty < \infty$.

Proof. See, e.g., [Pol83, Lem. 2.2.2], due to [Gla65]. □

Denote the *trench* (sublevel set) of the extended objective f_S at any level $\alpha \in \mathbb{R}$ by

$$T_\alpha := \{\, x : f_S(x) \le \alpha \,\}. \tag{2.1}$$

Recalling that f_S is closed and convex, note that the following are equivalent: (i) f_S is *coercive*, i.e., $\lim_{|x|\to\infty} f_S(x) = \infty$; (ii) f_S is *level-bounded*; i.e., T_α is bounded for all $\alpha \in \mathbb{R}$; (iii) the optimal set $S_* = \operatorname{Arg\,min} f_S$ is nonempty and bounded [Roc70, Thm. 27.2].

We shall need some elementary properties of the trenches of f_S and their neighborhoods.

LEMMA 2.4. *Suppose that* f_S *is coercive and its trench* T_β *is nonempty for some* $\beta \in \mathbb{R}$.

(i) *For each level* $\alpha \ge \beta$, *let*

$$\rho(\alpha) := \max_{x\in T_\alpha} d_{T_\beta}(x) = \min\{\rho \ge 0 : T_\alpha \subset T_\beta + B_\rho\} \quad \text{and} \quad T_\beta^\alpha := T_\beta + B_{\rho(\alpha)}; \tag{2.2}$$

thus $\rho(\alpha)$ *is the distance between* T_α *and* T_β, *whereas* T_β^α *is the smallest neighborhood of* T_β *containing* T_α, *so that* $T_\beta \subset T_\beta^\alpha \subset T_\beta + B_\rho$ *whenever* $\rho \ge \rho(\alpha)$. *Then* $\lim_{\alpha\downarrow\beta} \rho(\alpha) = 0$.

(ii) *If* f_S *is also continuous on its domain* S (*i.e.*, f *is continuous on* S), *then for every level* $\bar\alpha > \beta$ *there exists a radius* $\bar\rho > 0$ *such that* $S \cap (T_\beta + B_{\bar\rho}) \subset T_{\bar\alpha}$.

Proof. (i) Since f_S is closed and coercive, both T_β and T_α are compact, and $\rho(\alpha)$ is well defined by (2.2) (d_{T_β} is continuous) and nondecreasing (so is T_α by (2.1)). To show that $\lim_{\alpha\downarrow\beta} \rho(\alpha) = 0$ by contradiction, suppose there are sequences $\alpha_i \downarrow \beta$ and $y^i \in T_{\alpha_i}$ such that $d_{T_\beta}(y^i) \ge \rho > 0$. Since $T_{\beta+1}$ is bounded, we may assume without loss of generality that $y^i \to y^\infty$. Then $d_{T_\beta}(y^\infty) \ge \rho$, since d_{T_β} is continuous. However, $f_S(y^i) \le \alpha_i$ gives in the limit $f_S(y^\infty) \le \beta$ (f_S is closed) and hence $y^\infty \in T_\beta$, contradicting $d_{T_\beta}(y^\infty) \ge \rho$.

(ii) Otherwise there are $\rho_i \downarrow 0$, $y^i \in S \cap (T_\beta + B_{\rho_i}) \setminus T_{\bar\alpha}$, $z^i \in T_\beta$ such that $|y^i - z^i| \le \rho_i$. Since T_β is compact, we may assume without loss of generality that $z^i \to z^\infty \in T_\beta$. However, then $y^i \to z^\infty$ (since $|y^i - z^i| \to 0$) with $f_S(y^i) \ge \bar\alpha$ ($y^i \notin T_{\bar\alpha}$) and the continuity of f_S on S imply $f_S(z^\infty) \ge \bar\alpha$, which contradicts $z^\infty \in T_\beta$. □

LEMMA 2.5. *Suppose that* f_S *is coercive,* $\sigma \in [0, \infty)$, *and* $\alpha \in \mathbb{R}$. *Then the set*

$$T_{\alpha,\sigma} := \left\{ x : \sum_{i=1}^m f_i^S(x_i) \le \alpha \text{ for some } x_i \in x + B_\sigma \right\} \tag{2.3}$$

is bounded.

Proof. For each i, let $\hat f_i(x) := \inf_{y\in x+B_\sigma} f_i^S(y) = \inf_y\{f_i^S(y) + \mathrm{I}_{B_\sigma}(x-y)\}$ for all x. Since each f_i^S is closed proper convex, so is $\hat f_i$, and they have the same recession function (cf. [Roc70, Cors. 9.2.1 and 9.2.2]). Hence (cf. [Roc70, Thm. 9.3]) $f_S = \sum_i f_i^S$ and $\hat f := \sum_i \hat f_i$ have a common recession function and a common recession

cone. This cone is null because f_S is coercive, so $\hat{f}$ is coercive (cf. [Roc70, Thms. 8.4 and 8.7]); hence its level set $\{x : \hat{f}(x) \le \alpha\}$ is bounded. This set coincides with $T_{\alpha,\sigma}$, since $\hat{f}_i(x) \le \alpha_i$ iff $f_i^S(x_i) \le \alpha_i$ for some $x_i \in x + B_\sigma$, because f_i^S is closed and the ball $x + B_\sigma$ is compact. □

3. General convergence results. Throughout this section, and in the following sections until section 9, $\{x^k\}$, $\{\nu_k\}$, $\{\epsilon_k\}$, and $\{g^k\}$ denote the sequences involved in the (ordinary) subgradient iteration (1.2).

3.1. Basic estimates. Our convergence analysis hinges on the following three simple estimates.

LEMMA 3.1. *For each x and $k \ge 1$, we have*

$$|x^{k+1} - x|^2 - |x^k - x|^2 \le -2\nu_k \left[f(x^k) - f_S(x) - \epsilon_k - \tfrac{1}{2}|g^k|^2\nu_k \right], \tag{3.1}$$

$$\frac{\sum_{j=1}^k \nu_j f(x^j)}{\sum_{j=1}^k \nu_j} - f_S(x) \le \frac{\frac{1}{2}|x^1 - x|^2 + \sum_{j=1}^k \frac{1}{2}\nu_j^2|g^j|^2 + \sum_{j=1}^k \nu_j\epsilon_j}{\sum_{j=1}^k \nu_j}, \tag{3.2}$$

$$|x^{k+1} - x^k| \le \nu_k|g^k|. \tag{3.3}$$

Proof. Let $x \in S$, $r_k := |x^k - x|$. Using the nonexpansiveness of P_S and (1.2)–(1.3) gives

$$\begin{aligned} r_{k+1}^2 &\le |x^k - \nu_k g^k - x|^2 = r_k^2 - 2\nu_k \langle g^k, x^k - x\rangle + \nu_k^2|g^k|^2 \\ &\le r_k^2 + 2\nu_k \left[f_S(x) - f(x^k) + \epsilon_k \right] + \nu_k^2|g^k|^2, \end{aligned} \tag{3.4}$$

and hence (3.1). Summing up (3.1) yields (3.2). For $f_S(x) = \infty$, (3.1)–(3.2) are trivial. Finally, (3.3) follows from the nonexpansiveness of P_S and the fact that $x^k \in S$ in (1.2). □

Denoting the quantities involved in the basic estimate (3.1) by

$$\gamma_k := \tfrac{1}{2}|g^k|^2\nu_k \quad \text{and} \quad \delta_k := \gamma_k + \epsilon_k, \tag{3.5}$$

we have

$$|x^{k+1} - x|^2 - |x^k - x|^2 \le -2\nu_k \left[f(x^k) - f_S(x) - \delta_k \right] \quad \forall x. \tag{3.6}$$

Thus x^{k+1} becomes closer than x^k to points x such that $f(x^k) > f_S(x) + \delta_k$, and it is easy to see that the standard stepsize condition $\sum_k \nu_k = \infty$ yields $\underline{\lim}_k f(x^k) \le f_S(x) + \delta$ for all x and hence $\underline{\lim}_k f(x^k) \le f_* + \delta$ for $\delta := \overline{\lim}_k \delta_k$. (Of course, additional assumptions are needed to ensure $\delta < \infty$.) In fact, stronger results are derived in the next subsection by employing averages of $\{x^k\}$ and $\{\delta_k\}$ weighted by the stepsizes $\{\nu_k\}$.

3.2. Cesáro averages and ergodic convergence. Employing, as usual, an unbounded *summary stepsize*

$$\nu_{\text{sum}}^k := \sum_{j=1}^k \nu_j \to \infty \quad \text{as } k \to \infty, \tag{3.7}$$

we shall study the *Cesáro averages* of the sequences $\{x^k\}$ and $\{f(x^k)\}$ defined by

$$\bar{x}^k := \sum_{j=1}^{k} \nu_j x^j / \nu_{\text{sum}}^k \quad \text{and} \quad \bar{f}_k := \sum_{j=1}^{k} \nu_j f(x^j)/\nu_{\text{sum}}^k. \tag{3.8}$$

Note that, since $\nu_k > 0$ and $x^k \in S$, for all k, the convexity of f, S, and $|\cdot|$ yields

$$f(\bar{x}^k) \le \bar{f}_k, \quad \bar{x}^k \in S, \quad \text{and} \quad |\bar{x}^k| \le \max\{\, |x^j| : j = 1\colon k \,\}. \tag{3.9}$$

Using the Cesáro averages of the sequences $\{\gamma_k\}$, $\{\epsilon_k\}$, and $\{\delta_k\}$ (cf. (3.5)),

$$\bar{\gamma}_k := \sum_{j=1}^{k} \nu_j \gamma_j/\nu_{\text{sum}}^k, \quad \bar{\epsilon}_k := \sum_{j=1}^{k} \nu_j \epsilon_j/\nu_{\text{sum}}^k, \quad \text{and} \quad \bar{\delta}_k := \sum_{j=1}^{k} \nu_j \delta_j/\nu_{\text{sum}}^k = \bar{\gamma}_k + \bar{\epsilon}_k, \tag{3.10}$$

we may rewrite the estimate (3.2) in the Cesáro average form

$$\bar{f}_k - f_S(x) \le \tfrac{1}{2}|x^1 - x|^2/\nu_{\text{sum}}^k + \bar{\delta}_k \quad \forall x. \tag{3.11}$$

It is convenient to employ the shorthand notation

$$\bar{\gamma}_{\sup} := \overline{\lim_{k\to\infty}}\, \bar{\gamma}_k, \quad \bar{\epsilon}_{\sup} := \overline{\lim_{k\to\infty}}\, \bar{\epsilon}_k, \quad \bar{\delta}_{\sup} := \overline{\lim_{k\to\infty}}\, \bar{\delta}_k, \quad \text{and} \quad \bar{\delta}_{\inf} := \underline{\lim_{k\to\infty}}\, \bar{\delta}_k. \tag{3.12}$$

For each $\delta \ge 0$, denote the set of *δ-optimal points* of problem (1.1) by

$$S_\delta := \{\, x : f_S(x) \le f_* + \delta \,\}. \tag{3.13}$$

We now show that the algorithm attempts asymptotically to find points in the set $S_{\bar{\delta}_{\sup}}$.

THEOREM 3.2. *Assuming $\sum_{k=1}^{\infty} \nu_k = \infty$, define $\bar{\delta}_{\sup}$ and $\bar{\delta}_{\inf}$ by* (3.12) *and* (3.10). *Then we have the following statements:*

(i) $\underline{\lim}_{k\to\infty} f(\bar{x}^k) \le \underline{\lim}_{k\to\infty} \bar{f}_k \le f_* + \bar{\delta}_{\inf}$ *and* $\underline{\lim}_{k\to\infty} f(x^k) \le f_* + \bar{\delta}_{\inf}$.

(ii) $\overline{\lim}_{k\to\infty} f(\bar{x}^k) \le \overline{\lim}_{k\to\infty} \bar{f}_k \le f_* + \bar{\delta}_{\sup}$ *and* $\underline{\lim}_{k\to\infty} f(x^k) \le f_* + \bar{\delta}_{\sup}$.

(iii) *If* $\bar{\delta}_{\sup} = 0$, *then* $f(\bar{x}^k)$, $\bar{f}_k$, *and* $\inf_{l\ge k} f(x^l)$ *converge to* f_* *as* $k \to \infty$.

(iv) *All the cluster points of* $\{\bar{x}^k\}$ (*if any*) *lie in the* $\bar{\delta}_{\sup}$*-optimal set* $S_{\bar{\delta}_{\sup}}$.

(v) *If* $S_* = \emptyset$ *and* $\bar{\delta}_{\sup} = 0$, *then* $|\bar{x}^k| \to \infty$ *and* $\overline{\lim}_{k\to\infty} |x^k| = \infty$.

(vi) $\bar{\delta}_{\sup} \le \bar{\gamma}_{\sup} + \bar{\epsilon}_{\sup}$, $\bar{\delta}_{\sup} \le \overline{\lim}_{k\to\infty} \delta_k$, $\bar{\gamma}_{\sup} \le \overline{\lim}_{k\to\infty} \gamma_k$, *and* $\bar{\epsilon}_{\sup} \le \overline{\lim}_{k\to\infty} \epsilon_k$. *In particular,* $\bar{\gamma}_{\sup} = 0$ *if* $\lim_{k\to\infty} \nu_k |g^k|^2 = 0$ (*e.g.,* $\lim_{k\to\infty} \nu_k = 0$ *and* $\sup_k |g^k| < \infty$). *If* $\nu := \overline{\lim}_{k\to\infty} \nu_k$ *and* $C := \overline{\lim}_{k\to\infty} |g^k|$ *are finite, then* $\bar{\gamma}_{\sup} \le \frac{1}{2} C^2 \nu < \infty$.

Proof. (i) Since by assumption $\nu_{\text{sum}}^k \to \infty$, taking lower limits in (3.11) gives $\underline{\lim}_k \bar{f}_k \le f_S(x) + \bar{\delta}_{\inf}$ for each x, so $f_* := \inf f_S$ yields $\underline{\lim}_k \bar{f}_k \le f_* + \bar{\delta}_{\inf}$. The conclusion follows from the facts that $f(\bar{x}^k) \le \bar{f}_k$ for all k (cf. (3.9)) and $\underline{\lim}_k f(x^k) \le \underline{\lim}_k \bar{f}_k$ (cf. Lemma 2.1).

(ii) Argue as for (i), replacing lower limits by upper limits.

(iii) This follows from (ii), since (cf. (3.8)–(3.9)) $f(x^k), f(\bar{x}^k), \bar{f}_k \ge f_*$.

(iv) If $\{\bar{x}^k\}$ has a cluster point $\bar{x}^\infty$, then $f_S(\bar{x}^\infty) \le f_* + \bar{\delta}_{\sup}$ by (ii), since f_S is closed.

(v) If $|\bar{x}^k| \not\to \infty$, then $\{\bar{x}^k\}$ has a cluster point $\bar{x}^\infty$ in $S_0 = S_*$ by (iv), i.e., $S_* \ne \emptyset$. Hence if $S_* = \emptyset$, then $|\bar{x}^k| \to \infty$, with $|\bar{x}^k| \le \max_{j=1}^k |x^j|$ by (3.9).

(vi) This follows from (3.12), (3.10), (3.5), (3.7), and Lemma 2.1. ∎

Remark 3.3.

(i) Theorem 3.2 implies additional results for the *record* points

$$x^k_{\rm rec} \in \operatorname{Arg}\min_{\{x^j\}_{j=1}^k} f(x^j) \subset S \quad \text{with} \quad f(x^k_{\rm rec}) = \min_{j=1:\,k} f(x^j) \le \bar f_k, \tag{3.14}$$

where the inequality stems from (3.7)–(3.8). Specifically, $x^k_{\rm rec}$ may replace $\bar x^k$ throughout, also with $\bar\delta_{\rm sup}$ replaced by $\bar\delta_{\rm inf}$ in parts (iii)–(v). However, $\bar x^k = (\nu_k x^k + \nu^{k-1}_{\rm sum}\bar x^{k-1})/\nu^k_{\rm sum}$ may be updated at negligible cost *without* evaluating f, in contrast with $x^k_{\rm rec}$.

(ii) Theorem 3.2 handles both diminishing stepsizes ($\nu = 0$ in (vi)) and nonvanishing ones ($\nu > 0$), for which $\nu_k|g^k|^2 \to 0$ is unlikely in the nonsmooth case.

(iii) The second part of Theorem 3.2(ii) subsumes [Ber99, Ex. 6.3.13(a)] (where $\epsilon_k \to \bar\epsilon$ and $\nu_k|g^k|^2 \to 0$ so that $\bar\delta_{\rm sup} = \bar\epsilon$), which in turn generalizes [CoL93, Prop. 1.2] (where $\bar\epsilon = 0$); its first part subsumes [MiU82, Thm. 1] (where $\nu_k \to 0$, $\sup_k|g^k| < \infty$, and $\epsilon_k \equiv 0$).

3.3. Full convergence. To ensure convergence of $\{x^k\}$, we need stronger assumptions (relative to Theorem 3.2).

THEOREM 3.4. *Suppose* $\sum_{k=1}^\infty \nu_k = \infty$, $\sum_{k=1}^\infty \nu_k\delta_k < \infty$ (*cf.* (3.5)). *Then the conclusions of Theorem* 3.2(i–v) *hold with* $\bar\delta_{\rm sup} = 0$, *and the following statements are equivalent*:

(i) *The optimal set* S_* *is nonempty.*

(ii) $\{x^k\}$ *is bounded* (*where "*(i) $\Rightarrow$ (ii)*" does not require* $\sum_k \nu_k = \infty$).

(iii) $\{x^k\}$ *converges to some* $x^\infty \in S_*$.

Finally, if $\{x^k\}$ *converges to a point* x^∞, *then* $\{\bar x^k\}$ *converges to the same point.*

Proof. By (3.7), (3.10), and (3.12), $\sum_k \nu_k\delta_k < \infty$ yields $\bar\delta_{\rm sup} = 0$ for Theorem 3.2.

"(i) $\Rightarrow$ (ii)": Let $x \in S_*$. Then $f_S(x) \le f(x^k)$, so the basic estimate (3.6) yields

$$|x^{k+1} - x|^2 \le |x^k - x|^2 + 2\nu_k\delta_k \quad \forall k.$$

Hence Lemma 2.3 with $b_k := 0$ and $c_k := 2\nu_k\delta_k$ shows that $a_k := |x^k - x|$ converges. Thus $\{x^k\}$ is bounded. "(i) $\Leftarrow$ (ii)": If $\{x^k\}$ is bounded, then it has a cluster point $x^\infty \in S_*$, since $\underline{\lim}_k f_S(x^k) = f_*$ by Theorem 3.2(iii) and f_S is closed.

"(i) $\Rightarrow$ (iii)": As in the proof of "(i) $\Rightarrow$ (ii)", $|x^k - x|$ converges for each $x \in S_*$, and $\{x^k\}$ has a cluster point $x^\infty \in S_*$. Taking $x = x^\infty$, we get $\underline{\lim}_k |x^k - x| = 0$, and then $|x^k - x| \to 0$, i.e., $x^k \to x^\infty$. "(i) $\Leftarrow$ (iii)": The proof is trivial.

Finally, since $\nu^k_{\rm sum} \to \infty$, $x^k \to x^\infty$ yields $\bar x^k := \sum_{j=1}^k \nu_j x^j/\nu^k_{\rm sum} \to x^\infty$ (cf. Lemma 2.2). □

Remark 3.5.

(i) The assumption $\sum_k \nu_k\delta_k < \infty$ of Theorem 3.4 holds if $\sum_k \nu_k^2|g^k|^2 < \infty$ (e.g., $\sum_k \nu_k^2 < \infty$ and $\sup_k|g^k| < \infty$) and $\sum_k \nu_k\epsilon_k < \infty$.

(ii) For $\epsilon_k \equiv 0$, Theorem 3.4 subsumes [Ber99, Ex. 6.3.13(b)] (where the typo $\sum_k \nu_k^2 < \infty$ should be replaced by $\sum_k \nu_k^2|g^k|^2 < \infty$), [Sch83, Thm. on p. 538] (in which the claim $f(x^k) \to f_*$ is *not* proved), and [LPS96, Thm. 2.7] (where $\sum_k \nu_k^2 < \infty$, $\sup_k|g^k| < \infty$); the earliest and much cited [Pol78] result of [Lit68, Thm. 1] (claiming that $\lim_k f(x^k) = f_*$ for $\sum_k \nu_k^2 < \infty$, $\sup_k|g^k| < \infty$) has gaps in its proof, but a result similar to Theorem 3.4 follows from [ErS68] (with $\sum_k \nu_k^2 < \infty$, $\sup_k|g^k| < \infty$). For $S_* \ne \emptyset$ and $\nu_k \to 0$, Theorem 3.4 concerning Theorem 3.2(iv) recovers a part of [NeY78, Thm. (ii)]. Finally, Theorem 3.4 subsumes [LPS00, Thm. 8] (with $\sum_k \nu_k^2 < \infty$, $\sup_k|g^k| < \infty$, $\epsilon_k \to 0$).

For stepsizes such as $\nu_k := k^{-1}$, Theorem 3.4 may seem to require the boundedness of $\{g^k\}$; in fact, the norms $|g^k|$ may grow with x^k, but not too fast, as shown below.

THEOREM 3.6. *Suppose that* $\sum_{k=1}^{\infty} \nu_k = \infty$, $\sum_{k=1}^{\infty} \nu_k^2 < \infty$, $\sum_{k=1}^{\infty} \nu_k \epsilon_k < \infty$, *and the subgradients satisfy the linear growth condition*: *there exists a constant* $c < \infty$ *such that* $|g^k|^2 \le c(1 + |x^k|^2)$ *for all* k. *Then we have the following statements:*

(i) $\underline{\lim}_{k\to\infty} f(x^k) = f_*$.

(ii) *If* $S_* \neq \emptyset$, *then the assumptions of Theorem* 3.4 *are satisfied with* $\sup_k |g^k| < \infty$; *in particular,* $\{x^k\}$ *and* $\{\bar{x}^k\}$ *converge to some* $x^\infty \in S_*$ *and* $\lim_{k\to\infty} f(\bar{x}^k) = f_*$.

Proof. Suppose there exist $x \in S$ and $\bar{k}$ such that $f(x^k) \ge f(x)$ for all $k \ge \bar{k}$. Employing this inequality and the linear growth condition in the basic estimate (3.1), we obtain

$$\begin{aligned} |x^{k+1} - x|^2 &\le |x^k - x|^2 + \nu_k^2 c\left(1 + |x^k|^2\right) + 2\nu_k\epsilon_k - 2\nu_k\left[\, f(x^k) - f(x)\,\right] \\ &\le |x^k - x|^2 + \nu_k^2 c\left(1 + 2|x^k - x|^2 + 2|x|^2\right) + 2\nu_k\epsilon_k \\ &= |x^k - x|^2\left(1 + 2c\nu_k^2\right) + \left[c(1 + 2|x|^2)\nu_k^2 + 2\epsilon_k\nu_k\right], \end{aligned}$$

where we used the facts that $|x^k| \le |x^k - x| + |x|$ and $(a+b)^2 \le 2(a^2 + b^2)$. Hence Lemma 2.3 with $b_k := 1+2c\nu_k^2$ and $c_k := c(1+2|x|^2)\nu_k^2 + 2\epsilon_k\nu_k$ shows that $a_k := |x^k - x|$ converges. Thus $\{x^k\}$ is bounded, and $\sup_k |g^k|^2 \le c(1+\sup_k |x^k|^2) < \infty$ by the linear growth condition. Then $\sum_k \nu_k^2 < \infty$ implies $\sum_k \nu_k^2 |g^k|^2 < \infty$. Thus the assumptions of Theorem 3.4 are met, and Theorem 3.2(iii) yields $\underline{\lim}_k f(x^k) = f_*$. Since $x \in S$ was arbitrary, we obtain $\underline{\lim}_k f(x^k) \le \inf f_S = f_*$, i.e., (i). For (ii), use $x \in S_*$ above and Theorem 3.4. □

Remark 3.7. For $S = \mathbb{R}^n$ and $\epsilon_k \equiv 0$, Theorem 3.6 recovers [PoT73, Thm. 9.1] (in the finite-dimensional deterministic setting); note that in this case $f(x^k) \to f_*$ when $x^k \to x^\infty$ by continuity of f. Again, the earliest result of [Lit68, Thm. 2] has gaps in its proof.

4. Convergence in the coercive case. We now consider the case where "everything is bounded," including the solution set S_* and the algorithmic quantities δ_k and $|x^{k+1} - x^k|$. It turns out that the asymptotic objective accuracy $\delta := \overline{\lim}_k \delta_k$ and steplength $\sigma := \overline{\lim}_k |x^{k+1} - x^k|$ determine the neighborhood S_*^δ of S_* (cf. (4.1)) to which $\{x^k\}$ converges. The size of this neighborhood depends on the asymptotic steplength σ and on the shape of the δ-optimal set S_δ. The Cesáro averages $\{\bar{x}^k\}$ converge to the smaller set S_δ; thus averaging enhances stability.

THEOREM 4.1. *Suppose that* $\sum_{k=1}^{\infty} \nu_k = \infty$, $\delta := \overline{\lim}_{k\to\infty} \delta_k < \infty$, $\sigma := \overline{\lim}_{k\to\infty} |x^{k+1} - x^k| < \infty$, *and* f_S *is coercive. Then we have the following statements:*

(i) $\underline{\lim}_{k\to\infty} d_{S_\delta}(x^k) = 0$ *and* $\{x^k\}$ *has a cluster point in* S_δ. *Further, the assertions of Theorem* 3.2(ii)–(iii) *hold with* $\bar{\delta}_{\sup} \le \delta$.

(ii) $\lim_{k\to\infty} d_{S_*^\delta}(x^k) = 0$, *where* S_*^δ *is the neighborhood of* S_* *defined by* (*cf. Lemma* 2.4(i))

$$S_*^\delta := S_* + B_{\rho_\delta + \sigma} \quad \textit{with} \quad \rho_\delta := \max\{\, d_{S_*}(x) : x \in S_\delta \,\}. \tag{4.1}$$

Thus $\{x^k\}$ *is bounded and its cluster points belong to* S_*^δ.

(iii) $\{\bar{x}^k\}$ *is bounded, its cluster points lie in* S_δ, *and* $\lim_{k\to\infty} d_{S_\delta}(\bar{x}^k) = 0$.

(iv) *In general, for* $\gamma := \overline{\lim}_{k\to\infty} \gamma_k$, $\epsilon := \overline{\lim}_{k\to\infty} \epsilon_k$, $\nu := \overline{\lim}_{k\to\infty} \nu_k$, $C := \overline{\lim}_{k\to\infty} |g^k|$, *and* $\bar{\sigma} := \overline{\lim}_{k\to\infty} \nu_k |g^k|$, *we have* $\delta \le \gamma + \epsilon$, $\gamma \le \frac{1}{2} C^2 \nu$, *and* $\sigma \le \bar{\sigma} \le \min\{C\nu, (2\gamma\nu)^{1/2}\}$. *In particular,* $\gamma = 0$ *if* $\nu = 0$ *and* $C < \infty$, *whereas* $\sigma = 0$ *if* $\bar{\sigma} = 0$ (*e.g.,* $\nu = 0$ *and* $C < \infty$, *or* $\gamma = 0$ *and* $\nu < \infty$).

Proof. First, recall from section 2 that the closedness and coercivity of f_S imply that the sets $S_* \subset S_\delta \subset S_* + B_{\rho_\delta} \subset S_*^\delta$ are nonempty and compact (cf. (2.2), (3.13), and (4.1)).

(i) By our assumptions and Theorem 3.2(vi), $\bar{\delta}_{\sup} \le \delta$. Hence Theorem 3.2(ii) gives $\underline{\lim}_k f_S(x^k) \le f_* + \delta$. Pick a subsequence $\{x^{k_j}\}$ such that $\lim_j f_S(x^{k_j}) = \underline{\lim}_k f_S(x^k)$. Since f_S is coercive, $\{x^{k_j}\}$ is bounded. Assume without loss of generality that $x^{k_j} \to x^\infty$. Then $f_S(x^\infty) \le f_* + \delta$ (f_S is closed) gives $x^\infty \in S_\delta$ (cf. (3.13)), so $d_{S_\delta}(x^{k_j}) \to d_{S_\delta}(x^\infty) = 0$ by continuity of d_{S_δ}. Thus $\underline{\lim}_k d_{S_\delta}(x^k) = 0$.

(ii) Fixing $\rho > 0$, let

$$V_{2\rho} := S_*^\delta + B_{2\rho} = \{\, x : d_{S_*^\delta}(x) \le 2\rho \,\} \tag{4.2}$$

and

$$v_\rho := \min\{\, f_S(x) : d_{S_\delta}(x) \ge \rho \,\} - (\, f_* + \delta \,). \tag{4.3}$$

Since f_S is closed and coercive, whereas d_{S_δ} is continuous, the minimum in (4.3) is attained at some x, and $v_\rho > 0$. (Otherwise $f_S(x) \le f_* + \delta$ would give $x \in S_\delta$ and hence $d_{S_\delta}(x) = 0$, contradicting $\rho > 0$ in (4.3).)

Since $\delta := \overline{\lim}_k \delta_k$ and $\sigma := \overline{\lim}_k |x^{k+1} - x^k|$, there is $k_\rho < \infty$ such that

$$\delta_k \le \delta + v_\rho \quad \text{and} \quad |x^{k+1} - x^k| \le \sigma + \rho \quad \forall k \ge k_\rho. \tag{4.4}$$

Since $\underline{\lim}_k d_{S_\delta}(x^k) = 0$ by (i), there exists $k = k'_\rho \ge k_\rho$ such that $x^k \in S_\delta + B_\rho$; then $S_\delta \subset S_*^\delta$ implies $x^k \in V_{2\rho}$ (cf. (4.2)).

Assuming $x^k \in V_{2\rho}$ for some $k \ge k'_\rho$, we now show that $x^{k+1} \in V_{2\rho}$. If $d_{S_\delta}(x^k) \le \rho$, then from $S_\delta \subset S_* + B_{\rho_\delta}$, (4.1), and the second inequality of (4.4) we get

$$x^{k+1} \in (S_\delta + B_\rho) + B_{\sigma+\rho} \subset S_* + B_{\rho_\delta} + B_{\sigma+2\rho} = (S_* + B_{\rho_\delta+\sigma}) + B_{2\rho} = S_*^\delta + B_{2\rho},$$

so $x^{k+1} \in V_{2\rho}$ (cf. (4.2)). Thus suppose $d_{S_\delta}(x^k) > \rho$. Then, by (4.3),

$$f(x^k) \ge v_\rho + f_* + \delta. \tag{4.5}$$

Next, by (4.1) and (4.2),

$$V_{2\rho} = S_* + B_{\rho_\delta+\sigma+2\rho}, \tag{4.6}$$

so, since $x^k \in V_{2\rho}$, $|x^k - x| \le \rho_\delta + \sigma + 2\rho$ for $x = P_{S_*} x^k$. Using the basic estimate (3.6) with $f_S(x) = f_*$, the bound (4.5), and the first inequality of (4.4) yields

$$|x^{k+1} - x|^2 - |x^k - x|^2 \le -2\nu_k\,[\, v_\rho + \delta - \delta_k \,] \le 0.$$

Thus $|x^{k+1} - x| \le |x^k - x| \le \rho_\delta + \sigma + 2\rho$ with $x \in S_*$, so $x^{k+1} \in V_{2\rho}$ by (4.6).

Therefore, by induction for each $k \ge k'_\rho$, $x^k \in V_{2\rho}$ and hence (cf. (4.2)) $d_{S_*^\delta}(x^k) \le 2\rho$. Since $\rho > 0$ was arbitrary, $d_{S_*^\delta}(x^k) \to 0$. Thus, since S_*^δ is bounded, so is $\{x^k\}$, and its cluster points must lie in S_*^δ because $d_{S_*^\delta}(x^k) \to 0$, $d_{S_*^\delta}$ is continuous and S_*^δ is closed.

(iii) Since $\{x^k\}$ is bounded by (ii), so is $\{\bar{x}^k\}$ by (3.9). Pick $\bar{x}^{k_j}$ such that $\lim_j d_{S_\delta}(\bar{x}^{k_j}) = \overline{\lim}_k d_{S_\delta}(\bar{x}^k)$. Extracting a subsequence if necessary, suppose $\bar{x}^{k_j} \to \bar{x}^\infty$. By Theorem 3.2(iv) with $\bar{\delta}_{\sup} \le \delta$ (cf. the proof of (i)), $\bar{x}^\infty \in S_\delta$. Hence $\lim_j d_{S_\delta}(\bar{x}^{k_j}) = 0$ by the continuity of d_{S_δ}, and thus $\overline{\lim}_k d_{S_\delta}(\bar{x}^k) = 0$.

(iv) Recalling (3.3) and (3.5), use $|x^{k+1}-x^k| \le \nu_k|g^k|$ and $\nu_k^2|g^k|^2 = 2\nu_k\gamma_k$. ∎

COROLLARY 4.2. *Suppose that the sequences $\{\nu_k\}$, $\{|g^k|\}$, and $\{\epsilon_k\}$ are bounded, and the extended objective f_S is coercive. Then the sequence $\{x^k\}$ is bounded.*

Proof. This follows from Theorem 4.1(ii), (iv) if $\sum_k \nu_k = \infty$. Otherwise, i.e., if $\sum_k \nu_k < \infty$, then by summing the inequality $|x^{k+1} - x^k| \le \nu_k|g^k|$ (cf. (3.3)) and using the assumption that $\sup_k |g^k| < \infty$ we get $\sum_k |x^{k+1} - x^k| < \infty$; hence $\{x^k\}$ converges. ∎

Remark 4.3.

(i) Theorem 4.1(ii) may be augmented as follows: (ii$_1$) *if $\delta = \sigma = 0$, then $S_*^\delta = S_\delta = S_*$ and $\lim_{k\to\infty} d_{S_*}(x^k) = 0$*; (ii$_2$) *if f is continuous on S, then $\overline{\lim}_{k\to\infty} f(x^k) \le \max_{S\cap S_*^\delta} f$ (so that $\lim_{k\to\infty} f(x^k) = f_*$ if $\delta = \sigma = 0$)*. Indeed, if $\delta = \sigma = 0$, then $S_\delta = S_*$ by (3.13), $\rho_\delta = 0$, and $S_*^\delta = S_*$ by (4.1), since S_* is closed, whereas if f is continuous on S, then by picking x^{k_j} such that $\lim_j f(x^{k_j}) = \overline{\lim}_k f(x^k)$ and $x^{k_j} \to x^\infty \in S_*^\delta$, from $x^{k_j} \in S$ we get $x^\infty \in S$ (since S is closed) and $\overline{\lim}_{k\to\infty} f(x^k) = f(x^\infty) \le \max_{S\cap S_*^\delta} f$.

(ii) Theorem 4.1(ii) subsumes [LPS00, Thm. 3] (where $\epsilon_k \to 0$, $\nu_k \to 0$, and $\nu_k|g^k|^2 \to 0$), a "stationary" version of [ShW96, Thm. 2.2] (where $\epsilon_k \downarrow 0$, $\nu_k|g^k|^2 \to 0$, $\sup_k \nu_k < \infty$ yield $\delta = \sigma = 0$), [Nur79, Thm. 2.8] (where S is bounded, $\delta = \sigma = 0$) and a convex version of [MGN87, Thm. 9.1] (where $S = \mathbb{R}^n$, $\epsilon_k \equiv 0$, $\nu_k \to 0$). Further, it subsumes [KiA91, Thm. 2] (where $\epsilon_k \to 0$, $\nu_k \to 0$, $\sup_k |g^k| < \infty$); the latter is a (mis)quotation of [NuZ77, Thm. 2], which, however, uses scaled stepsizes (cf. Remark 7.4(ii)).

5. Convergence when the iterates are bounded. We now show that the case where all the algorithmic quantities (i.e., x^k, g^k, ϵ_k, and ν_k) are bounded is analogous to the coercive case analyzed in Theorem 4.1. Only the statement of the following result is fairly complicated, since it does not presume that $S_* \ne \emptyset$.

THEOREM 5.1. *Suppose that $\sum_{k=1}^\infty \nu_k = \infty$, $\nu := \overline{\lim}_{k\to\infty} \nu_k < \infty$, $\epsilon := \overline{\lim}_{k\to\infty} \epsilon_k < \infty$, $\{x^k\}$ is bounded, and $C := \overline{\lim}_{k\to\infty} |g^k| < \infty$. Then $\gamma := \overline{\lim}_{k\to\infty} \gamma_k \le \frac{1}{2}C^2\nu$, $\sigma := \overline{\lim}_{k\to\infty} |x^{k+1} - x^k| \le C\nu$, and $\delta := \overline{\lim}_{k\to\infty} \delta_k \le \gamma + \epsilon$. For any $R \ge \underline{R} := \sup_k |x^k|$, consider the restricted problem*

$$f_*' := \inf f_S' \quad \text{with} \quad f_S' := f_S + \mathrm{I}_{B_R}. \tag{5.1}$$

Let $S' := S \cap B_R$, $S_*' := \operatorname{Arg\,min} f_S'$, $S_\delta' := \{x : f_S'(x) \le f_*' + \delta\}$, and (cf. Lemma 2.4(i))

$$S_*^{\delta'} := S_*' + B_{\rho_\delta' + \sigma} \quad \text{with} \quad \rho_\delta' := \max\left\{ d_{S_*'}(x) : x \in S_\delta' \right\}. \tag{5.2}$$

Then $f_*' \ge f_*$, $S_*' \supseteq S_* \cap B_R$, and $S_\delta' \supseteq S_\delta \cap B_R$, with equalities holding iff $S_* \cap B_R \ne \emptyset$. In fact, if S_* is nonempty and bounded, and R is large enough (e.g., $B_R \supset S_\delta$), then $f_*' = f_*$, $S_*' = S_*$, $S_\delta' = S_\delta$, $\rho_\delta' = \rho_\delta$ (cf. (4.1)), and $S_*^{\delta'} = S_*^\delta \cap B_R$. Moreover, we have the following statements:

(i) *$\underline{\lim}_{k\to\infty} d_{S_\delta}(x^k) = 0$ and $\{x^k\}$ has a cluster point in S_δ. Further, the assertions of Theorem* 3.2(ii)–(iii) *hold with $\bar\delta_{\sup} \le \delta$.*

(ii) *$\lim_{k\to\infty} d_{S_*^{\delta'}}(x^k) = 0$ and the cluster points of $\{x^k\}$ lie in $S_*^{\delta'}$.*

(iii) *$\{\bar x^k\}$ is bounded, its cluster points lie in S_δ, and $\lim_{k\to\infty} d_{S_\delta}(\bar x^k) = 0$.*

(iv) *If $\delta = 0$, then $S_* \ne \emptyset$, and $\lim_{k\to\infty} f(x^k) = f_*$ if f is continuous on S and $\sigma = 0$.*

Proof. By (5.1), f_S' is closed and convex (so are f_S and B_R), proper (since its domain $S' := S \cap B_R$ contains $\{x^k\}$ by the choice of R), and coercive (S' is bounded),

so its optimal set $S'_* \subset B_R$ is nonempty and bounded. Of course, $f'_S \geq f_S$ and f'_S coincides with f_S on B_R. Hence $f'_* \geq f_*$, $S'_* \supseteq S_* \cap B_R$, and $S'_\delta \supseteq S_\delta \cap B_R$ (cf. (3.13)), with equalities holding throughout iff $S_* \cap B_R \neq \emptyset$. Indeed, if $f'_* = f_*$, then $\emptyset \neq S'_* \subset S_* \cap B_R$ and $S'_\delta \subset S_\delta \cap B_R$ from $f'_* + \delta < \infty$; conversely, if $f_S(x) = f_*$ for some $x \in B_R$, then $f_* = f'_S(x) \geq f'_* \geq f_*$ implies $f'_* = f_*$. Similarly, if S_* is nonempty and bounded, then, since S_δ is bounded, we may choose R such that $B_R \supset S_\delta \supset S_*$, in which case $S'_* = S_* \cap B_R = S_*$ and $S'_\delta = S_\delta \cap B_R = S_\delta$, so that $\rho'_\delta = \rho_\delta$ and $S_*^{\delta\prime} = S_*^\delta \cap B_R$ by (4.1) and (5.2).

Next, we may replace S and f_S in (1.2) by $S' := S \cap B_R$ and f'_S, since $\{x^k\} \subset S'$, whereas $g^k \in \partial_{\epsilon_k} f_S(x^k)$ implies $g^k \in \partial_{\epsilon_k} f'_S(x^k)$, using $f'_S(x^k) = f_S(x^k)$ and $f'_S \geq f_S$. Thus the algorithm works as if applied to problem (5.1), for which the assumptions of Theorem 4.1 hold with S_* replaced by S'_* (since $\nu < \infty$ and $C < \infty$). Therefore, the conclusions of Theorems 4.1 and 3.2(ii)–(iii) are valid with f_S replaced by f'_S, f_* by f'_*, etc. In particular, assertion (ii) follows from Theorem 4.1(ii), whereas Theorem 4.1(i), (iii) implies the first part of (i) as well as (iii) with S_δ replaced by S'_δ. For proving (i), (iii), and (iv), note that x^k and $f_S(x^k) = f'_S(x^k)$ are *independent* of R, for $R \geq \underline{R}$.

(i) Theorem 3.2(ii), (vi) with $\bar{\delta}_{\sup} \leq \delta$ gives $\underline{\lim}_k f_S(x^k) \leq f'_* + \delta$, using $f_S(x^k) = f'_S(x^k)$. Pick a subsequence $\{x^{k_j}\}$ such that $\lim_j f_S(x^{k_j}) = \underline{\lim}_k f_S(x^k)$. Since $\{x^{k_j}\}$ is bounded, we may assume that $x^{k_j} \to x^\infty$. Then by the closedness of f_S, $f_S(x^\infty) \leq f'_* + \delta$. Hence $f_S(x^\infty) \leq f_* + \delta$, since (cf. (5.1)) we can make f'_* arbitrarily close to f_* by increasing R. Thus $x^\infty \in S_\delta$ (cf. (3.13)), so $d_{S_\delta}(x^{k_j}) \to d_{S_\delta}(x^\infty) = 0$. By a similar argument, the assertions of Theorem 3.2(ii)–(iii) hold both with f_* replaced by f'_* and in their original form.

(iii) Since $\{x^k\} \subset B_R$, $\{\bar{x}^k\} \subset B_R$ by (3.9). Pick $\bar{x}^{k_j}$ such that $\lim_j d_{S_\delta}(\bar{x}^{k_j}) = \overline{\lim}_k d_{S_\delta}(\bar{x}^k)$. Extracting a subsequence, if necessary, suppose $\bar{x}^{k_j} \to \bar{x}^\infty$. As in the proof of (i), invoking Theorem 3.2(iv) with $\bar{\delta}_{\sup} \leq \delta$ we get $\bar{x}^\infty \in S'_\delta$ and then $\bar{x}^\infty \in S_\delta$. Hence $\lim_j d_{S_\delta}(\bar{x}^{k_j}) = 0$ by the continuity of d_{S_δ}, and thus $\overline{\lim}_k d_{S_\delta}(\bar{x}^k) = 0$.

(iv) If $\delta = 0$, then in the proof of (i) we have $x^\infty \in S_0 = S_*$ (cf. (3.13)), i.e., $S_* \neq \emptyset$. If additionally $\sigma = 0$ and f is continuous on S, then $\lim_k f(x^k) = f'_*$ by (ii) (cf. Remark 5.2(i) below), with $f'_* = f_*$ for R large enough so that $S_* \cap B_R \neq \emptyset$. □

Remark 5.2.

(i) Theorem 5.1(ii) may be augmented as follows: (ii$_1$) *if* $\delta = \sigma = 0$ (*e.g.,* $\nu = \epsilon = 0$), then $S_*^{\delta\prime} = S'_\delta = S'_*$ and $\lim_{k\to\infty} d_{S_*}(x^k) = 0$; (ii$_2$) *if* f *is continuous on* S, *then* $\overline{\lim}_{k\to\infty} f(x^k) \leq \max_{S\cap S_*^{\delta\prime}} f'_S$ (*so that* $\lim_{k\to\infty} f(x^k) = f'_*$ *if* $\delta = \sigma = 0$). Indeed, this follows as in Remark 4.3(i).

(ii) For $S = \mathbb{R}^n$, Theorem 5.1(i)–(ii) subsumes [Nur91, Thms. 2.3 and 2.4] and the results of [Nur82, sect. 6] (where $\nu_k \to 0$, either $\epsilon_k \to 0$ or $\epsilon_k \equiv \epsilon > 0$, $S_* \neq \emptyset$ is assumed *implicitly*, and the proofs are more complicated).

6. Bounding strategies. Our further results require the following definition.

Definition 6.1. *We say that the algorithm employs a* locally bounded oracle *if* $g^k = g(x^k, \epsilon_k)$ *for all* k, *where the mapping* $S \times \mathbb{R}_+ \ni (x, \epsilon) \mapsto g(x, \epsilon) \in \partial_\epsilon f_S(x)$ *is locally bounded* (*bounded on bounded subsets of its domain*).

This concept is quite natural in view of the following comments.

Remark 6.2.

(i) In most applications, one has an oracle (*black box*) that, given $(x, \epsilon) \in S \times \mathbb{R}_+$, delivers an approximate subgradient $g_f(x, \epsilon) \in \partial_\epsilon f(x)$. Recall that for a fixed ϵ, $\partial_\epsilon f(\cdot)$ is locally bounded on S if f is finite on a neighborhood of S, in which

case $\partial_\epsilon f(S)$ is bounded if S is bounded; also $\partial_\epsilon f(S)$ is bounded if f is finite-valued and polyhedral [HUL93, sect. XI.4.1]. In such cases one may use $g := g_f$, since $\partial_\epsilon f(\cdot) \subset \partial_\epsilon f_S(\cdot)$ on S. For some applications [KLL99a, sect. 9.4] one may choose a locally bounded g_f even when $\partial_\epsilon f(\cdot)$ is unbounded.

(ii) To handle the constraint $x \in S$ more efficiently, one may use the subgradient projection techniques of [KiU93], [Kiw96a, sect. 7], and [LPS96, sect. 3]. Thus, for $g_f(x,\epsilon) \in \partial_\epsilon f(x)$, we may let $g(x,\epsilon)$ be the projection of $g_f(x,\epsilon)$ onto the negative of the tangent cone of S at x so that $-g(x,\epsilon)$ is a feasible direction when S is polyhedral; e.g., for $S := \mathbb{R}^n_+$, $g(x,\epsilon)_j = \min\{g_f(x,\epsilon)_j, 0\}$ if $x_j = 0$, $g_f(x,\epsilon)_j$ otherwise. Then $g(x,\epsilon) \in \partial_\epsilon f_S(x)$, and the crucial property $|g(x,\epsilon)| \le |g_f(x,\epsilon)|$ ensures that g is locally bounded if g_f is bounded.

(iii) Note that if a locally bounded oracle is available, then f must be locally Lipschitz continuous on S [KLL99b, Rem. 3.9(ii)].

Of course, for a locally bounded oracle, $\{g^k\}$ is bounded if $\{x^k\}$ and $\{\epsilon_k\}$ are bounded. We now show that if the algorithm starts from any point in a fixed bounded trench of f_S and employs sufficiently small stepsizes and subgradient errors, then $\{x^k\}$ is bounded.

THEOREM 6.3. *Suppose f_S is coercive and the algorithm employs a locally bounded oracle. Fix any point $\bar{x} \in S$ and a bounding tolerance $\bar{\delta} \in (0,\infty)$. Then there exist stepsize and error thresholds $\bar{\nu}_{\max} > 0$ and $\bar{\epsilon}_{\max} > 0$ with the following property: If the algorithm starts from a point $x^1 \in T_{f(\bar{x})}$ (e.g., $x^1 = \bar{x}$) and employs stepsizes $\nu_k \le \bar{\nu}_{\max}$ and errors $\epsilon_k \le \bar{\epsilon}_{\max}$ for all k, then $\{x^k\}$ stays in the bounded trench $T_{f(\bar{x})+\bar{\delta}}$ so that $\{g^k\}$ is bounded.*

Proof. Let $\beta := f(\bar{x})$, $\bar{\alpha} := \beta + \bar{\delta}$. Since the oracle is locally bounded, f_S is continuous on S (cf. Remark 6.2(iii)). By Lemma 2.4(ii), there exists $\bar{\rho} > 0$ such that $S \cap (T_\beta + B_{2\bar{\rho}}) \subset T_{\bar{\alpha}}$, whereas by Lemma 2.4(i) there is $\alpha > \beta$ such that $T^\alpha_\beta \subset T_\beta + B_{\bar{\rho}}$; thus

$$S \cap \left(T^\alpha_\beta + B_{\bar{\rho}} \right) \subset S \cap \left(T_\beta + B_{2\bar{\rho}} \right) \subset T_{\bar{\alpha}}. \tag{6.1}$$

Let

$$\bar{\epsilon}_{\max} := \tfrac{1}{2}(\alpha - \beta), \tag{6.2}$$

$$C := \sup\left\{ |g(x,\epsilon)| : x \in S \cap (T_\beta + B_{2\bar{\rho}}), \epsilon \le \bar{\epsilon}_{\max} \right\}, \tag{6.3}$$

$$\bar{\nu}_{\max} := \min\left\{ \bar{\rho}/C, (\alpha - \beta)/C^2 \right\}. \tag{6.4}$$

Note that $C < \infty$, since T_β is bounded and $\bar{\epsilon}_{\max} < \infty$.

Since $\{x^k\} \subset S$ and $f(x^1) \le f(\bar{x}) =: \beta$, we have $x^1 \in S \cap (T_\beta + B_{2\bar{\rho}})$.

Assuming $x^k \in S \cap (T_\beta + B_{2\bar{\rho}})$ for some $k \ge 1$, we now show that $x^{k+1} \in S \cap (T_\beta + B_{2\bar{\rho}})$. Using the bound $|x^{k+1} - x^k| \le \nu_k |g^k|$ (cf. (3.3)) with $|g^k| = |g(x^k, \epsilon_k)| \le C$ (cf. (6.3)) and $\nu_k \le \bar{\nu}_{\max} \le \bar{\rho}/C$ (cf. (6.4)) gives $|x^{k+1} - x^k| \le \bar{\rho}$. Hence if $x^k \in T_\alpha$, then from $T_\alpha \subset T^\alpha_\beta$ (cf. (2.2)), the first inclusion of (6.1), and the fact that $x^{k+1} \in S$ we get

$$x^{k+1} \in S \cap \left(x^k + B_{\bar{\rho}} \right) \subset S \cap \left(T_\alpha + B_{\bar{\rho}} \right) \subset S \cap \left(T^\alpha_\beta + B_{\bar{\rho}} \right) \subset S \cap \left(T_\beta + B_{2\bar{\rho}} \right).$$

Next, suppose $x^k \notin T_\alpha$, i.e.,

$$f(x^k) > \alpha. \tag{6.5}$$

Since $x^k \in S \cap (T_\beta + B_{2\bar\rho})$, we have $|x^k - x| \le 2\bar\rho$ for $x = P_{T_\beta} x^k$. Next, by (6.2)–(6.4),

$$\epsilon_k \le \bar\epsilon_{\max} \le \tfrac{1}{2}(\alpha - \beta) \quad \text{and} \quad \tfrac{1}{2}|g^k|^2 \nu_k \le \tfrac{1}{2} C^2 \bar\nu_{\max} \le \tfrac{1}{2}(\alpha - \beta). \tag{6.6}$$

Using the estimate (3.1) with $f_S(x) \le \beta$ and the bounds (6.5) and (6.6), we obtain

$$|x^{k+1} - x|^2 - |x^k - x|^2 \le -2\nu_k \left[f(x^k) - f(x) - \epsilon_k - \tfrac{1}{2}|g^k|^2 \nu_k \right] \le 0.$$

Thus $|x^{k+1} - x| \le |x^k - x| \le 2\bar\rho$ with $x \in T_\beta$, so $x^{k+1} \in S \cap (T_\beta + B_{2\bar\rho})$.

Therefore, by induction, for all k we have $x^k \in S \cap (T_\beta + B_{2\bar\rho})$, and hence (cf. (6.3)) $|g^k| \le C$ and (cf. (6.1)) $x^k \in T_{\bar\alpha}$. ∎

In view of Theorem 6.3, we may employ the following *bounding strategy* that generates finitely many restarts indexed by $l = 1, 2, \ldots$. Fixing $\bar x \in S$ and $\bar\delta > 0$, pick positive sequences $\{\nu^l_{\max}\}$ and $\{\epsilon^l_{\max}\}$ such that $\nu^l_{\max} \to 0$ and $\epsilon^l_{\max} \to 0$ if $l \to \infty$. For the current $l \ge 1$, start the algorithm from $\bar x$ (or the best point found so far if $l > 1$), using stepsizes $\nu_k \le \nu^l_{\max}$ and errors $\epsilon_k \le \epsilon^l_{\max}$ until for some k (if any) it is discovered that

$$f(x^k) > f(\bar x) + \bar\delta, \tag{6.7}$$

in which case increase l by 1, restart the algorithm, etc.

A *special case* of the above strategy consists of picking sequences $\nu_k \to 0$ and $\epsilon_k \to 0$, and resetting x^{k+1} to $\bar x$ (or the best point found so far) if (6.7) holds. Ensuring that $\sup_k |g^k| < \infty$, this version meets the assumptions of Theorem 4.1 if $\sum_k \nu_k = \infty$ and of Theorem 3.4 if additionally $\sum_k \nu_k^2 < \infty$ and $\sum_k \nu_k \epsilon_k < \infty$. However, the general version allows us to satisfy the assumptions of Theorem 4.1 with $\overline{\lim}_k \nu_k > 0$ and $\overline{\lim}_k \epsilon_k > 0$.

To avoid calculating $f(x^k)$, the test (6.7) may be replaced by $|x^k| > R$ for R such that $T_{f(\bar x)+\bar\delta} \subset B_R$; this ensures the boundedness of $\{x^k\}$ and $\{g^k\}$ as before. However, finding such R may be difficult, so the following result motivates an alternative bounding strategy.

THEOREM 6.4. *Suppose f_S is coercive and the algorithm employs a locally bounded oracle. Then for each $\beta \in (f_*, \infty)$ and $\bar\epsilon_{\max} \in [0, \infty)$ there exists $\bar\nu_{\max} > 0$ such that if $f_S(x^1) \le \beta$, $\nu_k \le \bar\nu_{\max}$, and $\epsilon_k \le \bar\epsilon_{\max}$ for all k, then $\{x^k\}$ and $\{g^k\}$ are bounded.*

Proof. We show only how to modify the proof of Theorem 6.3. Let $\bar\alpha := \infty$, $\alpha > \beta + 2\bar\epsilon_{\max}$. Invoking Lemma 2.4(i), pick $\bar\rho > 0$ such that $T^\alpha_\beta \subset T_\beta + B_{\bar\rho}$. Then we have (6.1), whereas (6.2) is replaced by $\bar\epsilon_{\max} \le \frac{1}{2}(\alpha - \beta)$; the rest goes on as before. ∎

In view of Theorem 6.4, we may use the following bounding strategy that generates finitely many restarts indexed by $l = 1, 2, \ldots$. Fixing $\bar x \in S$ and $\bar\epsilon_{\max} \ge 0$, pick positive sequences $\nu^l_{\max} \to 0$ and $R_l \to \infty$. For the current $l \ge 1$, start the algorithm from $\bar x$ (or the best point found so far if $l > 1$), using stepsizes $\nu_k \le \nu^l_{\max}$ and errors $\epsilon_k \le \bar\epsilon_{\max}$; if

$$|x^k| > R_l \tag{6.8}$$

for some k, then increase l by 1, restart the algorithm, etc.

The test (6.8) may be replaced by $\max\{|x^k - x^1|, \nu_k |g^k|, |g^k|\} > R_l$.

This strategy also meets the assumptions of Theorem 4.1, if $\sum_k \nu_k = \infty$, and of Theorem 3.4 if additionally $\sum_k \nu_k^2 < \infty$ and $\sum_k \nu_k \epsilon_k < \infty$. Note that, in contrast with (6.7), its resetting test (6.8) does not require calculating $f(x^k)$.

Yet another bounding strategy stems from the following extension of Corollary 4.2.

THEOREM 6.5. *Suppose that* $\hat{\nu} := \sup_k \nu_k$, $\hat{\gamma} := \sup_k \gamma_k$, *and* $\hat{\epsilon} := \sup_k \epsilon_k$ *are finite and* f_S *is coercive. Then* $\{x^k\}$ *is bounded.*

Proof. We show only how to modify the proof of Theorem 6.3. Let $\beta := f(x^1)$, $\bar{\alpha} := \infty$, $\alpha > \beta + 2\max\{\hat{\epsilon}, \hat{\gamma}\}$. Invoking Lemma 2.4(i), pick $\bar{\rho} \geq (2\hat{\gamma}\hat{\nu})^{1/2}$ such that $T_\beta^\alpha \subset T_\beta + B_{\bar{\rho}}$. Then, by (3.3) and (3.5), we have $|x^{k+1} - x^k|^2 \leq \nu_k^2 |g^k|^2 = 2\nu_k\gamma_k$ and hence $|x^{k+1} - x^k| \leq \bar{\rho}$, $\epsilon_k \leq \frac{1}{2}(\alpha - \beta)$ and $\frac{1}{2}|g^k|^2\nu_k \leq \frac{1}{2}(\alpha - \beta)$ as in (6.6); the rest goes on as before. ∎

Theorem 6.5 suggests the following bounding strategy with resets indexed by $l = 1, 2, \ldots$. Fixing $\bar{x} \in S$, $\bar{\epsilon}_{\max} \in [0, \infty)$, and $\gamma_{\max} \in (0, \infty)$, pick a positive sequence $\nu_{\max}^l \to 0$. For the current $l \geq 1$, start the algorithm from $\bar{x}$ (or the best point found so far if $l > 1$), using stepsizes $\nu_k \leq \nu_{\max}^l$ and errors $\epsilon_k \leq \bar{\epsilon}_{\max}$; if $\gamma_k > \gamma_{\max}$ for some k, then increase l by 1, restart the algorithm, etc. Under the assumptions of Theorem 6.4, only finitely many resets occur (otherwise we would have $\hat{G} := \sup_k |g^k| < \infty$ and $\frac{1}{2}\hat{G}^2\nu_{\max}^l > \gamma_{\max}$ at each reset, contradicting $\nu_{\max}^l \to 0$), so Theorem 6.5 implies the boundedness of $\{x^k\}$. (A special case of this strategy consists of using sequences $\nu_k \to 0$ and $\epsilon_k \leq \bar{\epsilon}_{\max}$, and resetting x^{k+1} to x^1 whenever $\gamma_k > \gamma_{\max}$.) Alternatively, the test $\gamma_k > \gamma_{\max}$ may be replaced by $|g^k| > G_l$, where $G_l \to \infty$ as $l \to \infty$ (e.g., $G_{l+1} := \max\{|g^k|, 10G_l\}$).

Remark 6.6. For $S = \mathbb{R}^n$ and $\epsilon_k \equiv 0$, Theorem 6.3 subsumes in the convex case [MGN87, Lem. 9.1] (which employs (6.7) with $\bar{x} = x^1$), whereas Theorem 6.4 subsumes a result of [Sho79, p. 39]. We note that the proof of [MGN87, Lem. 9.1] is quite complicated, whereas that of [Sho79, p. 39] does not extend to the constrained case.

7. Using scaled stepsizes.

7.1. Extension of Ermoliev's framework. We now highlight an idea that is implicit in the pioneering paper of Ermoliev [Erm66, sect. 9]: to ensure convergence, the stepsize ν_k may be chosen as $\nu_k := \lambda_k\mu_k$, where λ_k is fairly arbitrary (e.g., $\lambda_k := k^{-1}$), but μ_k should damp the possible growth of $|g^k|$. We first discuss general conditions on the choice of μ_k and then provide several examples.

THEOREM 7.1. *Suppose that* $\epsilon := \overline{\lim}_{k\to\infty} \epsilon_k < \infty$ *and the algorithm employs stepsizes* $\nu_k := \lambda_k\mu_k$ *with* $\lambda_k > 0$, $\sum_{k=1}^\infty \lambda_k = \infty$, $\lambda := \overline{\lim}_{k\to\infty} \lambda_k < \infty$, *and* $\mu_k > 0$ *such that*

$$\bar{\gamma} := \overline{\lim_{k\to\infty}}\ \tfrac{1}{2}\mu_k|g^k|^2 < \infty, \tag{7.1}$$

$$\underline{\lim_{k\to\infty}}\ \mu_k > 0 \quad \textit{whenever} \quad \{x^k\} \quad \textit{is bounded.} \tag{7.2}$$

Then $\sum_{k=1}^\infty \nu_k = \infty$ *whenever* $\{x^k\}$ *is bounded. Further, we have the following statements:*

(i) $\underline{\lim}_{k\to\infty} f(x^k) \leq f_* + \delta$, *where* $\delta := \overline{\lim}_{k\to\infty} \delta_k \leq \gamma + \epsilon$ *with* $\gamma := \overline{\lim}_{k\to\infty} \gamma_k \leq \bar{\gamma}\lambda$.

(ii) *If* f_S *is coercive and* $\bar{\sigma} := \overline{\lim}_{k\to\infty} \nu_k|g^k|$ *is finite, which holds if*

$$\overline{\lim_{k\to\infty}}\ \mu_k|g^k| < \infty \quad \textit{or} \quad \mu := \overline{\lim_{k\to\infty}}\ \mu_k < \infty, \tag{7.3}$$

then the conclusions of Theorem 4.1 *hold with* $\nu := \overline{\lim}_{k\to\infty} \nu_k \le \lambda\mu$ *and*

$$\sigma := \overline{\lim_{k\to\infty}}\, |x^{k+1} - x^k| \le \bar\sigma \le \lambda \min\left\{ \overline{\lim_{k\to\infty}}\, \mu_k|g^k|, (2\mu\bar\gamma)^{1/2} \right\}. \tag{7.4}$$

(iii) *If additionally* $\sum_{k=1}^\infty \lambda_k^2 < \infty$ *and the assumptions* $\epsilon < \infty$ *and* $\bar\gamma < \infty$ *are replaced by* $\sum_{k=1}^\infty \nu_k\epsilon_k < \infty$ *and* $\sup_k \mu_k|g^k| < \infty$ (*retaining* $\sum_{k=1}^\infty \lambda_k = \infty$ *and* (7.2)) *then we have the following statements:*

(iii$_1$) $\underline{\lim}_{k\to\infty} f(x^k) = f_*$.

(iii$_2$) $S_* \ne \emptyset$ *iff* $\{x^k\}$ *is bounded.*

(iii$_3$) *If* $S_* \ne \emptyset$, *then the assumptions of Theorem* 3.4 *hold*; *in particular,* $\{x^k\}$ *and* $\{\bar x^k\}$ *converge to some* $x^\infty \in S_*$.

Proof. Note that $\sum_k \lambda_k = \infty$ and (7.2) imply $\sum_k \nu_k = \infty$ whenever $\{x^k\}$ is bounded.

(i) For contradiction, suppose there exist $x \in S$, $v > 0$, and k_v such that $f(x^k) \ge f(x) + \delta + v$ for all $k \ge k_v$. Pick $k'_v \ge k_v$ such that $\delta_k \le \delta + v$ for all $k \ge k'_v$. Then (3.6) yields $|x^{k+1} - x| \le |x^k - x|$ for all $k \ge k'_v$. Thus $\{x^k\}$ is bounded, so $\sum_k \nu_k = \infty$. Hence Theorem 3.2(ii), (vi) gives $\bar\delta_{\sup} \le \delta$ and $\underline{\lim}_k f(x^k) \le f_* + \delta$, a contradiction.

(ii) We have $\sigma \le \bar\sigma < \infty$ from $|x^{k+1} - x^k| \le \nu_k|g^k|$ (cf. (3.3)), $\bar\sigma \le \lambda \overline{\lim}_k \mu_k|g^k|$, and $\bar\sigma^2 \le \lambda^2\mu 2\bar\gamma$ by the definitions of $\bar\sigma$, ν_k, λ, $\bar\gamma$, and μ. Using (i) in the proof of Theorem 4.1(i) gives $\underline{\lim}_k d_{S_\delta}(x^k) = 0$. Then the proof of Theorem 4.1(ii) yields the boundedness of $\{x^k\}$, so $\sum_k \nu_k = \infty$. Hence we may invoke Theorem 3.2(ii), (vi) in the proof of Theorem 4.1(i), and Theorem 3.2(iv) in the proof of Theorem 4.1(iii).

(iii) Since $\tilde C := \sup_k \mu_k|g^k| < \infty$, we have $\sum_k \nu_k^2|g^k|^2 \le \tilde C^2 \sum_k \lambda_k^2 < \infty$. (iii$_1$) Suppose $\underline{\lim}_k f(x^k) > f_*$. Thus there are $x \in S$ and $\bar k$ such that $f(x^k) \ge f(x)$ for all $k \ge \bar k$. Then by the proof of "(i) $\Rightarrow$ (ii)" in Theorem 3.4, $\{x^k\}$ is bounded, so $\sum_k \nu_k = \infty$ and Theorems 3.4 and 3.2(iii) yield $\underline{\lim}_k f(x^k) = f_*$, a contradiction. (iii$_2$–iii$_3$) If $S_* \ne \emptyset$, then $\{x^k\}$ is bounded by Theorem 3.4. On the other hand, if $\{x^k\}$ is bounded, then $\sum_k \nu_k = \infty$, so the conclusion follows from Theorem 3.4. ∎

Remark 7.2. When $\sup_k \epsilon_k < \infty$, (7.2) holds if the oracle is locally bounded and

$$\underline{\lim_{k\to\infty}}\, \mu_k > 0 \quad \text{whenever} \quad \{g^k\} \quad \text{is bounded.} \tag{7.5}$$

Next, we exhibit several choices of the scaling coefficients μ_k for Theorem 7.1 that ensure convergence without *any* indirect assumptions on the boundedness of $\{g^k\}$ which are implicit in the results of sections 3 and 4, and hence do not need the bounding techniques of section 6.

Example 7.3. For a locally bounded oracle (with $\sup_k \epsilon_k < \infty$) and a constant $G > 0$, the requirements (7.1) and (7.3) of Theorem 7.1 and (7.5) are met by the scaling coefficients

$$\mu_k := \max\left\{ |g^k|, |g^k|^2/G \right\}^{-1} = \min\left\{ 1, G/|g^k| \right\} |g^k|^{-1}, \tag{7.6}$$

where G replaces $|g^k|$ if $|g^k| = 0$ (with $\mu_k|g^k|^2 \le G$, $\mu_k|g^k| \le 1$),

$$\mu_k := \max\left\{ 1, |g^k|^2/G^2 \right\}^{-1} = \min\left\{ 1, G^2/|g^k|^2 \right\} \tag{7.7}$$

(with $\mu_k|g^k|^2 \le G^2$, $\mu_k|g^k| \le G$), and

$$\mu_k := \max\left\{ G^2, |g^k|^2 \right\}^{-1} = \min\left\{ 1, G^2/|g^k|^2 \right\} G^{-2} \tag{7.8}$$

(with $\mu_k|g^k|^2 \le 1$, $\mu_k|g^k| \le G^{-1}$); yet another choice of [NuZ77, Thm. 2] with $G \ge 1$ is

$$\mu_k := \begin{cases} 1 & \text{if } |g^k| \le G, \\ |g^k|^{-2} & \text{otherwise.} \end{cases} \tag{7.9}$$

The requirements (7.5), (7.3), and $\sup_k \mu_k|g^k| < \infty$ of Theorem 7.1(iii) are met by

$$\mu_k := |g^k|^{-1}, \tag{7.10}$$

the classical scaling of Shor [Sho62], and its popular variants

$$\mu_k := \big(G + |g^k|\big)^{-1}, \ \mu_k := \max\big\{G, |g^k|\big\}^{-1}, \text{ or } \mu_k := \big(G^2 + |g^k|^2\big)^{-1/2} \tag{7.11}$$

(with $\mu_k|g^k| \le 1$), as well as by the choice of [Lis86]

$$\mu_k := \max\big\{\lambda_k, |g^k|\big\}^{-1} = \min\big\{\lambda_k^{-1}, |g^k|^{-1}\big\} \tag{7.12}$$

(using $\sup_k \lambda_k < \infty$ for (7.5)); note that if $C := \overline{\lim}_{k\to\infty} |g^k| < \infty$ (e.g., $\{x^k\}$ is bounded), then also (7.1) holds with $\bar\gamma \le \frac{1}{2}C$, as required in Theorem 7.1(i)–(ii) (and $\bar\sigma \le \lambda$ in (7.4)). Next,

$$\mu_k := |g^k|^{-2} \tag{7.13}$$

satisfies (7.1) (with $\bar\gamma \le 1/2$) and (7.5) as required in Theorem 7.1(i), as well as (7.3) if $\underline{\lim}_k |g^k| > 0$ (which typically holds in the nondifferentiable case). Thus (7.8) with a "small" G may be regarded as a regularized version of (7.13) that ensures (7.3), but

$$\mu_k := \max\big\{\lambda_k^2, |g^k|^2\big\}^{-1} \tag{7.14}$$

also meets the requirements of Theorem 7.1(i)–(ii) (with $\bar\gamma \le 1/2$, $\nu_k|g^k| \le 1$, $\bar\sigma \le 1$). Note that (7.6)–(7.11) may use a variable $G = G_k \in [G_{\min}, G_{\max}] \subset (0, \infty)$.

Remark 7.4.

(i) Theorem 7.1(i) and its proof correct the proof of [Erm66, sect. 9], where the assumption (7.2) was *implicit* (and the claim that $f(x^k) \to f_*$ was *not* proved). Equation (7.2) is also implicit in [Erm76, Thm. I.3.5] (where $\sup_k \mu_k|g^k| < \infty$ should be replaced by (7.1)) and in [Erm76, Thm. I.3.6] (where $\sup_k \mu_k < \infty$ is implicit); the latter is subsumed by Theorem 7.1(ii). Theorem 7.1(iii$_3$) subsumes [Erm76, Thm. III.1.4] (in the deterministic case).

(ii) Theorem 7.1(ii) subsumes [NuZ77, Thm. 2], which uses (7.9) and $\epsilon = \lambda = 0$. Theorem 7.1(iii) subsumes [Sch83, Lem. on p. 539] with $\mu_k := (G^2 + |g^k|^2)^{-1/2}$ and $\epsilon_k \equiv 0$, and [AIS98, Thm. 1], in which $\mu_k := \max\{1, |g^k|\}^{-1}$ and $\epsilon_k \le C_\epsilon \lambda_k$ with $C_\epsilon < \infty$. Theorem 7.1(iii$_1$) subsumes [Lis86, Thm. on p. 70], which uses (7.12) and $\epsilon_k \equiv 0$, whereas Theorem 7.1(iii$_3$) subsumes [LPS00, Thm. 10] (with $\mu_k := \max\{1, |g^k|\}^{-1}$, $\sum_k \lambda_k\epsilon_k < \infty$, $\epsilon_k \to 0$) and [DeV81, Thm. III.4.5], which uses (7.10) and $\epsilon_k \equiv 0$.

We also have an analogue of Theorem 5.1 for scaled stepsizes.

THEOREM 7.5. *Assume that* $\epsilon := \overline{\lim}_{k\to\infty} \epsilon_k < \infty$, $\{x^k\}$ *is bounded, and* $C := \overline{\lim}_{k\to\infty} |g^k| < \infty$ (*e.g., the oracle is locally bounded*). *Suppose that the algorithm employs stepsizes* $\nu_k := \lambda_k\mu_k$ *with* $\lambda_k, \mu_k > 0$, $\sum_{k=1}^\infty \lambda_k = \infty$, $\lambda := \overline{\lim}_{k\to\infty} \lambda_k < \infty$, $\underline{\lim}_{k\to\infty} \mu_k > 0$, *such that* $\bar\gamma := \overline{\lim}_{k\to\infty} \frac{1}{2}\mu_k|g^k|^2 < \infty$ *and* $\bar\sigma := \overline{\lim}_{k\to\infty} \nu_k|g^k| < \infty$.

Let $\mu := \overline{\lim}_{k\to\infty} \mu_k$ *and* $\nu := \overline{\lim}_{k\to\infty} \nu_k$. *Then the conclusions of Theorem* 5.1 *hold with* $\gamma \le \bar\gamma\lambda$, $\bar\gamma \le \frac{1}{2}C^2\mu$, $\sigma \le \bar\sigma \le \lambda \min\{\overline{\lim}_{k\to\infty} \mu_k|g^k|, (2\mu\bar\gamma)^{1/2}\}$, *and* $\nu \le \lambda\mu$.

Proof. Invoke Theorem 7.1(ii) in the proof of Theorem 5.1. ∎

Remark 7.6.

(i) For a locally bounded oracle, the requirements of Theorem 7.5 are met by the scaling coefficients given by (7.6)–(7.12).

(ii) Theorem 7.5 subsumes [MGN87, Thm. 9.2] in the convex case with $\lambda = \epsilon = 0$.

7.2. Analysis of Shor-type scalings. Additional results for the Shor-type scalings (7.10)–(7.12) require the following assumption.

Assumption 7.7. The objective f is finite-valued and $g^k \in \partial_{\epsilon_k} f(x^k)$ for all k.

Under Assumption 7.7, the objective f is continuous, as required for the following basic estimates inspired by [Nes84, Lem. 1].

LEMMA 7.8. *Suppose Assumption* 7.7 *holds. Fixing a point* $x \in S$, *define the function*

$$\omega_x(\rho) := \max_{x+B_\rho} f \quad \textit{for} \quad \rho \ge 0, \tag{7.15}$$

and let ρ_k^+ *be the distance from the point* x *to the halfspace* $\{y : \langle g^k, x^k - y\rangle \le 0\}$:

$$\rho_k^+ := \max\{\rho_k, 0\} \quad \textit{with} \quad \rho_k := \begin{cases} \langle g^k/|g^k|, x^k - x\rangle & \textit{if } g^k \ne 0, \\ 0 & \textit{otherwise.} \end{cases} \tag{7.16}$$

The function ω_x *is continuous and nondecreasing, and we have the estimate*

$$f(x^k) \le \omega_x(\rho_k^+) + \epsilon_k. \tag{7.17}$$

The stepsize $\nu_k := \lambda_k\mu_k$ *with* $\lambda_k > 0$ *and* $\mu_k \le |g^k|^{-1}$ (*as in* (7.10)–(7.12)) *produces*

$$|x^{k+1} - x|^2 - |x^k - x|^2 \le -2\nu_k|g^k|\left(\rho_k - \tfrac{1}{2}\nu_k|g^k|\right) \le -2\lambda_k\mu_k|g^k|\left(\rho_k - \tfrac{1}{2}\lambda_k\right). \tag{7.18}$$

Proof. Suppose $f(x) < f(x^k) - \epsilon_k$. (Otherwise (7.17) holds with $\omega_x(\rho_k^+) \ge f(x)$.) Then $\rho_k > 0$ (since $g^k \in \partial_{\epsilon_k} f(x^k)$). The point $\hat x := x + \frac{\rho_k}{|g^k|} g^k$ satisfies $|\hat x - x| = \rho_k$ and $\langle g^k, x^k - \hat x\rangle = 0$, so $f(\hat x) \le \omega_x(\rho_k)$ and $f(\hat x) \ge f(x^k) - \epsilon_k$ (from $g^k \in \partial_{\epsilon_k} f(x^k)$); thus (7.17) holds. For (7.18), rewrite (3.4) with $\nu_k := \lambda_k\mu_k$ and use $\mu_k|g^k| \le 1$. ∎

We have the following analogue of Theorem 7.1(i) for the scalings (7.10)–(7.12).

THEOREM 7.9. *Suppose Assumption* 7.7 *holds,* $\epsilon := \overline{\lim}_{k\to\infty} \epsilon_k < \infty$, *and the algorithm employs stepsizes* $\nu_k := \lambda_k\mu_k$ *with* $\lambda_k > 0$, $\sum_{k=1}^\infty \lambda_k = \infty$, $\lambda := \overline{\lim}_{k\to\infty} \lambda_k < \infty$, *and* μ_k *chosen as in* (7.10)–(7.12). *Then we have the following statements:*

(i) $\underline{\lim}_{k\to\infty} f(x^k) \le \inf_{x\in S} \max_{x+B_{\lambda/2}} f + \epsilon$.

(ii) *If* $\lambda = 0$ (*i.e.,* $\lim_{k\to\infty} \lambda_k = 0$), *then* $\underline{\lim}_{k\to\infty} f(x^k) \le f_* + \epsilon$.

(iii) *If* $S_* \ne \emptyset$, *then* $\underline{\lim}_{k\to\infty} f(x^k) \le \inf_{x\in S_*} \max_{x+B_{\lambda/2}} f + \epsilon \le \sup_{S_*+B_{\lambda/2}} f + \epsilon$.

Proof. We need only to prove item (i), since (ii) and (iii) follow immediately from (i).

First, suppose μ_k is chosen via (7.10). Then for $x \in S$ and ρ_k defined by (7.16) we have

$$\underline{\lim}_{k\to\infty}\ \rho_k \le \tfrac{1}{2}\lambda. \tag{7.19}$$

Indeed, summing up (7.18) with $\mu_k|g^k|$ replaced by 1 produces the Cesáro estimate

$$\bar\rho_k := \frac{\sum_{j=1}^k \lambda_j\rho_j}{\sum_{j=1}^k \lambda_j} \le \frac{|x^1-x|^2+\sum_{j=1}^k \lambda_j^2}{2\sum_{j=1}^k \lambda_j}, \tag{7.20}$$

which combined with $\sum_k \lambda_k = \infty$ yields $\underline{\lim}_k \rho_k \le \overline{\lim}_k \bar\rho_k \le \frac{1}{2}\lambda$ (cf. Lemma 2.1). By (7.17) and (7.19), we have $\underline{\lim}_{k\to\infty} f(x^k) \le \max_{x+B_{\lambda/2}} f + \epsilon$ for each $x \in S$, as required.

Similarly, for the remaining choices (7.11)–(7.12), assertion (i) is established if (7.19) holds, so suppose $\underline{\lim}_k \rho_k > \frac{1}{2}\lambda$ for some $x \in S$. Thus, since $\lambda := \overline{\lim}_k \lambda_k$, we have $\rho_k > \frac{1}{2}\lambda_k$ for large k and (7.18) shows that $\{x^k\}$ is bounded. We consider two cases.

First, suppose $\underline{\lim}_k |g^k| = 0$. Then a subsequence $g^{k_j} \to 0$, and taking limits in the subgradient inequality $f(y) \ge f(x^{k_j}) - \epsilon_{k_j} + \langle g^{k_j}, y - x^{k_j}\rangle$ gives $\underline{\lim}_k f(x^k) \le f(y) + \epsilon$ for each y; thus assertion (i) holds.

Second, suppose $\underline{\lim}_k |g^k| > 0$. Write $\nu_k := \lambda_k\mu_k$ as $\nu_k = \hat\lambda_k\hat\mu_k$ with $\hat\lambda_k := \lambda_k\mu_k|g^k|$ and $\hat\mu_k := |g^k|^{-1}$. Note that $\hat\lambda_k \le \lambda_k$ (since $\mu_k \le |g^k|^{-1}$) and $\underline{\lim}_k \mu_k|g^k| > 0$ for the choices (7.11)–(7.12) (using $\underline{\lim}_k |g^k| > 0$ and $\overline{\lim}_k \lambda_k < \infty$ for (7.12)). The first property gives $\hat\lambda := \overline{\lim}_k \hat\lambda \le \lambda$, whereas the second one combined with $\sum_k \lambda_k = \infty$ implies $\sum_k \hat\lambda_k = \infty$. Hence by replacing λ_k, μ_k by $\hat\lambda_k$, $\hat\mu_k$ in the argument of the first paragraph we obtain assertion (i) with λ replaced by $\hat\lambda$; since $\hat\lambda \le \lambda$, (i) must hold for λ as well. □

A result on finite convergence is given in part (ii) of the following corollary.

COROLLARY 7.10. *Under the assumptions of Theorem* 7.9*, suppose that the optimal set S_* is nonempty and $\epsilon_k \equiv 0$ so that $\lambda := \overline{\lim}_{k\to\infty} \lambda_k$ determines the asymptotic accuracy. Then we have the following statements:*

(i) *For every $\hat\delta > 0$, if λ is small enough so that $\omega_x(\frac{1}{2}\lambda) < f_* + \hat\delta$ for some $x \in S_*$ (cf.* (7.15)*), then $\underline{\lim}_{k\to\infty} f(x^k) < f_* + \hat\delta$.*

(ii) *For every $\rho > \frac{1}{2}\lambda$ and $x \in S_*$, if $\omega_x(\rho) > f_*$ or the Shor scaling* (7.10) *is used, then there is an iteration $\hat k$ such that $f(x^{\hat k}) = f(\hat x)$ for a point $\hat x$ satisfying $|\hat x - x| < \rho$; in particular, if $x + B_\rho \subset S_*$ and the Shor scaling* (7.10) *is employed, then $x^{\hat k} \in S_*$.*

Proof. (i) By (7.15) and Theorem 7.9(iii), $\underline{\lim}_k f(x^k) \le \omega_x(\frac{1}{2}\lambda)$.

(ii) The function ω_x is increasing for ρ such that $\omega_x(\rho) > f(x) = f_*$ (since any maximizer y of (7.15) satisfies $|y - x| = \rho$ by convexity), so $\underline{\lim}_k f(x^k) \le \omega_x(\frac{1}{2}\lambda) < \omega_x(\rho)$ yields the existence of $\hat k$ such that $f(x^{\hat k}) < \omega_x(\rho)$. For the scaling (7.10), since $\underline{\lim}_k \rho_k \le \frac{1}{2}\lambda < \rho$ by (7.19), for $\hat k$ such that $\rho_{\hat k}^+ < \rho$ we have $f(x^{\hat k}) \le \omega_x(\rho_{\hat k}^+)$ by (7.17). The existence of $\hat x$ follows from the continuity of f in (7.15), with $f(\hat x) = f_*$ if $x + B_\rho \subset S_*$. □

The Shor-type scalings (7.10)–(7.12) have the following analogue of Theorem 7.1(ii).

THEOREM 7.11. *Suppose Assumption* 7.7 *holds, $\epsilon := \overline{\lim}_{k\to\infty} \epsilon_k < \infty$, the algorithm employs stepsizes $\nu_k := \lambda_k\mu_k$ with $\lambda_k > 0$, $\sum_{k=1}^\infty \lambda_k = \infty$, $\lambda := \overline{\lim}_{k\to\infty} \lambda_k < \infty$, μ_k chosen as in* (7.10)–(7.12)*, and f_S is coercive. Then $\sigma := \overline{\lim}_{k\to\infty} |x^{k+1} - x^k| \le \lambda$. Let*

$$\hat\delta := \hat\gamma + \epsilon \quad \textit{with} \quad \hat\gamma := \max_{S_* + B_{\lambda/2}} f - f_*. \tag{7.21}$$

Then we have the following statements:

(i) $\underline{\lim}_{k\to\infty} d_{S_{\hat\delta}}(x^k) = 0$ *and* $\{x^k\}$ *has a cluster point in* $S_{\hat\delta}$.

(ii) $\lim_{k\to\infty} d_{S_*^{\hat\delta}}(x^k) = 0$, *where* $S_*^{\hat\delta}$ *is the neighborhood of* S_* *defined by* (*cf. Lemma* 2.4(i))

$$S_*^{\hat\delta} := S_* + B_{\rho_{\hat\delta}+\sigma} \quad \textit{with} \quad \rho_{\hat\delta} := \max\left\{ d_{S_*}(x) : x \in S_{\hat\delta} \right\}. \tag{7.22}$$

Thus $\{x^k\}$ *is bounded and its cluster points belong to* $S_*^{\hat\delta}$.

(iii) $C := \overline{\lim}_{k\to\infty} |g^k|$ *is finite and the conclusions of Theorem* 7.1(i)–(ii) *hold with* $\bar\gamma \le \frac{1}{2}C$; *in particular, the conclusions of Theorem* 4.1 *hold with* $\gamma \le \frac{1}{2}C\lambda$ *and* $\sigma \le \lambda$ *so that assertions* (i) *and* (ii) *hold with* $\hat\delta$ *replaced by* $\min\{\delta, \hat\delta\}$, *where* $\delta := \overline{\lim}_k \delta_k \le \frac{1}{2}C\lambda + \epsilon$.

Proof. As in the proof of Theorem 4.1, the closedness and coercivity of f_S imply that the sets $S_* \subset S_{\hat\delta} \subset S_* + B_{\rho_{\hat\delta}} \subset S_*^{\hat\delta}$ are nonempty and compact (with $\hat\gamma < \infty$ because f is continuous). Further, (3.3) implies $|x^{k+1} - x^k| \le \lambda_k\mu_k|g^k| \le \lambda_k$, and hence $\sigma \le \lambda$.

(i) By Theorem 7.9(iii) and (7.21), we have $\underline{\lim}_k f(x^k) \le f_* + \hat\gamma + \epsilon = f_* + \hat\delta$, so the conclusion follows upon replacing δ by $\hat\delta$ in the proof of Theorem 4.1(i).

(ii) Fixing $v > 0$, let $\lambda_v := \lambda + v$, $\gamma_v := \max_{S_*+B_{\lambda_v/2}} f - f_*$, $\delta_v := \gamma_v + \epsilon + v$, $\alpha := \alpha_v := f_* + \delta_v$, $\rho_\alpha := \max_{T_\alpha} d_{S_{\hat\delta}}$ (so that $T_\alpha \subset S_{\hat\delta} + B_{\rho_\alpha}$; cf. (2.1), (2.2)), and (cf. (7.22))

$$V_v := S_*^{\hat\delta} + B_{\rho_\alpha+v} = S_* + B_{\rho_{\hat\delta}+\sigma+\rho_\alpha+v}. \tag{7.23}$$

By (7.21), $\gamma_v \ge \hat\gamma$, $\delta_v > \hat\delta$, and $\alpha_v > f_* + \hat\delta$. Since S_* is compact and f is continuous, for $v \downarrow 0$ we have $\gamma_v \downarrow \hat\gamma$, $\delta_v \downarrow \hat\delta$, $\alpha_v \downarrow f_* + \hat\delta$, and $\rho_\alpha \downarrow 0$ (cf. Lemma 2.4(i) with $\beta := f_* + \hat\delta$).

Since $\lambda := \overline{\lim}_k \lambda_k$, $\epsilon := \overline{\lim}_k \epsilon_k$ and $\sigma := \overline{\lim}_k |x^{k+1} - x^k|$, there is $k_v < \infty$ such that

$$\lambda_k \le \lambda_v, \quad \epsilon_k \le \epsilon + v, \quad \text{and} \quad |x^{k+1} - x^k| \le \sigma + v \quad \forall k \ge k_v. \tag{7.24}$$

Since $\underline{\lim}_k d_{S_{\hat\delta}}(x^k) = 0$ by (i), there exists $k = k'_v \ge k_v$ such that $x^k \in S_{\hat\delta} + B_v$; then $S_{\hat\delta} \subset S_*^{\hat\delta}$ implies $x^k \in V_v$ (cf. (7.23)).

Assuming $x^k \in V_v$ for some $k \ge k'_v$, we now show that $x^{k+1} \in V_v$. If $x^k \in T_\alpha$, then from the third inequality of (7.24), $T_\alpha \subset S_{\hat\delta} + B_{\rho_\alpha}$, and $S_{\hat\delta} \subset S_* + B_{\rho_{\hat\delta}}$ (cf. (7.22)) we get

$$x^{k+1} \in T_\alpha + B_{\sigma+v} \subset S_{\hat\delta} + B_{\rho_\alpha+\sigma+v} \subset S_* + B_{\rho_{\hat\delta}} + B_{\rho_\alpha+\sigma+v} = S_* + B_{\rho_{\hat\delta}+\sigma+\rho_\alpha+v},$$

so $x^{k+1} \in V_v$ (cf. (7.23)). Thus suppose $x^k \notin T_\alpha$. Then, by the second inequality of (7.24),

$$f(x^k) - \epsilon_k > \alpha - \epsilon_k = f_* + \gamma_v + \epsilon + v - \epsilon_k \ge f_* + \gamma_v = \max_{S_*+B_{\lambda_v/2}} f,$$

so for $x = P_{S_*}x^k$, by Lemma 7.8, we have $\omega_x(\frac{1}{2}\lambda_v) < f(x^k) - \epsilon_k \le \omega_x(\rho_k^+)$, $\rho_k > \frac{1}{2}\lambda_v$, and $|x^{k+1} - x| \le |x^k - x|$ because $\lambda_k \le \lambda_v$ in (7.18) due to the first inequality of (7.24). Since $x \in S_*$ and $x^k \in V_v$, the inequality $|x^{k+1} - x| \le |x^k - x|$ and (7.23) yield $x^{k+1} \in V_v$.

Therefore, by induction for each $k \geq k'_v$, $x^k \in V_v$ and hence (cf. (7.23)) $d_{S_*^{\hat{\delta}}}(x^k) \leq \rho_\alpha + v$. Since $\rho_\alpha \downarrow 0$ as $v \downarrow 0$, $d_{S_*^{\hat{\delta}}}(x^k) \to 0$. The rest follows as in the proof of Theorem 4.1(ii).

(iii) We have $\sup_k |g^k| < \infty$, since $\{x^k\}$ is bounded, $\sup_k \epsilon_k < \infty$, and the oracle is locally bounded under Assumption 7.7 (cf. Remark 6.2(i)). The conclusion follows from Theorem 7.1 and the discussion of (7.10)–(7.12) in Example 7.3. ∎

Remark 7.12.

(i) Theorem 7.11(ii) may be augmented as follows: (ii$_1$) *if* $\lambda = \epsilon = 0$, *then* $S_*^{\hat{\delta}} = S_{\hat{\delta}} = S_*$ *and* $\lim_{k\to\infty} d_{S_*}(x^k) = 0$; (ii$_2$) $\overline{\lim}_{k\to\infty} f(x^k) \leq \max_{S\cap S_*^{\hat{\delta}}} f$ (*so that* $\lim_{k\to\infty} f(x^k) = f_*$ *if* $\lambda = \epsilon = 0$). Indeed, this follows as in Remark 4.3(i).

(ii) For $\lambda > 0$ (i.e., nonvanishing stepsizes), the asymptotic accuracy determined by $\hat{\gamma}$ in (7.21) may depend on the behavior of f outside the feasible set S, whereas the corresponding bound of Theorem 4.1 expressed by $\gamma \leq \frac{1}{2}\lambda \overline{\lim}_k |g^k|$ depends on the properties of f seen by the algorithm inside S; the bound of Theorem 7.11(iii) using $\min\{\delta, \hat{\delta}\}$ combines the best of both worlds.

(iii) The estimate (7.17) extends [Nes84, Lem. 1] (to $\epsilon_k > 0$). Theorem 7.9 subsumes [Pol67, Thm. 1] (which uses (7.10) and $\epsilon_k \equiv 0$). For the Shor scaling (7.10), Corollary 7.10 subsumes [Sho79, Thm. 2.1 and Cors. 1–2] (where $\lambda_k \equiv \lambda > 0$) and [DeV81, Cor. III.4.1] (where $\lambda = 0$), whereas Theorem 7.11(i)–(ii) subsumes [DeV81, Thms. III.4.1–4] and some results of [DeV81, sect. IV.5]; the proof of a related result [LPS00, Thm. 6] is wrong.

7.3. Shor's bounding strategy. The following result helps in analyzing the bounding strategy of Shor [Sho79, Thm. 2.4].

PROPOSITION 7.13. *Suppose that Assumption* 7.7 *holds and* f_S *is coercive. Fix any point* $\bar{x} \in S$, *a step bound* $\bar{\rho} \in (0, \infty)$, *and an error threshold* $\bar{\epsilon}_{\max} \in [0, \infty)$. *If* $f_S(x^1) \leq f(\bar{x})$, $\nu_k|g^k| \leq \bar{\rho}$, *and* $\epsilon_k \leq \bar{\epsilon}_{\max}$ *for all* k, *then* $\{x^k\}$ *and* $\{g^k\}$ *are bounded.*

Proof. Let $\alpha := \max_{\bar{x}+B_{\bar{\rho}}} f + \bar{\epsilon}_{\max}$. Since $f(x^1) \leq f(\bar{x})$, we have $x^1, \bar{x} \in T_\alpha$ (cf. (2.1)). First, suppose $x^k \in T_\alpha$. Since $|x^{k+1} - x^k| \leq \nu_k|g^k| \leq \bar{\rho}$ by (3.3) and our assumption,

$$|x^{k+1} - \bar{x}| \leq |x^k - \bar{x}| + |x^{k+1} - x^k| \leq \operatorname{diam}(T_\alpha) + \bar{\rho} \quad \text{if} \quad x^k \in T_\alpha. \tag{7.25}$$

Next, suppose $x^k \notin T_\alpha$. Then $f(x^k) > \max_{\bar{x}+B_{\bar{\rho}}} f + \epsilon_k$, since $\epsilon_k \leq \bar{\epsilon}_{\max}$. Thus for $x = \bar{x}$ in Lemma 7.8, we have $f(x^k) > \omega_x(\bar{\rho}) + \epsilon_k$ (cf. (7.15)), so (7.17) yields $\rho_k > \bar{\rho}$, and then (3.4) or, equivalently, the first inequality of (7.18) with $\nu_k|g^k| \leq \bar{\rho}$ gives $|x^{k+1} - \bar{x}| \leq |x^k - \bar{x}|$. Combining this with (7.25) yields $|x^k - \bar{x}| \leq \operatorname{diam}(T_\alpha) + \bar{\rho}$ for all k, since $x^1, \bar{x} \in T_\alpha$. ∎

In the framework of Proposition 7.13, we may use the following bounding strategy that generates finitely many restarts indexed by $l = 1, 2, \ldots$. Fixing $\bar{x} \in S$, $\bar{\rho} > 0$, and $\bar{\epsilon}_{\max} \geq 0$, pick a positive sequence $\nu^l_{\max} \to 0$. For the current $l \geq 1$, start the algorithm from $\bar{x}$ (or the best point found so far if $l > 1$), using stepsizes $\nu_k \leq \nu^l_{\max}$ and errors $\epsilon_k \leq \bar{\epsilon}_{\max}$; if $\nu_k|g^k| > \bar{\rho}$ for some k, then increase l by 1, restart the algorithm, etc. Since the number of restarts is finite by Theorem 6.4, this strategy ensures the boundedness of $\{x^k\}$ and $\{g^k\}$. A special case of this strategy consists of picking a sequence $\nu_k \to 0$ and resetting x^{k+1} to x^1 whenever $\nu_k|g^k| > \bar{\rho}$ (as in [Sho79, Thm. 2.4]).

Remark 7.14. Proposition 7.13 also fills a gap in the proof of [Sho79, Thm. 2.4].

7.4. Fejér-type stepsizes. We now highlight a property of the *quadratic* scalings (7.6)–(7.9) and (7.13)–(7.14) based on $|g^k|^2$ which distinguishes them from the *linear* scalings (7.10)–(7.12) that use $|g^k|$.

COROLLARY 7.15. *Suppose that* $\epsilon := \overline{\lim}_{k\to\infty} \epsilon_k < \infty$ *and the algorithm employs a locally bounded oracle and stepsizes* $\nu_k := \lambda_k\mu_k$ *with* $\lambda_k > 0$, $\sum_{k=1}^{\infty} \lambda_k = \infty$, *and* μ_k *chosen as in* (7.6)–(7.9) *or* (7.13)–(7.14). *If* $\lambda := \overline{\lim}_{k\to\infty} \lambda_k$ *is finite, then* $\underline{\lim}_{k\to\infty} f(x^k) \le f_* + \bar\gamma\lambda + \epsilon$, *where (cf.* (7.1)*)* $\bar\gamma$ *is at most* $\frac{1}{2}G$ *for* μ_k *chosen via* (7.6), *and* $\frac{1}{2}G^2$ *for* (7.7), $\frac{1}{2}$ *for* (7.8) *and* (7.13)–(7.14), *and* $\frac{1}{2}G^2$ *for* (7.9). *Consequently, we have* $\inf_k f(x^k) \le f_* + \frac{1}{2}\bar\gamma\lambda + \epsilon$ *if* λ *is finite whenever* $\inf_k f(x^k) > -\infty$.

Proof. This follows from Theorem 7.1(i) and the discussion in Example 7.3. ∎

Remark 7.16.

(i) Corollary 7.15 says that for the quadratic scalings (7.6)–(7.9) and (7.13)–(7.14), the asymptotic objective accuracy can be controlled by choosing the stepsize value λ a priori. In contrast, the asymptotic accuracy for the linear scalings (7.10)–(7.12) depends on the value of $\inf_{x\in S} \max_{x+B_{\lambda/2}} f$ (cf. Thm 7.9), which may be hard to guess.

(ii) The following adaptive choice of λ_k meets the requirements of Corollary 7.15. Select $\lambda_{\min} \in (0,\infty)$, $\kappa \in (0,1)$, and $\lambda_1 \ge \lambda_{\min}$. For each k, letting $f^k_{\rm rec} := \min_{j=1}^k f(x^j)$, choose

$$\lambda_{k+1} \in \begin{cases} [\lambda_{\min},\infty) & \text{if} \quad f(x^{k+1}) \le f^k_{\rm rec} - \lambda_{\min}, \\ [\lambda_{\min}, \max\{\lambda_{\min}, \kappa\lambda_k\}] & \text{if} \quad f(x^{k+1}) > f^k_{\rm rec} - \lambda_{\min}. \end{cases} \tag{7.26}$$

Clearly, either $f^k_{\rm rec} \downarrow -\infty$ (and hence $f_* = -\infty$) or $\lambda_k = \lambda_{\min}$ for all large k.

Our quadratic scalings are related to *Fejér* stepsizes that reduce the distance to the solution set S_*. The latter stem from the observation that for $x \in S_*$ and $\epsilon_k = 0$, the optimal stepsize ν_k that minimizes the right-hand side of the estimate (3.1) has the form $\nu_k = \lambda_k\mu_k$ with $\lambda_k = f(x^k) - f_*$ and $\mu_k = |g^k|^{-2}$. Such stepsizes are analyzed below.

THEOREM 7.17. *Suppose that* $f_* > -\infty$ *and the algorithm employs a locally bounded oracle and stepsizes*

$$\nu_k := \kappa_k \left[\, f(x^k) - f_* \,\right] |g^k|^{-2} \quad \textit{with} \quad \kappa_k \in [\kappa_{\min}, \kappa_{\max}] \subset (0,2). \tag{7.27}$$

(i) *If* $\epsilon := \overline{\lim}_{k\to\infty} \epsilon_k$ *is finite, then* $\underline{\lim}_{k\to\infty} f(x^k) \le f_* + \frac{2}{2-\kappa_{\max}}\epsilon$.

(ii) *If the solution set* S_* *is nonempty and for all* k

$$\epsilon_k \le \tfrac{1}{2}\kappa_\epsilon(2-\kappa_k)\left[\, f(x^k) - f_* \,\right] \quad \textit{with} \quad \kappa_\epsilon \in [0,1), \tag{7.28}$$

then $\{x^k\}$ *converges to some solution* $x^\infty \in S_*$ *and* $\lim_{k\to\infty} f(x^k) = f_*$.

Proof. (i) For contradiction, suppose $\frac{2-\kappa_{\max}}{2}\,\underline{\lim}_{k\to\infty} \lambda_k > \epsilon$, where $\lambda_k := f(x^k) - f_*$. Since $\epsilon := \overline{\lim}_k \epsilon_k \ge 0$ and $f_* := \inf_S f$, there exist $\kappa \in (0,1)$, $x \in S$, and k_ϵ such that

$$\kappa\tfrac{2-\kappa_{\max}}{2}\lambda_k \ge f(x) - f_* + \epsilon_k \quad \forall k \ge k_\epsilon. \tag{7.29}$$

Using the fact that $\lambda_k := f(x^k) - f_* \ge 0$, (7.27), (7.29), and again (7.27) in (3.1)

yields

$$\begin{aligned}|x^{k+1}-x|^2-|x^k-x|^2 &\le -2\nu_k\left[\, f_*-f(x)-\epsilon_k+f(x^k)-f_*-\tfrac{1}{2}\nu_k|g^k|^2\,\right]\\ &= -2\nu_k\left[\, f_*-f(x)-\epsilon_k+\lambda_k-\tfrac{1}{2}\kappa_k\lambda_k\,\right]\\ &\le -2\nu_k(1-\kappa)\tfrac{2-\kappa_{\max}}{2}\lambda_k\\ &\le -\kappa_{\min}(1-\kappa)(2-\kappa_{\max})\lambda_k^2/|g^k|^2<0 \qquad \forall k\ge k_\epsilon. \end{aligned} \tag{7.30}$$

By (7.30), $\{x^k\}$ is bounded and $\sum_k \lambda_k^2/|g^k|^2<\infty$. Hence $\sup_k|g^k|<\infty$ (because the oracle is locally bounded and $\epsilon<\infty$) and $\lambda_k^2/|g^k|^2\to 0$ yields $\lambda_k\to 0$, a contradiction.

(ii) For any $x\in S_*$, using (7.28) and (7.27) as in (7.30) yields

$$\begin{aligned}|x^{k+1}-x|^2-|x^k-x|^2 &\le -2\nu_k\left[\,\tfrac{1}{2}(2-\kappa_k)\lambda_k-\epsilon_k\,\right]\\ &\le -2\nu_k(1-\kappa_\epsilon)\tfrac{2-\kappa_{\max}}{2}\lambda_k\\ &\le -\kappa_{\min}(1-\kappa_\epsilon)(2-\kappa_{\max})\lambda_k^2/|g^k|^2<0 \qquad \forall k\ge 1, \end{aligned} \tag{7.31}$$

so again $\{x^k\}$ is bounded and $\lambda_k/|g^k|\to 0$. Then $\epsilon:=\overline{\lim}_k\,\epsilon_k<\infty$ by (7.28) (since f is continuous because the oracle is bounded), and as in (i) we get $\lambda_k:=f(x^k)-f_*\to 0$. Further, $\{x^k\}$ has a cluster point $x^\infty\in S$ with $f(x^\infty)\le f_*$ (since S and f are closed), i.e., $x^\infty\in S_*$. Setting $x=x^\infty$ in (7.31) shows that $|x^k-x^\infty|\downarrow 0$, i.e., $x^k\to x^\infty$. □

Remark 7.18. In contrast to standard results, Theorem 7.17(i) *does not* assume nonemptiness of the solution set S_*. Theorem 7.17(ii) subsumes [Pol69, Thm. 1] (where $\epsilon_k\equiv 0$) and [Brä95, Thm. 2.4] (for a special oracle). As in [Brä95, sect. 2], condition (7.28) may be replaced by $\inf_k\kappa_k(2-\kappa_k-2\epsilon_k/\lambda_k)>0$ with (7.27) relaxed to $\kappa_k\in[0,2]$.

Since the optimal value f_* in (7.27) is usually unknown, it may be replaced by a *target level* $f_{\rm lev}^k:=f_{\rm rec}^k-\tilde\delta_k$ with $\tilde\delta_k$ updated as in (7.26); the resulting scheme is analyzed below.

THEOREM 7.19. *Suppose that* $\epsilon:=\overline{\lim}_{k\to\infty}\epsilon_k<\infty$ *and the algorithm employs a locally bounded oracle and stepsizes* $\nu_k:=\lambda_k\mu_k$ *with*

$$\lambda_k:=f(x^k)-f_{\rm lev}^k,\quad f_{\rm lev}^k:=f_{\rm rec}^k-\tilde\delta_k, \tag{7.32}$$

$$\mu_k:=\kappa_k|g^k|^{-2},\quad \kappa_k\in[\kappa_{\min},\kappa_{\max}], \tag{7.33}$$

where $f_{\rm rec}^k:=\min_{j=1}^k f(x^j)$, $0<\kappa_{\min}\le\kappa_{\max}\le 2$, *and* $\tilde\delta_k>0$ *is such that* $\tilde\delta:=\overline{\lim}_{k\to\infty}\tilde\delta_k\in(0,\infty)$ *whenever* $f_{\rm rec}^\infty:=\inf_k f(x^k)>-\infty$ (*e.g.,* $\tilde\delta_k\equiv\tilde\delta>0$). *Then either* $f_{\rm rec}^\infty=-\infty=f_*$ *or* $f_{\rm rec}^\infty\le f_*+\tilde\delta+\epsilon$ *with* $\tilde\delta<\infty$.

Proof. If $f_{\rm rec}^\infty=-\infty$, then $f_*\le\inf_k f(x^k)=-\infty$, so assuming $f_{\rm rec}^\infty>-\infty$, suppose $f_{\rm rec}^\infty>f_*+\epsilon+\tilde\delta$. Then there exist $x\in S$ and $v>0$ such that $f_{\rm rec}^k\ge f(x)+\epsilon+\tilde\delta+v$ for all k, so using (7.32) with $\tilde\delta:=\overline{\lim}_k\tilde\delta_k$ and $\epsilon:=\overline{\lim}_k\epsilon_k$ we deduce the existence of k_v such that

$$f_{\rm lev}^k-f(x)-\epsilon_k=f_{\rm rec}^k-f(x)-\tilde\delta_k-\epsilon_k\ge\tilde\delta-\tilde\delta_k+\epsilon-\epsilon_k+v\ge\tfrac{1}{2}v \tag{7.34}$$

for all $k\ge k_v$. Since $\lambda_k\ge\tilde\delta_k>0$ by (7.32) and $\mu_k|g^k|^2\le\kappa_{\max}$ by (7.33), we have $\nu_k|g^k|^2\le\kappa_{\max}\lambda_k$. Hence using (7.32), (7.34), and $\kappa_{\max}\le 2$ in the estimate (3.1)

yields

$$\begin{aligned}|x^{k+1}-x|^2-|x^k-x|^2 &\le -2\nu_k\left[\,f^k_{\rm lev}-f(x)-\epsilon_k+f(x^k)-f^k_{\rm lev}-\tfrac12\nu_k|g^k|^2\,\right]\\ &\le -2\nu_k\left[\,f^k_{\rm lev}-f(x)-\epsilon_k+\lambda_k-\tfrac12\kappa_{\max}\lambda_k\,\right]\\ &\le -\nu_k\left[\,v+(2-\kappa_{\max})\lambda_k\,\right]\le -v\nu_k<0 \qquad \forall k\ge k_v.\end{aligned} \tag{7.35}$$

By (7.35), $\{x^k\}$ is bounded and $\sum_k\nu_k<\infty$. Hence $\hat G:=\sup_k|g^k|<\infty$ (because the oracle is locally bounded and $\epsilon<\infty$) and $\lim_k\nu_k=0$. However, $\nu_k:=\lambda_k\mu_k\ge\tilde\delta_k\kappa_{\min}\hat G^{-2}$ by (7.32)–(7.33), where $\kappa_{\min}>0$, so we get $\tilde\delta:=\overline{\lim}_k\,\tilde\delta_k=0$, a contradiction. □

Remark 7.20.

(i) The following adaptive choice of $\tilde\delta_k$ meets the requirements of Theorem 7.19. Select $\tilde\delta_{\min}\in(0,\infty)$, $\kappa\in(0,1)$, and $\tilde\delta_1\ge\tilde\delta_{\min}$. For each k, choose

$$\tilde\delta_{k+1}\in\begin{cases}[\tilde\delta_{\min},\infty) & \text{if}\quad f(x^{k+1})\le f^k_{\rm rec}-\tilde\delta_{\min},\\ \left[\tilde\delta_{\min},\max\left\{\tilde\delta_{\min},\kappa\tilde\delta_k\right\}\right] & \text{if}\quad f(x^{k+1})>f^k_{\rm rec}-\tilde\delta_{\min}.\end{cases} \tag{7.36}$$

Clearly, either $f^k_{\rm rec}\downarrow-\infty$ (and hence $f_*=-\infty$) or $\tilde\delta_k=\tilde\delta_{\min}$ for all large k.

(ii) A special case of (7.36) introduced in [NeB01, eq. (2.19)] is to set $\tilde\delta_{k+1}:=\eta\tilde\delta_k$ if $f(x^{k+1})\le f^k_{\rm lev}$, $\tilde\delta_{k+1}:=\max\{\tilde\delta_{\min},\kappa\tilde\delta_k\}$ otherwise, where $\eta\in[1,\infty)$. For this case Theorem 7.19 subsumes [NeB01, Rem. 2.1] (where $\epsilon_k\equiv0$ and $\kappa_{\max}<2$ in (7.33)). In the exact case ($\epsilon_k\equiv0$) similar schemes with nonvanishing level gaps are considered in [Kiw96b, Thm. 4.4], [Kiw98, Thm. 4.2], and [SCT00]; vanishing level gaps are studied in [Brä93, GoK99, KLL99b, NeB01].

8. Efficiency estimates. In order to derive efficiency estimates, in this section we assume that the optimal set S_* is nonempty and that the sequences $\{x^k\}$, $\{g^k\}$, and $\{\epsilon_k\}$ are bounded.

For some stepsizes, sharper estimates may be derived by replacing the index $j=1$ in (3.2), (3.7), (3.8), and (3.10) by $j=k'$, where k' depends on k, e.g., $k':=\lceil\frac12k\rceil$. Thus for

$$\bar f_k:=\sum_{j=k'}^k\nu_jf(x^j)/\nu^k_{\rm sum},\ \bar x^k:=\sum_{j=k'}^k\nu_jx^j/\nu^k_{\rm sum},\ \bar\epsilon_k:=\sum_{j=k'}^k\nu_j\epsilon_j/\nu^k_{\rm sum},\ \nu^k_{\rm sum}:=\sum_{j=k'}^k\nu_j, \tag{8.1}$$

replacing 1 by k' in (3.2) and using $x:=P_{S_*}x^{k'}$ yields the estimate

$$\bar f_k-f_*\le\Delta_k+\bar\epsilon_k\quad\text{with}\quad\Delta_k:=\frac{d^2_{S_*}(x^{k'})+\sum_{j=k'}^k\nu_j^2|g^j|^2}{2\sum_{j=k'}^k\nu_j}. \tag{8.2}$$

This is indeed an *accuracy estimate*, since we still have (cf. (3.9), (3.14))

$$f(\bar x^k)\le\bar f_k\quad\text{and}\quad\min\left\{\,f(x^j):j=k'\colon k\,\right\}\le\bar f_k. \tag{8.3}$$

Our efficiency estimates involve the (problem and algorithm-dependent) quantities

$$\hat D:=\sup_k d_{S_*}(x^k)\quad\text{and}\quad\hat G:=\sup_k|g^k|. \tag{8.4}$$

To provide freedom for implementations, we allow for additional scaling factors

$$D_k \in [\, D_{\min}, D_{\max}\,] \subset (0,\infty) \quad \text{and} \quad G_k \in [\, G_{\min}, G_{\max}\,] \subset (0,\infty). \tag{8.5}$$

For a fixed $s \in [1/2, 1]$, we consider the following stepsizes and their *efficiency factors*:

$$\nu_k := \frac{D_k k^{-s}}{\max\{|g^k|, |g^k|^2/G_k\}} \text{ with } c_{(8.6)} := \max\{\hat{G}, G_{\min}, \hat{G}^2/G_{\min}\}\frac{\hat{D}^2 + D_{\max}^2}{D_{\min}}, \tag{8.6}$$

$$\nu_k := \frac{D_k k^{-s}}{\max\{G_k, |g^k|^2/G_k\}} \text{ with } c_{(8.7)} := \max\{G_{\max}, \hat{G}^2/G_{\min}\}\frac{\hat{D}^2 + D_{\max}^2}{D_{\min}}, \tag{8.7}$$

$$\nu_k := \frac{D_k k^{-s}}{|g^k|} \text{ with } c_{(8.8)} := \max\{\hat{G}, G_{\min}\}\frac{\hat{D}^2 + D_{\max}^2}{D_{\min}}, \tag{8.8}$$

$$\nu_k := \frac{D_k k^{-s}}{G_k} \text{ with } c_{(8.9)} := G_{\max}\frac{\hat{D}^2 + D_{\max}^2(\hat{G}/G_{\min})^2}{D_{\min}}, \tag{8.9}$$

where $|g^k|$ is replaced by $G_{\min}$ if $|g^k| = 0$. For such stepsizes, the sums involved in (8.2) may be bounded via the following lemma.

LEMMA 8.1. *For $k \geq 1$ and $s \in [1/2, 1]$, we have the following statements:*

(i) $\sum_{j=\lceil \frac{1}{2}k\rceil}^{k} j^{-2s} \leq 1 + \ln 2$ *and* $\sum_{j=\lceil \frac{1}{2}k\rceil}^{k} j^{-s} \geq (2 - 2^{1/2})(k+1)^{1-s}$.

(ii) $\sum_{j=1}^{k} j^{-2s} \leq \min\{\frac{2s}{2s-1}, 1 + \ln k\}$ *and* $\sum_{j=1}^{k} j^{-s} \geq \max\{\ln(k+1), (2 - 2^{1/2})(k+1)^{1-s}\}$.

Proof. For $s \in (1/2, 1)$, this follows from standard integration arguments (cf. [Nes89, p. 157]), using the facts that $\frac{2^{s-1}-1}{s-1} \geq 2 - 2^{1/2}$ for (i), $\frac{k^{1-2s}-1}{1-2s} \leq \ln k$, and $\frac{(k+1)^{1-s}-1}{1-s} \geq \ln(k+1)$ for (ii); the rest follows by continuity. ∎

We may now state our efficiency estimates for the stepsizes (8.6)–(8.9).

THEOREM 8.2. *For a fixed $s \in [1/2, 1]$, consider any stepsize rule from* (8.6)–(8.9) *and its efficiency factor c (e.g., $c := c_{(8.6)}$ for* (8.6)*). Then for each k we have*

$$\bar{f}_k - f_* \leq \bar{\epsilon}_k + \begin{cases} \dfrac{(1+\ln 2)c}{(4 - 2^{3/2})(k+1)^{1-s}} & \text{if} \quad k' = \left\lceil \dfrac{1}{2}k \right\rceil, \\ \dfrac{\min\left\{\frac{2s}{2s-1}, 1 + \ln k\right\} c}{\max\left\{2\ln(k+1), (4 - 2^{3/2})(k+1)^{1-s}\right\}} & \text{if} \quad k' = 1. \end{cases} \tag{8.10}$$

If the errors satisfy $\epsilon_k \leq C_\epsilon k^{-s}$ for some constant C_ϵ, and the stepsizes are chosen via (8.7) *or* (8.9)*, then we also have*

$$\bar{\epsilon}_k \leq \begin{cases} \dfrac{(1+\ln 2)C_\epsilon c_\epsilon}{(2 - 2^{1/2})(k+1)^{1-s}} & \text{if} \quad k' = \left\lceil \dfrac{1}{2}k \right\rceil, \\ \dfrac{\min\left\{\frac{2s}{2s-1}, 1 + \ln k\right\} C_\epsilon c_\epsilon}{\max\left\{\ln(k+1), (2 - 2^{1/2})(k+1)^{1-s}\right\}} & \text{if} \quad k' = 1, \end{cases} \tag{8.11}$$

where $c_\epsilon := \frac{\max\{G_{\max}, \hat{G}^2/G_{\min}\}D_{\max}}{G_{\min}D_{\min}}$ for (8.7) *and $c_\epsilon := \frac{D_{\max}G_{\max}}{D_{\min}G_{\min}}$ for* (8.9)*; also* (8.11) *holds with $c_\epsilon := \frac{\max\{G_{\min}, \hat{G}^2/G_{\min}\}D_{\max}}{G_{\min}D_{\min}}$ for* (8.6) *and $c_\epsilon := \frac{\max\{\hat{G}, G_{\min}\}D_{\max}}{G_{\min}D_{\min}}$ for* (8.8) *provided that $|g^k|$ is replaced by $\max\{|g^k|, G_{\min}\}$ in the stepsizes of* (8.6) *and* (8.8)*, in which case the bound* (8.10) *remains valid.*

Proof. For (8.10), it suffices to bound Δ_k in (8.2) by using $d_{S_*}(x^{k'}) \le \hat{D}$ (cf. (8.4)), and then $|g^k| \le \hat{G}$ and (8.5) together with Lemma 8.1 for the sums. For (8.11), the sums of $\bar{\epsilon}_k$ (cf. (8.1)) are estimated in a similar way. ∎

The estimates (8.10) and (8.11) combine nicely into an *overall* efficiency estimate.

Remark 8.3.

(i) It follows from general complexity results [BTMN01, Prop. 4.1] that for $\epsilon_k \equiv 0$ and n large enough, a *lower* bound on $\min_{j=1}^k f(x^j) - f_*$ is of order $O(k^{-1/2})$. Since (8.3) and (8.10) imply an *upper* bound of the same order for $s = 1/2$ and $k' = \lceil \frac{1}{2}k \rceil$, this choice is *optimal* from the complexity viewpoint. The switch from $k' = \lceil \frac{1}{2}k \rceil$ to $k' = 1$ degrades the bound moderately to $O(k^{-1/2} \ln k)$, but the popular choice of $s = 1$ has a much worse bound of $O(1/\ln k)$. On the other hand, for $s = 1/2$ we cannot have $\sum_k \nu_k^2 < \infty$ as required for convergence of $\{x^k\}$ in Theorem 3.4; however, choosing s slightly larger than $1/2$ combines the best of both worlds: convergence of $\{x^k\}$ and efficiency of order $O(k^{s-1})$ comparable to $O(k^{-1/2})$.

(ii) The stepsize (8.6) corresponds to (7.6) (with $\lambda_k := D_k k^{-s}$), (8.7) corresponds to both (7.7) and (7.8) (with $\lambda_k := (D_k/G_k)k^{-s}$ and $\lambda_k := D_k G_k k^{-s}$, respectively), and (8.8) corresponds to (7.10). For these stepsizes Theorems 7.1 and 7.11 ensure finiteness of $\hat{D}$ and $\hat{G}$ in (8.4) under reasonable conditions. The stepsize (8.9) may need the bounding strategies of section 6, e.g., for picking $D_{\max}$ small enough.

(iii) The efficiency factors of (8.6)–(8.9) are of order $2\hat{G}\hat{D}$ when $D_{\min} \approx D_{\max} \approx \hat{D}$, $G_{\min} \approx G_{\max} \approx \hat{G}$, but in general the values of $\hat{D}$ and $\hat{G}$ in (8.4) are stepsize-dependent.

In the language of Theorem 7.1(i), nonvanishing stepsizes ensure only asymptotic objective accuracy of order $\tilde{\delta} \approx \bar{\gamma}\lambda$ (for ϵ_k sufficiently small). In this context, efficiency is understood in terms of bounds on the relative accuracy $(\Delta_k - \tilde{\delta})/\tilde{\delta}$ (cf. (8.2)–(8.3)). Roughly speaking, for reasonable stepsizes such bounds have the form $(\hat{\Delta}/2\tilde{\delta})^2/k$, where $\hat{\Delta}$ measures the variation of f; a more precise statement is given below.

PROPOSITION 8.4. *For fixed $\lambda > 0$, $G > 0$, $D := d_{S_*}(x^1)$, and $\hat{G} := \sup_k |g^k|$, the stepsizes ν_k exhibited below have the following given efficiency bounds on Δ_k (cf.* (8.2)–(8.3) *with $k' = 1$):*

$$\nu_k := \frac{\lambda}{\max\{|g^k|, |g^k|^2/G\}} \quad \Rightarrow \quad \Delta_k \le \frac{1}{2}G\lambda\left(1 + \frac{\max\{\hat{G}, G\}^2 D^2}{(G\lambda)^2 k}\right), \tag{8.12}$$

$$\nu_k := \frac{\lambda}{\max\{1, |g^k|^2/G^2\}} \quad \Rightarrow \quad \Delta_k \le \frac{1}{2}G^2\lambda\left(1 + \frac{\max\{\hat{G}, G\}^2 D^2}{(G^2\lambda)^2 k}\right), \tag{8.13}$$

$$\nu_k := \frac{\lambda}{\max\{G^2, |g^k|^2\}} \quad \Rightarrow \quad \Delta_k \le \frac{1}{2}\lambda\left(1 + \frac{\max\{\hat{G}, G\}^2 D^2}{\lambda^2 k}\right), \tag{8.14}$$

$$\nu_k := \frac{\lambda}{|g^k|^2} \quad \Rightarrow \quad \Delta_k \le \frac{1}{2}\lambda\left(1 + \frac{\hat{G}^2 D^2}{\lambda^2 k}\right), \tag{8.15}$$

$$\nu_k := \frac{\lambda}{|g^k|} \quad \Rightarrow \quad \Delta_k \le \frac{1}{2}\hat{G}\lambda\left(1 + \frac{\hat{G}^2 D^2}{(\hat{G}\lambda)^2 k}\right), \tag{8.16}$$

$$\nu_k := \lambda \quad \Rightarrow \quad \Delta_k \le \frac{1}{2}\hat{G}^2\lambda\left(1 + \frac{\hat{G}^2 D^2}{(\hat{G}^2\lambda)^2 k}\right). \tag{8.17}$$

Here we assume that $|g^k|$ is replaced by G in (8.12) *and* (8.15)–(8.16) *whenever $|g^k| = 0$ and for* (8.15)–(8.16) *that G is reset to $|g^k|$ when $|g^k|$ becomes nonzero.*

Proof. Recalling the definition (8.2) of Δ_k, simple calculations yield the conclusion. ∎

9. Analysis of the incremental subgradient method.

9.1. Basic incremental estimates. Throughout this section, $\{x^k\}$, $\{\nu_k\}$, $\{x_i^k\}$, $\{\epsilon_i^k\}$, and $\{g_i^k\}$ denote the sequences involved in the incremental subgradient iteration (1.4). Further, for each k, we let

$$f_{\rm inc}^k := \sum_{i=1}^m f_i(x_i^k), \tag{9.1}$$

$$\epsilon_k := \sum_{i=1}^m \epsilon_i^k, \tag{9.2}$$

$$\bar{C}_k := \sum_{i=1}^m \bar{C}_{ik} \quad \text{with} \quad \bar{C}_{ik} := \max\left\{\, |g_i^k|, |\bar{g}_i^k| \,\right\} \quad \text{for some} \quad \bar{g}_i^k \in \partial f_i^S(x^k). \tag{9.3}$$

Note that the *incremental* objective value $f_{\rm inc}^k$ is a natural estimate for $f(x^k)$, and the additional subgradients $\bar{g}_i^k$ provide only bounds on $f(x^k) - f_{\rm inc}^k$ (cf. (9.8), (9.11)).

We start by extending the basic estimates of Lemma 3.1 to the incremental case.

LEMMA 9.1. *For each x and $k \geq 1$, we have*

$$|x^{k+1} - x|^2 - |x^k - x|^2 \leq -2\nu_k \left[\, f(x^k) - f_S(x) - \epsilon_k - \tfrac{1}{2}\bar{C}_k^2\nu_k \,\right], \tag{9.4}$$

$$\frac{\sum_{j=1}^k \nu_j f(x^j)}{\sum_{j=1}^k \nu_j} - f_S(x) \leq \frac{\frac{1}{2}|x^1 - x|^2 + \sum_{j=1}^k \frac{1}{2}\nu_j^2\bar{C}_j^2 + \sum_{j=1}^k \nu_j\epsilon_j}{\sum_{j=1}^k \nu_j}, \tag{9.5}$$

$$|x^{k+1} - x^k| \leq \nu_k\bar{C}_k, \tag{9.6}$$

$$|x^{k+1} - x|^2 - |x^k - x|^2 \leq -2\nu_k \left[f_{\rm inc}^k - f_S(x) - \epsilon_k - \tfrac{1}{2}\nu_k \sum_{i=1}^m |g_i^k|^2 \right], \tag{9.7}$$

$$f(x^k) - f_{\rm inc}^k \leq \nu_k \sum_{i=1}^m \bar{C}_{ik} \sum_{j=1}^{i-1} |g_j^k| \leq \nu_k \sum_{i=1}^m \bar{C}_{ik} \sum_{j=1}^{i-1} \bar{C}_{jk}, \tag{9.8}$$

$$f_{\rm inc}^k - f(x^k) - \epsilon_k \leq \nu_k \sum_{i=1}^m |g_i^k| \sum_{j=1}^{i-1} |g_j^k| \leq \nu_k \sum_{i=1}^m \bar{C}_{ik} \sum_{j=1}^{i-1} \bar{C}_{jk}, \tag{9.9}$$

$$|x_i^k - x^k| \leq \nu_k \sum_{j=1}^{i-1} |g_j^k| \leq \nu_k \sum_{j=1}^{i-1} \bar{C}_{jk} \quad \textit{for } i = 1\colon m+1. \tag{9.10}$$

Proof. Let $x \in S$, $r_{ik} := |x_i^k - x|$. Using the nonexpansiveness of P_S and (1.4) gives

$$\begin{aligned} r_{i+1,k}^2 &\le |x_i^k - \nu_k g_i^k - x|^2 = r_{ik}^2 - 2\nu_k \langle g_i^k, x_i^k - x\rangle + \nu_k^2 |g_i^k|^2 \\ &\le r_{ik}^2 + 2\nu_k \left[f_i(x) - f_i(x_i^k) + \epsilon_i^k \right] + \nu_k^2 |g_i^k|^2; \end{aligned}$$

sum up and use $r_k := |x^k - x|$, $x^{k+1} := x_{m+1}^k$, and (9.1)–(9.2) to get (9.7). Since $|x_{i+1}^k - x^k| \le |x_i^k - x^k| + |x_{i+1}^k - x_i^k|$, where $|x_{i+1}^k - x_i^k| \le \nu_k |g_i^k|$ by (1.4), (9.10) follows by induction. Summing $f_i(x^k) - f_i(x_i^k) \le \langle \bar g_i^k, x^k - x_i^k\rangle$ (cf. (9.3)) and using (9.1) and (9.10), we obtain

(9.11)
$$f(x^k) - f_{\rm inc}^k = \sum_i \left[f_i(x^k) - f_i(x_i^k) \right] \le \sum_i |\bar g_i^k| |x_i^k - x^k| \le \nu_k \sum_i \bar C_{ik} \sum_{j<i} |g_j^k|$$

and hence (9.8); similarly, summing $f_i(x_i^k) - f_i(x^k) - \epsilon_i^k \le \langle g_i^k, x_i^k - x^k\rangle$ (cf. (1.4)) gives (9.9). Then (9.7), (9.8), and (9.3) yield (9.4), since $2\sum_i \bar C_{ik} \sum_{j<i} \bar C_{jk} + \sum_i \bar C_{ik}^2 = \bar C_k^2$. Summing up (9.4) gives (9.5). For $f_S(x) = \infty$, (9.4), (9.5), and (9.7) are trivial. Finally, (9.6) follows from (9.10) with $i = m+1$, using $x^{k+1} := x_{m+1}^k$ and (9.3). ∎

9.2. General incremental convergence results. All the convergence results of sections 3 and 4 extend easily to the incremental method.

COROLLARY 9.2. *Theorems* 3.2, 3.4, 3.6, 4.1, *and Corollary* 4.2 *hold for the incremental subgradient method* (1.4) *with* $|g^k|$ *replaced by* $\bar C_k$ (*so that* $\gamma_k := \frac12 \bar C_k^2 \nu_k$ *in* (3.5) *and* $C := \overline{\lim}_{k\to\infty} \bar C_k$ *in Theorems* 3.2(vi) *and* 4.1(iv)).

Proof. Comparing (3.1)–(3.3) with (9.4)–(9.6), we may replace $|g^k|$ by $\bar C_k$ in the proofs of sections 3.2–3.3 and section 4. ∎

We now give a more refined version of Corollary 4.2 for the incremental case that employs a slightly weaker assumption (boundedness of $|g_i^k|$ instead of $\max\{|g_i^k|, |\bar g_i^k|\}$).

LEMMA 9.3. *Suppose that* f_S *is coercive,* $\hat\nu := \sup_k \nu_k < \infty$, $\hat\epsilon := \sup_k \epsilon_k < \infty$, *and* $C_i := \sup_k |g_i^k| < \infty$ *for all* i. *Then* $\{x^k\}$ *and* $\{x_i^k\}$ *are bounded for all* i.

Proof. Let $x \in S_*$, $C := \sum_i C_i$, $\sigma := C\hat\nu$, and $\alpha := f_* + \hat\epsilon + \frac12 C^2 \hat\nu$. Since $f(x) = f_*$ and f_S is coercive, x lies in the bounded set $T_{\alpha,\sigma}$ (cf. (2.3)). First, suppose that $f_{\rm inc}^k \le \alpha$. By (9.10) with $\nu_k \le \hat\nu$, we have $\max_i |x_i^k - x^k| \le \nu_k C \le \sigma$. Hence $x^k \in T_{\alpha,\sigma}$ (cf. (2.3) and (9.1)) and $|x^{k+1} - x^k| \le \sigma$ (since $x^{k+1} := x_{m+1}^k$). Thus

$$|x^{k+1} - x| \le |x^k - x| + |x^{k+1} - x^k| \le \operatorname{diam}(T_{\alpha,\sigma}) + \sigma \quad \text{if} \quad f_{\rm inc}^k \le \alpha. \tag{9.12}$$

Second, if $f_{\rm inc}^k > \alpha$, i.e., $f_{\rm inc}^k > f(x) + \hat\epsilon + \frac12 C^2\hat\nu$, then by using the bounds $\nu_k \le \hat\nu$, $\epsilon_k \le \hat\epsilon$, and $\sum_i |g_i^k|^2 \le \sum_i C_i^2 \le C^2$ in (9.7), we obtain

(9.13)
$$|x^{k+1} - x|^2 - |x^k - x|^2 \le -2\nu_k \left[\tfrac12 C^2 \hat\nu + \hat\epsilon - \epsilon_k - \tfrac12 \nu_k C^2 \right] \le 0 \quad \text{if} \quad f_{\rm inc}^k > \alpha.$$

Combining (9.12) and (9.13) gives $|x^k - x| \le \max\{\operatorname{diam}(T_{\alpha,\sigma}) + \sigma, |x^1 - x|\}$ for all k. Thus $\{x^k\}$ is bounded, and so are $\{x_i^k\}$ for all i, since $\max_i |x_i^k - x^k| \le \sigma$. ∎

Of course, in the incremental case Definition 6.1 is replaced by the following definition.

DEFINITION 9.4. *We say that the algorithm employs a* locally bounded oracle *if* $g_i^k = g_i(x^k, \epsilon_i^k)$ *and* $\bar{g}_i^k = g_i(x^k, 0)$ *for all* i *and* k, *where the mappings* $S \times \mathbb{R}_+ \ni (x, \epsilon) \mapsto g_i(x, \epsilon) \in \partial_\epsilon f_i^S(x)$ *are locally bounded.*

The following result complements Lemma 9.3 and enables us to extend Theorem 5.1 to the incremental method.

LEMMA 9.5. *Suppose that* $\{x^k\}$ *is bounded and* $\hat{\nu} := \sup_k \nu_k < \infty$. *Then we have the following statements:*

(i) *If the oracle is locally bounded and* $\hat{\epsilon} := \sup_k \epsilon_k < \infty$, *then* $\{x_i^k\}$ *is bounded for all* i, *and* $\sup_k \bar{C}_k < \infty$.

(ii) *If* $\sup_k \bar{C}_k < \infty$, *then* $\{x_i^k\}$ *is bounded for all* i.

Proof. (i) By Definition 9.4, $\{\bar{g}_i^k = g_i(x^k, 0)\}$ is bounded for all i. Assuming $C_j := \sup_k \bar{C}_{jk} < \infty$ for $j < i$, by (9.10) we have $|x_i^k - x^k| \le \hat{\nu} \sum_{j<i} C_j$ ($x_i^k = x^k$ if $i = 1$). Thus $\{x_i^k\}$ is bounded, and so is $\{g_i^k = g_i(x_i^k, \epsilon_i^k)\}$ because the oracle is locally bounded. Hence, by (9.3), $C_i := \sup_k \bar{C}_{ik}$ is finite. The rest follows by induction, with $\sup_k \bar{C}_k \le \sum_i C_i$.

(ii) This follows from (9.3) and (9.10) with $\nu_k \le \hat{\nu}$. ∎

COROLLARY 9.6. *Theorem* 5.1 *holds for the incremental subgradient method* (1.4) *with* $|g^k|$ *replaced by* $\bar{C}_k$ (*so that* $\gamma_k := \frac{1}{2}\bar{C}_k^2 \nu_k$) *and* $\underline{R}$ *redefined as* $\underline{R} := \sup_{i,k} |x_i^k|$.

Proof. The assumptions of Theorem 5.1 and Lemma 9.5 yield $\underline{R} < \infty$. Next, in the proof of Theorem 5.1, we may replace S and f_i^S in (1.4) by $S' := S \cap B_R$ and $f_i^{S'} := f_i^S + \mathrm{I}_{\mathrm{B}_R}$, since $\{x_i^k\} \subset S'$, whereas $g_i^k \in \partial_{\epsilon_i^k} f_i^S(x_i^k)$ implies $g_i^k \in \partial_{\epsilon_i^k} f_i^{S'}(x_i^k)$. In view of Corollary 9.2, the proof may be finished as before. ∎

Theorems 7.17 and 7.19 also may be extended to the incremental case.

COROLLARY 9.7. *Theorems* 7.17 *and* 7.19 *hold for the incremental subgradient method* (1.4) *if* $|g^k|$ *in* (7.27) *and* (7.33) *is replaced by a constant* $C \in (0, \infty)$ *such that* $C \ge \sup_k \bar{C}_k$.

Proof. Replace $|g^k|$ by C in the original proofs, invoking (9.4) instead of (3.1). ∎

Remark 9.8. Our framework is more general than that of [NeB01, sect. 2], where each f_i is finite-valued and $g_i^k \in \partial f_i(x_i^k)$ in (1.4); i.e., $\epsilon_i^k \equiv 0$ and the oracle is locally bounded. The basic assumption of [NeB01, Ass. 2.1] is $\sup_k \bar{C}_i^k < \infty$ for all i. Theorem 3.2(ii), (vi) subsumes [NeB01, Props. 2.1–2.2] (with $C := \sup_k \bar{C}_k$), Theorem 3.4 subsumes [NeB01, Prop. 2.4], and Theorem 4.1(ii) subsumes [NeB01, Prop. 2.3] (with $\nu = 0$). Corollary 9.7 subsumes [NeB01, Props. 2.5–2.6].

9.3. Incremental bounding strategies. We now extend Theorems 6.3 and 6.4 to the incremental case.

THEOREM 9.9. *Suppose* f_S *is coercive and the algorithm employs a locally bounded oracle. Fix any point* $\bar{x} \in S$ *and a tolerance* $\bar{\delta} \in (0, \infty)$. *Then there exist thresholds* $\bar{\nu}_{\max} > 0$ *and* $\bar{\epsilon}_{\max} > 0$ *with the following property*: *If the algorithm starts from a point* $x^1 \in T_{f(\bar{x})}$ (*e.g.*, $x^1 = \bar{x}$) *and employs stepsizes* $\nu_k \le \bar{\nu}_{\max}$ *and errors* $\epsilon_k \le \bar{\epsilon}_{\max}$ *for all* k, *then* x^k *stays in the bounded trench* $T_{f(\bar{x})+\bar{\delta}}$ *and* $f_{\mathrm{inc}}^k \le f(\bar{x}) + 2\bar{\delta}$ *for all* k, *and there exist* $C_i < \infty$ *such that* $\bar{C}_{ik} := \max\{|g_i^k|, |\bar{g}_i^k|\} \le C_i$ *and* $|x_i^k - x^k| \le \nu_k \sum_{j<i} C_j$ *for all* k *and* i.

Proof. Let $\beta := f(\bar{x})$, $\bar{\alpha} := \beta + \bar{\delta}$. Since the oracle is locally bounded, f_S is continuous on S (cf. Remark 6.2(iii)). By Lemma 2.4(ii), there exists $\bar{\rho} > 0$ such that $S \cap (T_\beta + B_{3\bar{\rho}}) \subset T_{\bar{\alpha}}$, whereas by Lemma 2.4(i) there is $\alpha \in (\beta, \bar{\alpha})$ such that

$T_\beta^\alpha \subset T_\beta + B_{\bar\rho}$; thus

$$S \cap \left(T_\beta^\alpha + B_{\bar\rho} \right) \subset S \cap \left(T_\beta + B_{2\bar\rho} \right) \subset S \cap \left(T_\beta + B_{3\bar\rho} \right) \subset T_{\bar\alpha}. \tag{9.14}$$

Let

$$\bar\epsilon_{\max} := \tfrac{1}{2}(\alpha - \beta), \tag{9.15}$$

$$C := \sum_i C_i \quad \text{with} \quad C_i := \sup \left\{ |g_i(x,\epsilon)| : x \in S \cap (T_\beta + B_{3\bar\rho}), \epsilon \le \bar\epsilon_{\max} \right\}, \tag{9.16}$$

$$\bar\nu_{\max} := \min \left\{ \bar\rho / C, (\alpha - \beta)/C^2 \right\}. \tag{9.17}$$

Note that $C < \infty$, since T_β is bounded and $\bar\epsilon_{\max} < \infty$.

Since $\{x^k\} \subset S$ and $f(x^1) \le f(\bar x) =: \beta$, we have $x^1 \in S \cap (T_\beta + B_{2\bar\rho})$.

Assuming $x^k \in S \cap (T_\beta + B_{2\bar\rho})$ for some $k \ge 1$, we now show that $x^{k+1} \in S \cap (T_\beta + B_{2\bar\rho})$. First, note that, by induction as for (9.10), we have $|g_i^k| \le C_i$ for $i = 1\colon m$ and

$$|x_i^k - x^k| \le \nu_k \sum_{j<i} |g_j^k| \le \bar\nu_{\max} \sum_{j<i} C_j \le \bar\rho \quad \text{for} \quad i = 1\colon m+1. \tag{9.18}$$

Indeed, suppose (9.18) holds for some $i \le m$. (Recall that $x_1^k = x^k$.) Then $|x_{i+1}^k - x^k| \le |x_i^k - x^k| + |x_{i+1}^k - x_i^k|$, where $|x_{i+1}^k - x_i^k| \le \nu_k |g_i^k|$ by (1.4) with $|g_i^k| = |g_i(x_i^k, \epsilon_i^k)| \le C_i$ (cf. (9.16)) because $\epsilon_i^k \le \bar\epsilon_{\max}$ and $x_i^k \in T_\beta + B_{3\bar\rho}$ from $x^k \in T_\beta + B_{2\bar\rho}$ and $|x_i^k - x^k| \le \bar\rho$. Thus (9.18) holds for i increased by 1, with the final inequality due to (9.17). Further, (9.3) and (9.16) give $\bar C_{ik} \le C_i$ and $\bar C_k \le C$, using $|\bar g_i^k| = |g_i(x^k, 0)| \le C_i$. If $x^k \in T_\alpha$, then $T_\alpha \subset T_\beta^\alpha$ (cf. (2.2)), and the first inclusion of (9.14) and (9.18) with $x^{k+1} := x_{m+1}^k$ yield

$$x^{k+1} \in S \cap \left(x^k + B_{\bar\rho} \right) \subset S \cap \left(T_\alpha + B_{\bar\rho} \right) \subset S \cap \left(T_\beta^\alpha + B_{\bar\rho} \right) \subset S \cap \left(T_\beta + B_{2\bar\rho} \right).$$

Next, suppose $x^k \notin T_\alpha$, i.e.,

$$f(x^k) > \alpha. \tag{9.19}$$

Since $x^k \in S \cap (T_\beta + B_{2\bar\rho})$, we have $|x^k - x| \le 2\bar\rho$ for $x = P_{T_\beta} x^k$. By (9.15) and (9.17),

$$\epsilon_k \le \bar\epsilon_{\max} \le \tfrac{1}{2}(\alpha - \beta) \quad \text{and} \quad \tfrac{1}{2}\bar C_k^2 \nu_k \le \tfrac{1}{2} C^2 \bar\nu_{\max} \le \tfrac{1}{2}(\alpha - \beta). \tag{9.20}$$

Using the estimate (9.4) with $f_S(x) \le \beta$ and the bounds (9.19) and (9.20), we obtain

$$|x^{k+1} - x|^2 - |x^k - x|^2 \le -2\nu_k \left[f(x^k) - f(x) - \epsilon_k - \tfrac{1}{2}\bar C_k^2 \nu_k \right] \le 0.$$

Thus $|x^{k+1} - x| \le |x^k - x| \le 2\bar\rho$ with $x \in T_\beta$, so $x^{k+1} \in S \cap (T_\beta + B_{2\bar\rho})$.

Therefore, by induction, we have $x^k \in S \cap (T_\beta + B_{2\bar\rho}) \subset T_{\bar\alpha}$ (cf. (9.14)), $\bar C_{ik} \le C_i$, and (9.18) for all k. Finally, using (9.9) with $f(x^k) \le \bar\alpha$ and $\sum_i \bar C_{ik} \sum_{j<i} \bar C_{jk} \le \frac{1}{2}\bar C_k^2$ together with (9.20) gives $f_{\text{inc}}^k \le \bar\alpha + \alpha - \beta \le \beta + 2\bar\delta$, since $\alpha < \bar\alpha := \beta + \bar\delta$. ∎

THEOREM 9.10. *Suppose f_S is coercive and the algorithm employs a locally bounded oracle. Then for each $\beta \in (f_*, \infty)$ and $\bar\epsilon_{\max} \in [0, \infty)$ there exists $\bar\nu_{\max} > 0$*

such that if $f_S(x^1) \le \beta$, $\nu_k \le \bar{\nu}_{\max}$, *and* $\epsilon_k \le \bar{\epsilon}_{\max}$ *for all* k, *then* $\{x_i^k\}$, $\{g_i^k\}$ *and* $\{\bar{g}_i^k\}$ *are bounded for all* i.

Proof. Modify the proof of Theorem 9.9 as in the proof of Theorem 6.4. □

In view of Theorems 9.9–9.10, for the incremental method we may use the bounding strategy with the resetting test (6.7) or the strategy inspired by Theorem 6.4 with the test (6.8) replaced by $\max_i |x_i^k| > R_l$.

Yet another bounding strategy stems from the following result.

LEMMA 9.11. *Suppose that* f_S *is coercive and there exist* $\alpha \in \mathbb{R}$ *and* $\sigma \in \mathbb{R}_+$ *such that* $f_{\rm inc}^k \le \alpha$ *and* $\max_i |x_i^k - x^k| \le \sigma$ *for all* k. *Then* $\{x^k\}$ *is bounded.*

Proof. By (2.3) and (9.1), $\{x^k\}$ lies in the bounded set $T_{\alpha,\sigma}$ (cf. Lemma 2.5). □

Lemma 9.11 suggests the following bounding strategy with resets indexed by $l = 1, 2, \ldots$. Fixing $\bar{x} \in S$, $\bar{\delta} \in (0, \infty)$, and $\bar{\sigma} \in (0, \infty)$, pick positive sequences $\nu_{\max}^l \to 0$ and $\epsilon_{\max}^l \to 0$ as $l \to \infty$. For the current $l \ge 1$, start the algorithm from $\bar{x}$ (or the best point found so far if $l > 1$), using stepsizes $\nu_k \le \nu_{\max}^l$ and errors $\epsilon_k \le \epsilon_{\max}^l$; if for some k

$$f_{\rm inc}^k > f(\bar{x}) + 2\bar{\delta} \quad \text{or} \quad \max_i |x_i^k - x^k| > \bar{\sigma}, \tag{9.21}$$

then increase l by 1, restart the algorithm, etc. Under the assumptions of Theorem 9.9, only finitely many resets occur, so Lemmas 9.5(i) and 9.11 imply the boundedness of $\{x_i^k\}$ and $\{\bar{C}_k\}$. (A special case of this strategy consists of using sequences $\nu_k \to 0$ and $\epsilon_k \to 0$, and resetting x^{k+1} to x^1 whenever (9.21) holds.)

9.4. Incremental efficiency estimates. Following section 8, in this subsection we assume that the optimal set S_* is nonempty, and that the sequences $\{x^k\}$, $\{\bar{C}_k\}$ (cf. (9.3)), and $\{\epsilon_k\}$ are bounded. Thus, replacing (8.4) by

$$\hat{D} := \sup_k d_{S_*}(x^k) \quad \text{and} \quad \hat{G} := \sup_k \bar{C}_k, \tag{9.22}$$

we have

$$\bar{C}_k := \sum_{i=1}^m \bar{C}_{ik} \le \hat{G} \le m\hat{G}_{\max} \quad \text{with} \quad |g_i^k| \le \bar{C}_{ik} \le \hat{G}_{\max} := \max_i \sup_k \bar{C}_{ik}. \tag{9.23}$$

We now give estimates for the Cesáro averages of the objective values $\bar{f}_k$ (cf. (8.1)), the Cesáro averages of the incremental objective values (cf. (9.1)) defined by

$$\bar{f}_{\rm inc}^k := \sum_{j=k'}^k \nu_j f_{\rm inc}^j / \nu_{\rm sum}^k \quad \text{with} \quad \nu_{\rm sum}^k := \sum_{j=k'}^k \nu_j, \tag{9.24}$$

and the objective values of the *incremental record points* (cf. [BTMN01, sect. 5])

$$\breve{x}^k := x^{\breve{k}} \quad \text{with} \quad \breve{k} \in \operatorname{Arg\,min}\{ f_{\rm inc}^j : k' \le j \le k \}. \tag{9.25}$$

LEMMA 9.12. *In the notation of* (8.1), (9.23), (9.24), *and* (9.25), *we have*

$$\bar f_k - f_* \le \Delta_k + \bar\epsilon_k, \quad \Delta_k := \frac{d_{S_*}^2(x^{k'}) + \hat G^2 \sum_{j=k'}^k \nu_j^2}{2\sum_{j=k'}^k \nu_j}, \tag{9.26}$$

$$\bar f_{\rm inc}^k - f_* \le \bar\Delta_k + \bar\epsilon_k, \quad \bar\Delta_k := \frac{d_{S_*}^2(x^{k'}) + \min\{\hat G^2, m\hat G_{\max}^2\} \sum_{j=k'}^k \nu_j^2}{2\sum_{j=k'}^k \nu_j}, \tag{9.27}$$

$$f(\breve x^k) - f_* \le \breve\Delta_k + \bar\epsilon_k, \quad \breve\Delta_k := \bar\Delta_k + \frac{m-1}{2m}\hat G^2 \max_{j=k':\,k} \nu_j. \tag{9.28}$$

Proof. Replace $|g^j|$ by $\bar C_j$ in (8.2) (cf. (9.5) and the proof of Corollary 9.2) and use (9.23) to get (9.26). Summing up (9.7) and using (9.24) and (8.1) (for $\bar\epsilon_k$) yields

$$\bar f_{\rm inc}^k - f_S(x) \le \frac{|x^{k'} - x|^2 + \sum_{j=k'}^k \nu_j^2 \sum_{i=1}^m |g_i^j|^2}{2\sum_{j=k'}^k \nu_j} + \bar\epsilon_k \quad \forall x. \tag{9.29}$$

Letting $x := P_{S_*} x^{k'}$ in (9.29) and bounding $\sum_i |g_i^j|^2 \le \min\{m\hat G_{\max}^2, \hat G^2\}$ (cf. (9.23)), we get (9.27). Next, we have $f_{\rm inc}^{\breve k} = \min_{j=k'}^k f_{\rm inc}^j \le \bar f_{\rm inc}^k$ by (9.24) and (9.25), whereas by (9.8) and (9.23)

$$f(\breve x^k) = f(x^{\breve k}) \le f_{\rm inc}^{\breve k} + \nu_{\breve k} \sum_{i=1}^m \bar C_{i\breve k} \sum_{j=1}^{i-1} \bar C_{j\breve k} \le f_{\rm inc}^{\breve k} + \nu_{\breve k} \tfrac12 \hat G^2 (1 - \tfrac1m)$$

(since $\sum_i \bar C_{i,\breve k}^2 \ge \frac1m \bar C_{\breve k}^2$); combining these bounds with (9.27) gives (9.28). ∎

The estimate (9.26) bounds the objective values $f(\bar x^k) \le \bar f_k$ and $f(x_{\rm rec}^k) \le \bar f_k$ of the Cesáro points $\bar x^k$ and the record points $x_{\rm rec}^k$ (cf. (3.14), (8.3)).

We may now present efficiency estimates for stepsizes analogous to those of (8.9).

THEOREM 9.13. *Consider the following two stepsize rules and their efficiency factors:*

$$\nu_k := \frac{D_k k^{-s}}{G_k} \quad \textit{with} \quad c_{(9.30)} := G_{\max} \frac{\hat D^2 + D_{\max}^2 (\hat G / G_{\min})^2}{D_{\min}}, \tag{9.30}$$

$$\nu_k := \frac{D_k k^{-s}}{m G_k} \quad \textit{with} \quad c_{(9.31)} := m G_{\max} \frac{\hat D^2 + D_{\max}^2 (\hat G_{\max} / G_{\min})^2}{D_{\min}}, \tag{9.31}$$

where $s \in [1/2, 1]$, $\hat D$, $\hat G$ *and* $\hat G_{\max}$ *are defined by* (9.22)–(9.23), *and* D_k *and* G_k *are scaling factors that satisfy* (8.5). *Then for each rule we have for all* k

(9.32)
$$\bar f_k - f_* \le \bar\epsilon_k + \begin{cases} \dfrac{(1+\ln 2)c}{(4-2^{3/2})(k+1)^{1-s}} & \textit{if} \quad k' = \left\lceil \dfrac12 k \right\rceil, \\[2ex] \dfrac{\min\left\{\frac{2s}{2s-1}, 1+\ln k\right\} c}{\max\left\{2\ln(k+1), (4-2^{3/2})(k+1)^{1-s}\right\}} & \textit{if} \quad k' = 1, \end{cases}$$

where $c := c_{(9.30)}$ *for the rule* (9.30) *and* $c := c_{(9.31)}$ *for the rule* (9.31). *Moreover, for the incremental record points* $\breve x^k$ *defined by* (9.25) *with* $k' = \lceil \frac12 k \rceil$, *we have for*

each k

(9.33)
$$f(\breve{x}^k) - f_* \le \bar{\epsilon}_k + \frac{(1+\ln 2)c}{(4-2^{3/2})(k+1)^{1-s}} + \frac{D_{\max}}{2^{1-s}G_{\min}k^s} \begin{cases} \frac{m-1}{m}\hat{G}^2 & \text{for (9.30)}, \\ (m-1)\hat{G}^2_{\max} & \text{for (9.31)}, \end{cases}$$

$$f(\breve{x}^k) - f_* \le \bar{\epsilon}_k + \frac{(1+\ln 2)c}{(4-2^{3/2})k^{1/2}} \quad \text{for} \quad s = 1/2, \tag{9.34}$$

where $c := \frac{3}{2}c_{(9.30)}$ *for the rule* (9.30) *and* $c := c_{(9.31)}$ *for the rule* (9.31). *Further, if* $C_\epsilon := \sup_k k^s\epsilon_k$ *is finite, then the estimate* (8.11) *holds with* $c_\epsilon := \frac{D_{\max}G_{\max}}{D_{\min}G_{\min}}$ *so that* $\bar{\epsilon}_k$ *in* (9.32)–(9.34) *has the same order in* k *as its right neighbors.*

Proof. It suffices to bound Δ_k in (9.26) and $\breve{\Delta}_k$ in (9.28) by using $d_{S_*}(x^{k'}) \le \hat{D}$ (cf. (9.22)) and (8.5) together with Lemma 8.1 for the sums. ∎

Remark 9.14.

(i) For both stepsize rules (9.30)–(9.31), D_k should be a guess for $d_{S_*}(x^k)$ (or for the "diameter of the picture"), but for the first one G_k should be a guess for $\hat{G}$ (e.g., $\sum_i |g_i^{k-1}|$), whereas for the second one G_k should be a guess for $\hat{G}_{\max}$ (e.g., $\max_i |g_i^{k-1}|$).

(ii) For comparisons, suppose the feasible set S is bounded and the subgradients of each objective f_i are exact ($\epsilon_i^k \equiv 0$) and bounded by its Lipschitz constant L_{f_i} on S so that $\hat{D}$ may be replaced by $\operatorname{diam}(S)$, $\hat{G}$ by $\sum_i L_{f_i}$, and $\hat{G}_{\max}$ by $\max_i L_{f_i}$. Further, assume that $D_{\min}$ and $D_{\max}$ are of order $\hat{D}$, $G_{\min}$ and $G_{\max}$ are of order $\hat{G}$ for (9.30) and $\hat{G}_{\max}$ for (9.31) so that $c_{(9.30)} \approx 2\operatorname{diam}(S)\sum_i L_{f_i}$ and $c_{(9.31)} \approx 2\operatorname{diam}(S)m\max_i L_{f_i}$. Under similar assumptions, the nonincremental version has $c_{(8.9)} \approx 2\operatorname{diam}(S)L_f$, where L_f is the Lipschitz constant of f on S. Of course, $L_f \le \sum_i L_{f_i} \le m\max_i L_{f_i}$. Assuming that $\max_i L_{f_i} \le L_f$ (as in [BTMN01, Thm. 5.1]), the efficiency estimates for the incremental version given in Theorem 9.13 are at most m times larger than those for the ordinary version stated in Theorem 8.2; yet their ratio decreases when $\sum_i L_{f_i}$ gets closer to L_f; i.e., all f_i become "similar." Such "similarity" features help the incremental version to be competitive in practice [BTMN01, NeB01].

(iii) Remark 8.3(i) on the choice of s and k' remains valid.

(iv) In the exact case of $\epsilon_k \equiv 0$, our estimate (9.34) for the stepsize rule (9.31) is similar to that of [BTMN01, Thm. 5.1] (for the Euclidean norm).

For nonvanishing stepsizes $\nu_k \equiv \nu$, the asymptotic objective accuracy is of order $\frac{1}{2}\hat{G}^2\nu \le \frac{1}{2}m^2\hat{G}^2_{\max}\nu$ (cf. Corollary 9.2, Thm. 3.2, and (9.22)–(9.23)), and the relative accuracy may be estimated as in Proposition 8.4 (cf. (8.17)).

PROPOSITION 9.15. *For a fixed stepsize* $\nu_k \equiv \nu > 0$, *we have the following efficiency bounds on* Δ_k *and* $\breve{\Delta}_k$ *defined by* (9.26) *and* (9.28) *with* $k' = 1$:

$$\Delta_k \le \frac{1}{2}\hat{G}^2\nu\left(1 + \frac{\hat{G}^2D^2}{(\hat{G}^2\nu)^2k}\right), \tag{9.35}$$

$$\breve{\Delta}_k \le \frac{1}{2}\hat{G}^2\nu\left(1 + \frac{m-1}{m} + \frac{\hat{G}^2D^2}{(\hat{G}^2\nu)^2k}\right), \tag{9.36}$$

$$\max\left\{\Delta_k, \check{\Delta}_k\right\} \le \frac{1}{2}m^2\hat{G}_{\max}^2\nu\left(1+\frac{m^2\hat{G}_{\max}^2 D^2}{(m^2\hat{G}_{\max}^2\nu)^2 k}\right), \tag{9.37}$$

where $D := d_{S_*}(x^1)$, $\hat{G}_{\max} := \sup_{i,k}\bar{C}_{ik}$, *and* $\hat{G} := \sup_k \bar{C}_k \le m\hat{G}_{\max}$.

Proof. This follows easily from the definitions (9.26) and (9.28). ∎

Acknowledgments. I would like to thank the Associate Editor and the two anonymous referees for helpful comments. Further, I would like to acknowledge extensive discussions with Ya. I. Alber, D. Bertsekas, O. Burdakov, C. Lemaréchal, and A. Nedić.

REFERENCES

[AIS98] Ya. I. Alber, A. N. Iusem, and M. V. Solodov, *On the projected subgradient method for nonsmooth convex optimization in a Hilbert space*, Math. Programming, 81 (1998), pp. 23–35.

[Ber97] D. P. Bertsekas, *A new class of incremental gradient methods for least squares problems*, SIAM J. Optim., 7 (1997), pp. 913–926.

[Ber99] D. P. Bertsekas, *Nonlinear Programming*, 2nd ed., Athena Scientific, Belmont, MA, 1999.

[BeT00] D. P. Bertsekas and J. N. Tsitsiklis, *Gradient convergence in gradient methods with errors*, SIAM J. Optim., 10 (2000), pp. 627–642.

[Brä93] U. Brännlund, *On Relaxation Methods for Nonsmooth Convex Optimization*, Ph.D. thesis, Department of Mathematics, Royal Institute of Technology, Stockholm, 1993.

[Brä95] U. Brännlund, *A generalized subgradient method with relaxation step*, Math. Programming, 71 (1995), pp. 207–219.

[BSS93] M. S. Bazaraa, H. D. Sherali, and C. M. Shetty, *Nonlinear Programming: Theory and Algorithms*, 2nd ed., Wiley, New York, 1993.

[BTMN01] A. Ben-Tal, T. Margalit, and A. Nemirovski, *The ordered subsets mirror descent optimization method with applications to tomography*, SIAM J. Optim., 12 (2001), pp. 79–108.

[CoL93] R. Correa and C. Lemaréchal, *Convergence of some algorithms for convex minimization*, Math. Programming, 62 (1993), pp. 261–275.

[DeV81] V. F. Demyanov and L. V. Vasilev, *Nondifferentiable Optimization*, Nauka, Moscow, 1981 (in Russian); Optimization Software Inc., New York, 1985 (in English).

[DuS88] N. Dunford and J. T. Schwartz, *Linear Operators, Part* I*: General Theory*, Wiley-Interscience, New York, 1988.

[Erm66] Yu. M. Ermoliev, *Methods of solution of nonlinear extremal problems*, Kibernetika, no. 4 (1966), pp. 1–17 (in Russian); Cybernetics, 2 (1966), pp. 1–16 (in English).

[Erm76] Yu. M. Ermoliev, *Stochastic Programming Methods*, Nauka, Moscow, 1976 (in Russian).

[ErS68] Yu. M. Ermoliev and N. Z. Shor, *A random search method for a two-stage stochastic programming problem and its generalization*, Kibernetika, no. 1 (1968), pp. 90–92 (in Russian).

[Gai94] A. A. Gaivoronski, *Convergence properties of backpropagation for neural nets via theory of stochastic gradient methods. Part* 1, Optim. Methods Softw., 4 (1994), pp. 117–134.

[Gla65] E. G. Gladishev, *On stochastic approximation*, Theory Probab. Appl., 10 (1965), pp. 297–300 (in Russian).

[GoK99] J.-L. Goffin and K. C. Kiwiel, *Convergence of a simple subgradient level method*, Math. Program., 85 (1999), pp. 207–211.

[Gri94] L. Grippo, *A class of unconstrained minimization methods for neural network training*, Optim. Methods Softw., 4 (1994), pp. 135–150.

[HUL93] J.-B. Hiriart-Urruty and C. Lemaréchal, *Convex Analysis and Minimization Algorithms*, Springer-Verlag, Berlin, 1993.

[KiA91] S. Kim and H. Ahn, *Convergence of a generalized subgradient method for nondifferentiable optimization*, Math. Programming, 50 (1991), pp. 75–80.

[Kib79] V. M. Kibardin, *Decomposition into functions in the minimization problem*, Avtomat. i Telemekh., no. 9 (1979), pp. 66–79 (in Russian); Automat. Remote Control, 40 (1980), pp. 1311–1323 (in English).

[KiU93] S. Kim and B.-S. Um, *An improved subgradient method for constrained nondifferentiable optimization*, Oper. Res. Lett., 14 (1993), pp. 61–64.

[Kiw96a] K. C. Kiwiel, *The efficiency of subgradient projection methods for convex optimization, Part* II*: Implementations and extensions*, SIAM J. Control Optim., 34 (1996), pp. 677–697.

[Kiw96b] K. C. Kiwiel, *A Subgradient Method with Bregman Projections for Convex Constrained Nondifferentiable Minimization*, Tech. report, Systems Research Institute, Warsaw, Poland, 1996.

[Kiw98] K. C. Kiwiel, *Subgradient method with entropic projections for convex nondifferentiable minimization*, J. Optim. Theory Appl., 96 (1998), pp. 159–173.

[KLL99a] K. C. Kiwiel, T. Larsson, and P. O. Lindberg, *Dual Properties of Ballstep Subgradient Methods, with Applications to Lagrangian Relaxation*, Tech. report LiTH-MAT-R-1999-24, Department of Mathematics, Linköping University, Linköping, Sweden, 1999. Revised 2002.

[KLL99b] K. C. Kiwiel, T. Larsson, and P. O. Lindberg, *The efficiency of ballstep subgradient level methods for convex optimization*, Math. Oper. Res., 24 (1999), pp. 237–254.

[Lis86] S. A. Lisina, *A subgradient method for minimizing a convex function for the case when the infimum is not achieved*, Vestnik Leningrad. Univ. Mat. Mekh. Astronom., no. 4 (1986), pp. 70–74 (in Russian).

[Lit68] B. M. Litvakov, *Convergence of recurrent algorithms for pattern recognition learning*, Avtomat. i Telemekh., no. 3 (1968), pp. 142–150 (in Russian); Automat. Remote Control, 29 (1968), pp. 121–128 (in English).

[LPS96] T. Larsson, M. Patriksson, and A.-B. Strömberg, *Conditional subgradient optimization - theory and applications*, European J. Oper. Res., 88 (1996), pp. 382–403.

[LPS00] T. Larsson, M. Patriksson, and A.-B. Strömberg, *On the Convergence of Conditional ϵ-Subgradient Methods for Convex Programs and Convex-Concave Saddle-Point Problems*, Tech. report, Department of Mathematics, Chalmers University of Technology, Linköping, Sweden, 2000.

[Luo91] Z. Q. Luo, *On the convergence of the LMS algorithm with adaptive learning rate for linear feedforward networks*, Neural Computation, 3 (1991), pp. 226–245.

[LuT94] Z.-Q. Luo and P. Tseng, *Analysis of an approximate gradient projection method with applications to the backpropagation algorithm*, Optim. Methods Softw., 4 (1994), pp. 85–101.

[MaS94] O. L. Mangasarian and M. V. Solodov, *Serial and parallel backpropagation convergence via nonmonotone perturbed minimization*, Optim. Methods Softw., 4 (1994), pp. 103–116.

[MGN87] V. S. Mikhalevich, A. M. Gupal, and V. I. Norkin, *Methods of Nonconvex Optimization*, Nauka, Moscow, 1987 (in Russian).

[Min86] M. Minoux, *Mathematical Programming, Theory and Algorithms*, John Wiley and Sons, Chichester, UK, 1986.

[MiU82] F. Mirzoakhmedov and S. P. Uryasev, *Adaptive stepsize control for the stochastic approximation algorithm*, Zh. Vychisl. Mat. i Mat. Fiz., 23 (1982), pp. 1314–1325 (in Russian).

[NeB01] A. Nedić and D. P. Bertsekas, *Incremental subgradient methods for nondifferentiable optimization*, SIAM J. Optim., 12 (2001), pp. 109–138.

[Nes84] Yu. E. Nesterov, *Minimization methods for nonsmooth convex and quasiconvex functions*, Èkonom. i Mat. Metody, 20 (1984), pp. 519–531 (in Russian); Matekon, 29 (1984), pp. 519–531 (in English).

[Nes89] Yu. E. Nesterov, *Effective Methods in Nonlinear Programming*, Radio i Sviaz, Moscow, 1989 (in Russian).

[NeY78] A. S. Nemirovskii and D. B. Yudin, *Cesaro convergence of the gradient method for approximating saddle points of convex-concave functions*, Dokl. Akad. Nauk SSSR, 239 (1978), pp. 1056–1059 (in Russian).

[Nur79] E. A. Nurminskii, *Numerical Methods for Solving Deterministic and Stochastic Minimax Problems*, Naukova Dumka, Kiev, 1979 (in Russian).

[Nur82] E. A. Nurminski, *Subgradient method for minimizing weakly convex functions and ϵ-subgradient methods of convex optimization*, in Progress in Nondifferentiable Optimization, E. A. Nurminski, ed., CP–82–S8, International Institute for Applied Systems Analysis, Laxenburg, Austria, 1982, pp. 97–123.

[Nur91] E. A. Nurminskii, *Numerical Methods for Convex Optimization*, Nauka, Moscow, 1991 (in Russian).

[NuZ77] E. A. Nurminskii and A. A. Zhelikhovskii, *ϵ-Quasigradient method for solving non-*

smooth extremal problems, Kibernetika, no. 1 (1977), pp. 109–113 (in Russian); Cybernetics, 13 (1977), pp. 109–114 (in English).

[Pol67] B. T. Polyak, *A general method for solving extremum problems*, Dokl. Akad. Nauk SSSR, 174 (1967), pp. 33–36 (in Russian); Soviet Math. Dokl., 8 (1967), pp. 593–597 (in English).

[Pol69] B. T. Polyak, *Minimization of unsmooth functionals*, Zh. Vychisl. Mat. i Mat. Fiz., 9 (1969), pp. 509–521 (in Russian); U.S.S.R. Comput. Math. and Math. Phys., 9 (1969), pp. 14–29 (in English).

[Pol78] B. T. Polyak, *Subgradient methods: A survey of Soviet research*, in Nonsmooth Optimization, C. Lemaréchal and R. Mifflin, eds., Pergamon Press, Oxford, UK, 1978, pp. 5–29.

[Pol83] B. T. Polyak, *Introduction to Optimization*, Nauka, Moscow, 1983 (in Russian); Optimization Software Inc., New York, 1987 (in English).

[PoT73] B. T. Polyak and Ya. Z. Tsypkin, *Pseudogradient adaptation and training algorithms*, Avtomat. i Telemekh., no. 3 (1973), pp. 45–68 (in Russian); Automat. Remote Control, 34 (1973), pp. 377–397 (in English).

[Roc70] R. T. Rockafellar, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.

[Sch83] K. Schulz, *A note on the convergence of subgradient optimization methods*, Math. Operationsforsch. Statist. Ser. Optim., 14 (1983), pp. 537–541.

[SCT00] H. D. Sherali, G. Choi, and C. H. Tuncbilek, *A variable target value method for nondifferentiable optimization*, Oper. Res. Lett., 26 (2000), pp. 1–8.

[Sho62] N. Z. Shor, *An application of the generalized gradient descent method to the solution of a network transportation problem*, in Proceedings of the Scientific Seminar on Theoretical and Application Problems of Cybernetics and Operations Research, no. 1, Scientific Council on Cybernetics of the Academy of Sciences of the Ukrainian SSR, Kiev, 1962, pp. 9–17 (in Russian).

[Sho79] N. Z. Shor, *Minimization Methods for Non-Differentiable Functions*, Naukova Dumka, Kiev, 1979 (in Russian); Springer-Verlag, Berlin, 1985 (in English).

[ShW96] A. Shapiro and Y. Wardi, *Convergence analysis of gradient descent stochastic algorithms*, J. Optim. Theory Appl., 91 (1996), pp. 439–454.

SIAM J. OPTIM.
Vol. 14, No. 3, pp. 841–868

SIMULTANEOUS DATA PERTURBATIONS AND ANALYTIC CENTER CONVERGENCE*

A. HOLDER†

Abstract. The central path is an infinitely smooth parameterization of the nonnegative real line, and its convergence properties have been investigated since the mid 1980s. However, the central "path" followed by an infeasible-interior-point method relies on three parameters instead of one, and hence is a surface instead of a path. The additional parameters are included to allow for simultaneous perturbations in the cost vectors and right-hand side vectors. This paper provides a detailed analysis of the perturbed central path that is followed by infeasible-interior-point methods, and we characterize when such a path converges. We develop a set (Hausdorff) convergence property and show that the central paths impose an equivalence relation on the set of admissible cost vectors. We conclude with a technique to test for convergence under arbitrary, simultaneous data perturbations.

Key words. interior point methods, sensitivity analysis, central path, linear programming

AMS subject classification. 90

DOI. 10.1137/S1052623402409319

1. Introduction.

Interior point algorithms have "revolutionized" the field of mathematical programming [25], and a class of these algorithms, known as path-following interior-point algorithms, follows the *central path* toward the optimal set. The central path has been studied extensively; thus, instead of citing the numerous articles on the subject, we direct interested readers to the three texts of Roos, Terlaky, and Vial [18], Wright [26], and Ye [27], each of which contains an extensive bibliography and a complete development of the central path.

With the amount of literature available on the central path, one may perceive that there is little left to understand. However, this is not the case, especially in semidefinite optimization, where the general convergence of the central path has only recently been established [9]. One of the main goals of this paper is to characterize the convergence of a central "path" that depends on multiple parameters. Several researchers have investigated such convergence [1, 10, 14, 15] (also see [16]), but none of their works completely characterized the convergence of the perturbed central path followed by many interior-point algorithms. We approach the problem as a sensitivity analysis question, and our analysis provides both a characterization of convergence, which subsequently provides insight into algorithm design, and information about the stability of solutions. Another strength of our analysis is that it is relatively simple, requiring only an understanding of real analysis and linear programming (its weakness is that the notation is a bit cumbersome). In related work, Yildirim and Todd propose an interior-point approach to sensitivity analysis in linear and semidefinite optimization [29]. They extend their approach to degenerate linear programs in [30]. The asymptotic analysis in the semidefinite case is the topic of [28].

Consider the primal and dual linear programs

$$(1.1)\quad (LP)\ \max\{cx : Ax = b,\ x \geq 0\} \quad \text{and} \quad (LD)\ \min\{yb : yA + s = c,\ s \geq 0\},$$

*Received by the editors June 6, 2002; accepted for publication (in revised form) August 28, 2003; published electronically March 23, 2004.
http://www.siam.org/journals/siopt/14-3/40931.html
†Department of Mathematics, Trinity University, San Antonio, TX 78212 (aholder@trinity.edu).
This research was conducted at Trinity University and the University of Mississippi and was partially supported through ONR grant N00014-01-1-0917.

where $A \in \mathbb{R}^{m\times n}$ has full row rank, $b \in \mathbb{R}^m$, $c \in \mathbb{R}^n$, and y, s, and c are row vectors. The primal and dual feasible regions are denoted by $\mathcal{P}$ and $\mathcal{D}$, respectively, and their strict interiors are $\mathcal{P}^o = \{x \in \mathcal{P} : x > 0\}$ and $\mathcal{D}^o = \{(y, s) \in \mathcal{D} : s > 0\}$. The primal and dual optimal sets are $\mathcal{P}^*$ and $\mathcal{D}^*$. We assume throughout that Slater's interiority condition holds, i.e., $\mathcal{P}^o \neq \emptyset$ and $\mathcal{D}^o \neq \emptyset$. The necessary and sufficient conditions for optimality are

$$Ax = b, \quad x \geq 0, \quad yA + s = c, \quad s \geq 0, \quad x_i s_i = 0, \quad i = 1, 2, \ldots, n.$$

The central path is formed by replacing the complementarity constraint, $x_i s_i = 0$, with $x_i s_i = \mu > 0$. The fact that A has full row rank implies that for each positive μ there is a unique solution, denoted $(x(\mu), y(\mu), s(\mu))$, to the system

$$Ax = b, \quad x \geq 0, \quad yA + s = c, \quad s \geq 0, \quad x_i s_i = \mu, \quad i = 1, 2, \ldots, n. \tag{1.2}$$

An important observation is that the equations in (1.2) are the necessary and sufficient Lagrange conditions for the penalized linear programs

$$\min \left\{ cx - \mu \sum_{i=1}^{n} \ln(x_i) : x \in \mathcal{P}^o \right\} \qquad \text{and} \qquad \max \left\{ yb + \mu \sum_{i=1}^{n} \ln(s_i) : (y, s) \in \mathcal{D}^o \right\}. \tag{1.3}$$

The logarithmic barrier function in these programs is unique in that it is the only barrier function that yields the Lagrange conditions in (1.2) [13]. The logarithmic barrier function is also used to define the *analytic center* of a bounded polyhedron in the following way. Let $\mathcal{S} = \{x : Ax = b, x \geq 0\}$ be a bounded polyhedron, and let I index the components of x that are positive for some feasible element, i.e., $I = \{i : x_i > 0 \text{ for some } x \in \mathcal{S}\}$. The analytic center of $\mathcal{S}$ is the unique optimizer of

$$\max \left\{ \sum_{i \in I} \ln(x_i) : x \in \mathcal{S},\ x_i > 0,\ i \in I \right\}.$$

The analytic centers of $\mathcal{P}$ and $\mathcal{D}$ are denoted by $\bar{x}$ and $(\bar{y}, \bar{s})$, provided that either $\mathcal{P}$ or $\mathcal{D}$ is bounded. Frisch [4] and Huard [11] were the first to develop algorithms using analytic centers, and Sonnevend reintroduced this concept to the mathematical programming community in [19, 20, 21, 22, 23, 24].

A result first proved by McLinden [12] is that the central path converges to an optimal analytic center as $\mu \downarrow 0$. (Note that we distinguish between a $\downarrow$ and a $\rightarrow$, the former indicating that the limit is approached from above.) To make this precise, we first define the *optimal partition*, denoted by $(B|N)$, as

$$B = \{i : x_i > 0 \text{ for some } x \in \mathcal{P}^*\} \qquad \text{and} \qquad N = \{1, 2, 3, \ldots, n\} \backslash B.$$

Allowing a set subscript on a vector (or matrix) to be the subvector (or submatrix) comprised of the coordinates (or columns) corresponding to the elements in the set, we have that the optimal partition characterizes the optimal sets

$$\mathcal{P}^* = \{x \in \mathcal{P} : x_N = 0\} = \{x : A_B x_B = b,\ x_B \geq 0,\ x_N = 0\} \quad \text{and}$$
$$\mathcal{D}^* = \{(y, s) \in \mathcal{D} : s_B = 0\} = \{(y, s) : yA_B = c_B,\ yA_N + s_N = c_N,\ s_N \geq 0,\ s_B = 0\}.$$

It is well known that the nonemptiness of the strict interiors of the primal and dual feasible regions is equivalent to the boundedness of both $\mathcal{P}^*$ and $\mathcal{D}^*$ [18]. The *central*

solution, written (x^*, y^*, s^*), is the analytic center of $\mathcal{P}^*$ and $\mathcal{D}^*$, which means that x^* and (y^*, s^*) are the unique solutions to

$$\max\left\{\sum_{i\in B}\ln(x_i) : x \in \mathcal{P}^*,\ x_B > 0\right\},$$
$$\max\left\{\sum_{i\in N}\ln(s_i) : (y,s) \in \mathcal{D}^*,\ s_N > 0\right\}.$$

McLinden showed in 1980 that the central path converges to (x^*, y^*, s^*) as $\mu \downarrow 0$, a result that is stated in Theorem 1.1.

THEOREM 1.1 (see McLinden [12]). *We have that*

$$\lim_{\mu\downarrow 0}(x(\mu), y(\mu), s(\mu)) = (x^*, y^*, s^*).$$

Furthermore, if $\mathcal{P}$ is bounded, $\lim_{\mu\to\infty} x(\mu) = \bar{x}$, *and if $\mathcal{D}$ is bounded,* $\lim_{\mu\to\infty}(y(\mu), s(\mu)) = (\bar{y}, \bar{s})$.

Originally, interior-point algorithms assumed the existence of a strictly feasible primal and dual pair. However, subsequent interior-point algorithms allowed infeasible starting points, with the idea of starting with any (x^0, y^0, s^0) such that both x^0 and s^0 are positive and define the following primal and dual residuals:

$$r_b = Ax^0 - b \qquad \text{and} \qquad r_c = y^0A + s^0 - c. \tag{1.4}$$

These residuals are scaled and added to b and c in (1.2) to obtain

$$Ax = b + \rho r_b, \quad x \geq 0, \quad yA + s = c + \tau r_c, \quad s \geq 0, \quad x_i s_i = \mu, \quad i = 1, 2, \ldots, n. \tag{1.5}$$

For $\rho = \tau = 1$, (x^0, y^0, s^0) is strictly feasible. The problem is that, unless the residuals are zero, the right-hand side vector and cost vector are different from those of the original problem. So, infeasible-interior-point algorithms start with the perturbed data $b + \rho r_b$ and $c + \tau r_c$ and then decrease ρ and τ to zero while decreasing μ to zero. However, this means that the central path no longer relies on the single parameter μ but on the three parameters μ, ρ, and τ. Unfortunately, convergence is not guaranteed as μ, ρ, and τ decrease to zero, as shown in [10].

Explaining the convergence behavior of $(x(\mu), y(\mu), s(\mu))$ under data perturbations falls under the auspices of sensitivity analysis, and this is precisely the perspective from which we approach the problem. Because we are interested in how the central path relies on b and c, we extend our notation so that $(x(\mu, b, c), y(\mu, b, c), s(\mu, b, c))$ is the unique solution to the equations in (1.2). We point out that, because $x(\mu, b, c)$ is the optimizer of the first math program in (1.3), we have for any positive α that $x(\mu, b, c) = x(\alpha\mu, b, \alpha c)$ (simply multiply the objective function by α). Similarly, $(y(\mu, b, c), s(\mu, b, c)) = (y(\alpha\mu, \alpha b, c), s(\alpha\mu, \alpha b, c))$ for $\alpha > 0$. For the data b and c, the central path, primal central path, and dual central path are, respectively,

$$\begin{aligned} CP_{(b,c)} &\equiv \{(x(\mu, b, c), y(\mu, b, c), s(\mu, b, c)) : \mu > 0\}, \\ PCP_{(b,c)} &\equiv \{x(\mu, b, c) : \mu > 0\}, \\ DCP_{(b,c)} &\equiv \{(y(\mu, b, c), s(\mu, b, c)) : \mu > 0\}. \end{aligned}$$

In general, we consider sequences b^k and c^k, the use of which allows for arbitrary, simultaneous, and independent perturbations in b and c. Obviously, these data perturbations encompass the linear changes found in (1.5). Because x, y, and s no longer depend on a single parameter, we are, technically, dealing with a surface and not a path. However, for intuitive and geometric reasons, we refer to a *perturbed central path* and choose sequences $x(\mu^k, b^k, c^k)$ from $PCP_{(b^k,c^k)}$.

As we shall see, allowing nonlinear perturbations in the cost coefficients significantly increases the difficulty of characterizing the convergence of the perturbed central path, and we often end up dealing with linear changes. When this is the case, we let $b^k = b(\rho^k) = b + \rho^k \delta b$ and $c^k = c(\tau^k) = c + \tau^k \delta c$, where the direction vectors δb and δc are understood. Other notational extensions are described in Table 1.1.

TABLE 1.1
Notation accounting for the dependence on b and c.

Notation	Explanation	Notation	Explanation
$\mathcal{P}_b$	Primal feasible region	$\mathcal{D}_c$	Dual feasible region
$\mathcal{P}_b^o$	Strict interior of $\mathcal{P}_b$	$\mathcal{D}_c^o$	Strict interior of $\mathcal{D}_c$
$\mathcal{P}^*_{(b,c)}$	Primal optimal set	$\mathcal{D}^*_{(b,c)}$	Dual optimal set
$(\mathcal{P}^*_{(b,c)})^o$	Strict interior of $\mathcal{P}^*_{(b,c)}$	$(\mathcal{D}^*_{(b,c)})^o$	Strict interior of $\mathcal{D}^*_{(b,c)}$
$\bar{x}(b)$	Analytic center of $\mathcal{P}_b$	$(\bar{y}(c), \bar{s}(c))$	Analytic center of $\mathcal{D}_c$
$x^*(b,c)$	Analytic center of $\mathcal{P}^*_{(b,c)}$	$(y^*(b,c), s^*(b,c))$	Analytic center of $\mathcal{D}^*_{(b,c)}$
$(B(b,c)\|N(b,c))$	Optimal partition		

All scalar sequences are in $\mathbb{R}^*_+ = \{\nu \in \mathbb{R} : \nu \geq 0\} \cup \{\infty\}$, which means that every scalar sequence has a cluster point (one of which may be ∞). The row, column, and null spaces of a matrix are denoted by $\text{row}(A)$, $\text{col}(A)$, and $\text{null}(A)$, respectively, and the projection of v onto the vector space W is denoted by $\text{proj}_W v$. The capitalization of a vector indicates the diagonal matrix formed from the vector. So, X is a diagonal matrix whose diagonal components are $x_1, x_2, \ldots, x_n$. The vector e is the all ones vector, where length is decided by the context of its use. The standard Big-O, o, Ω, and Θ notation is used [17]. Other notation is consistent with the *Mathematical Programming Glossary* [5].

We accomplish three primary goals in this paper. First, we characterize the convergence of $x(\mu^k, b^k, c+\tau^k \delta c)$ as $\mu^k \downarrow 0$, $b^k \to b$, and $\tau^k \downarrow 0$ by providing necessary and sufficient conditions on (μ^k, b^k, τ^k). Notice that nonlinear perturbations in b are allowed (but only linear changes in c). This result completely describes the convergence of the perturbed central path followed by all infeasible-path-following-interior-point algorithms. Second, we provide a set (Hausdorff) convergence result for the perturbed Central path. This result shows that while the sequence $x(\mu^k, b^k, c + \tau^k \delta c)$ may not converge, the sequence of perturbed central paths does converge. Third, we remove the restriction that the perturbation in c must be linear, and we develop a process to calculate the limit of $x(\mu^k, b^k, c^k)$.

Before we begin, we point out that partial solutions are found in the literature. In [14], Mizuno, Todd, and Ye provide necessary conditions for the cluster points of the perturbed central path to be contained in the interior of the optimal set and the boundary of the optimal set. Bonnans and Potra [1] consider the case of a single shifted center within a specific algorithm environment for the horizontal

linear complementarity problem. However, these results do not permit independent changes in b and c because the single parameter that is used controls the perturbations in b and c. Monteiro and Tsuchiya [15] show that $x^*(\mu^k, b, c + \mu^k \&c)$ converges as $\mu^k \downarrow 0$ but, as in [1], this analysis relies on a single parameter. Holder, Sturm, and Zhang [10] show that for any positive η, $x(\eta\mu^k, b + \rho^k \&b, c + \mu^k \&c)$ converges as $(\mu^k, \rho^k) \downarrow 0$ and $x(\mu^k, b + \rho^k \&b, c)$ converges as $(\mu^k, \rho^k) \downarrow 0$. Moreover, they prove that if $\tau^k = o(\mu^k)$, then $x(\mu^k, b + \rho^k \&b, c + \tau^k \&c)$ converges as $(\mu^k, \rho^k, \tau^k) \downarrow 0$. The results in [10] and [15] provide the actual limit when convergence is guaranteed. As one can see, there are many *sufficient* conditions that guarantee the convergence of $x(\mu^k, b+\rho^k \&b, c+\tau^k \&c)$. Our goal is different in that we want to characterize the convergence of $x(\mu^k, b^k, c+\tau^k \&c)$ by providing *necessary* and *sufficient* conditions. A strength of our analysis is that we explain the entire set of cluster points of $x(\mu^k, b^k, c + \tau^k \&c)$.

2. Preliminary results. This section contains foundational material for subsequent sections, and several of the results in this section are simple to prove. While many of these results are used in the literature, some proofs are not readily available, and we include such proofs for completeness. If a result is proven elsewhere, we simply cite that reference. Readers familiar with the central path literature will feel comfortable browsing through the notation and results of this section.

We begin with a study of the data that we are allowed to operate over. We say that b and c are *admissible* if the strict interiors of the primal and dual are nonempty. The admissible data sets are denoted by

$$\begin{aligned}
\mathcal{G} &\equiv \{(b, c) \in \mathbb{R}^m \times \mathbb{R}^n : \mathcal{P}_b^o \neq \emptyset,\ \mathcal{D}_c^o \neq \emptyset\},\\
\mathcal{G}^1 &\equiv \{b \in \mathbb{R}^m : \mathcal{P}_b^o \neq \emptyset\},\\
\mathcal{G}^2 &\equiv \{c \in \mathbb{R}^n : \mathcal{D}_c^o \neq \emptyset\}.
\end{aligned}$$

Our definition of admissible does not correspond with the traditional definition of admissible, which states that (*LP*) and (*LD*) have finite optimal solutions. Our definition is more restrictive because only data for which $\mathcal{P}_b^o$ and $\mathcal{D}_c^o$ are not empty are included. The first result shows that $\mathcal{G}$ is open, which subsequently implies that arbitrarily small perturbations of b and c remain admissible.

THEOREM 2.1. *$\mathcal{G}$ is an open set.*

Proof. Let $(\hat{b}, \hat{c}) \in \mathcal{G}$. Then, there exists $\hat{x}$ and $(\hat{y}, \hat{s})$ such that $A\hat{x} = \hat{b}$, $\hat{x} > 0$, $\hat{y}A + \hat{s} = \hat{c}$, and $\hat{s} > 0$. Let U be an open set in $\mathbb{R}^n$ that contains $\hat{x}$ and has the property that $x \in U$ implies $x > 0$. Since the rank of A is m, the linear transformation $T : \mathbb{R}^n \to \mathbb{R}^m : x \to Ax$ is onto. Furthermore, since T is a continuous mapping, the open mapping theorem implies that $T(U)$ is open. Let $\epsilon = \min\{\hat{s}_i : i = 1, 2, \ldots, m\}$, and define $V = \{c : \|c - \hat{c}\| < \epsilon\}$. Then $(\hat{b}, \hat{c}) \in T(U) \times V \subset \mathcal{G}$, and the result follows since $T(U) \times V$ is open. □

If $x(\mu^k, b^k, c^k) \to \hat{x}$, we have that $b^k = Ax(\mu^k, b^k, c^k) \to A\hat{x}$, which means that the convergence of b^k is a necessary condition of the convergence of $x(\mu^k, b^k, c^k)$. As such, we make the following assumption throughout.

Assumption 1. We assume throughout that (b, c) and (b^k, c^k) are in $\mathcal{G}$. Moreover, we assume that $b^k \to b$ (but we do *not* necessarily assume that $c^k \to c$).

Also, for notational convenience we assume that $(B|N)$ is the optimal partition for (b, c), i.e., $(B|N) = (B(b, c)|N(b, c))$. The dependence that the optimal partition has on b and c is indicated only for the perturbed data b^k and c^k. Sonnevend [19] showed that $x(\mu, b, c)$ is an analytic function over $\mathbb{R}_{++} \times \mathcal{G}$ (where we abuse the notation so

that the 2-tuple $(\mu,(b,c))$ is understood to be the 3-tuple (μ,b,c)). Hence,

$$\mu^0 > 0 \Rightarrow \lim_{(\mu^k,b^k,c^k)\to(\mu^0,b,c)} x(\mu^k,b^k,c^k) = x(\mu^0,b,c). \tag{2.1}$$

The next two results show that either the primal objective function is strictly decreasing along the central path or that the central path degenerates to a single element.

THEOREM 2.2 (see Fiacco and McCormick [3]). *For $0<\mu^1<\mu^2$, we have that $c\notin \mathrm{row}(A)$ if and only if*

$$cx^*(b,c) < cx(\mu^1,b,c) < cx(\mu^2,b,c) < c\bar{x}(b).$$

Similarly, for $0<\mu^1<\mu^2$, we have that $b\neq 0$ if and only if

$$y^*(b,c)b > y(\mu^1,b,c)b > y(\mu^2,b,c)b > \bar{y}(c)b.$$

THEOREM 2.3 (see Roos, Terlaky, and Vial [18]). *The following are equivalent:*

1. *cx is constant on $\mathcal{P}_b$.*
2. *$x(\mu^1,b,c)=x(\mu^2,b,c)$ for all $0<\mu^1<\mu^2$.*
3. *$x(\mu^1,b,c)=x(\mu^2,b,c)$ for some $0<\mu^1<\mu^2$.*
4. *$c\in\mathrm{row}(A)$.*
5. *$s(\mu,b,c)=\mu s(1,b,c)$ for all $0<\mu$.*

An observation that we use later is that if $c\in\mathrm{row}(A)$ and $(b,c)\in\mathcal{G}$, then $\mathcal{P}_b$ is bounded. This follows because $\mathcal{P}_b$ is bounded if and only if there does not exist $d\!x$ such that $Ad\!x=0$, $d\!x\geq 0$, and $d\!x\neq 0$. From Gordon's theorem of the alternative (a variant of Farkas's lemma) this is the same as $\mathcal{P}_b$ being bounded if and only if there is a row vector y such that $yA>0$. Suppose that $c\in\mathrm{row}(A)$, so that $\hat{y}A=c$ for some $\hat{y}$. Then, for any positive μ, we have that $0<s(\mu,b,c)=c-y(\mu,b,c)A=(\hat{y}-y(\mu,b,c))A$, and hence $\mathcal{P}_b$ is bounded.

We now direct our attention toward linear perturbations. Recall that, for the understood directions of change $\delta\!b$ and $\delta\!c$, we defined $b(\rho)$ as $b+\rho\delta\!b$ and $c(\tau)$ as $c+\tau\delta\!c$. Directions of change for which the optimal partition is invariant for sufficiently small ρ and τ are of particular interest, and we define

$$\begin{aligned}
\mathcal{H}(b,c) = \{&(\delta\!b,\delta\!c) : \text{there exists } \tilde{\rho}>0 \text{ and } \tilde{\tau}>0 \text{ such that for all } 0\leq(\rho,\tau)<(\tilde{\rho},\tilde{\tau}),\\
&(B(b(\rho),c(\tau))|N(b(\rho),c(\tau))) = (B(b,c)|N(b,c))\},\\
\mathcal{H}^1(b,c) = \{&\delta\!b : (\delta\!b,0)\in\mathcal{H}(b,c)\},\\
\mathcal{H}^2(b,c) = \{&\delta\!c : (0,\delta\!c)\in\mathcal{H}(b,c)\}.
\end{aligned}$$

Properties of these sets are found in [6] and [7]. The next lemma shows that the optimal partition characterizes $\mathcal{H}(b,c)$, $\mathcal{H}^1(b,c)$, and $\mathcal{H}^2(b,c)$.

LEMMA 2.4. *We have that $\mathcal{H}^1(b,c)=\mathrm{col}(A_B)$ and that $\mathcal{H}^2(b,c)=\{\delta\!c\in\mathbb{R}^n : \delta\!c_B\in\mathrm{row}(A_B)\}$.*

Proof. The partition $(B|N)$ is optimal for the right-hand side $b(\rho)$ if and only if the following system is consistent:

$$A_Bx_B=b(\rho),\quad x_B>0,\quad yA_B=c_B,\quad \text{and}\quad yA_N<c_N.$$

If $\delta\!b\in\mathrm{col}(A_B)$, there exists x' such that $A_B(\rho x')=\rho\delta\!b$. Since $x^*_B(b,c)-\rho x'$ is positive for sufficiently small ρ, the above conditions remain consistent for arbitrarily small ρ.

Hence, $\mathrm{col}(A_B) \subseteq \mathcal{H}^1(b,c)$. If the optimal partition is invariant for sufficiently small ρ, then there exists $x_B(\rho)$ such that $A_B x_B(\rho) = b(\rho)$. Since $A_B(x_B(\rho) - x_B^*(b,c)) = \rho\delta b$, we have that δb is in $\mathrm{col}(A_B)$.

The argument for $\mathcal{H}^2(b,c)$ is similar, with the difference being that the optimality conditions are

$$A_B x_B = b, \quad x_B > 0, \quad yA_B = c_B(\tau), \quad yA_N < c_N(\tau). \qquad \square$$

The remainder of this section is concerned with establishing the existence of limits. Lemmas 2.5 and 2.7 provide bounds so that sequences have cluster points, and Lemma 2.6 and Theorem 2.9 use these bounds to establish limits. Consider the level set

$$\mathcal{L}(b,c,M) = \{(x,y,s) \in \mathcal{P}_b \times \mathcal{D}_c : sx \le M\}.$$

The next lemma shows that the union over k of the level sets $\mathcal{L}(b^k,c^k,M)$ is bounded, provided that c^k is bounded. The level set argument is similar to Theorem I.4 in [18] and Lemma 4.2 in [10], with the differences being that c^k need not converge and that independent, arbitrary perturbations in b and c are allowed. (Theorem I.4 does not permit data perturbations, and Lemma 4.2 allows only linear changes in b and c that converge.)

LEMMA 2.5. *If c^k is bounded, then for $M \ge 0$ we have that $\bigcup_k \mathcal{L}(b^k,c^k,M)$ is bounded.*

Proof. Let $M \ge 0$ and $\mu^0 > 0$. Also, let $x^k = x(\mu^0,b^k,c^k)$ and $s^k = s(\mu^0,b^k,c^k)$. Then, for any $x \in \mathcal{P}_{b^k}$ and $(y,s) \in \mathcal{D}_{c^k}$, we have that $x^k - x \in \mathrm{null}(A)$, $s^k - s \in \mathrm{row}(A)$, and

$$0 = (s^k - s)(x^k - x) = s^k x^k - sx^k - s^k x + sx. \tag{2.2}$$

So, for any $(x,y,s) \in \mathcal{L}(b^k,c^k,M)$, we have that

$$s_i^k x_i \le s^k x + sx^k = s^k x^k + sx \le s^k x^k + M.$$

Since $s^k > 0$ and $s^k x^k = \mu^0 n$, we have that $x_i \le (M + \mu^0 n)/s_i^k$. A similar argument shows that $s_i \le (M + \mu^0 n)/x_i^k$. Since y relates to s in a one-to-one, linear fashion, we have for each k that $\mathcal{L}(b^k,c^k,M)$ is bounded.

To establish that $\cup_k \mathcal{L}(b^k,c^k,M)$ is bounded, we first show that $x(\mu^0,b^k,c^k)$ and $s(\mu^0,b^k,c^k)$ are $\Omega(1)$. Suppose, for the sake of contradiction, that there is a subsequence (μ^0,b^{k_j},c^{k_j}) such that $x_i(\mu^0,b^{k_j},c^{k_j}) \downarrow 0$ for some i. Since c^{k_j} is bounded, it contains a convergent subsequence, and we assume without loss of generality that $c^{k_j} \to c$. However, this provides a contradiction because from (2.1) we have that $x(\mu^0,b^{k_j},c^{k_j}) \to x(\mu^0,b,c) > 0$. Hence, $x(\mu^0,b^k,c^k) = \Omega(1)$. An analogous argument shows that $s(\mu^0,b^k,c^k) = \Omega(1)$. We now have that there are positive λ^1 and λ^2 such that $x_i(\mu^0,b^k,c^k) > \lambda^1$ and $s_i(\mu^0,b^k,c^k) > \lambda^2$. So,

$$x_i \le \frac{M + \mu^0 n}{s_i^k} < \frac{2(M + \mu^0 n)}{\lambda^2} \qquad \text{and} \qquad s_i \le \frac{M + \mu^0 n}{x_i^k} < \frac{2(M + \mu^0 n)}{\lambda^1}.$$

Since these bounds are independent of k, we have that $\bigcup_k \mathcal{L}(b^k,c^k,M)$ is bounded. $\square$

Lemma 2.5 does not require that c^k converge but only that it be bounded. From this result we have that if $\mu^k \downarrow 0$ and c^k is bounded, then the sequence

$$(x(\mu^k,b^k,c^k), y(\mu^k,b^k,c^k), s(\mu^k,b^k,c^k))$$

has a cluster point. However, an example in [10] shows that these sequences need not converge, which means that a straightforward extension of Theorem 1.1 is not available. The next lemma shows that x_N and s_B approach zero with μ.

LEMMA 2.6. *If $\mu^k \downarrow 0$ and $c^k \to c$, we have that $x_N(\mu^k, b^k, c^k) \to 0$ and $s_B(\mu^k, b^k, c^k) \to 0$.*

Proof. Lemma 2.5 implies that $(x(\mu^k, b^k, c^k), y(\mu^k, b^k, c^k), s(\mu^k, b^k, c^k))$ has a convergent subsequence, say,

$$\lim_{i\to\infty} \left(x(\mu^{k_i}, b^{k_i}, c^{k_i}), y(\mu^{k_i}, b^{k_i}, c^{k_i}), s(\mu^{k_i}, b^{k_i}, c^{k_i})\right) = (\hat{x}, \hat{y}, \hat{s})\,.$$

Set $x^i = x(\mu^{k_i}, b^{k_i}, c^{k_i})$, $y^i = y(\mu^{k_i}, b^{k_i}, c^{k_i})$, and $s^i = s(\mu^{k_i}, b^{k_i}, c^{k_i})$. Since

$$\left.\begin{array}{rcll} Ax^i &=& b^{k_i}, & x^i > 0 \\ y^i A + s^i &=& c^{k_i}, & s^i > 0 \\ s^i x^i &=& n\mu^{k_i} & \end{array}\right\} \Rightarrow \left\{\begin{array}{rcll} A\hat{x} &=& b, & \hat{x} \geq 0 \\ \hat{y}A + \hat{s} &=& c, & \hat{s} \geq 0 \\ \hat{s}\hat{x} &=& 0, & \end{array}\right. \tag{2.3}$$

we have that $\hat{x} \in \mathcal{P}^*_{(b,c)} = \{x \in \mathcal{P}_b : x_N = 0\}$ and $(\hat{y}, \hat{s}) \in \mathcal{D}^*_{(b,c)} = \{(y, s) \in \mathcal{D}_{\bar{c}} : s_B = 0\}$, which proves the result. □

If $\mu^k \downarrow 0$, $c^k \to c$, and $x(\mu^k, b^k, c^k) \to \hat{x}$, Lemma 2.6 identifies a subvector of $\hat{x}$ that is zero. Unfortunately, this is not necessarily the largest subvector of $\hat{x}$ that is zero, an issue that we address in section 5.

The final objective of this section is to develop sufficient conditions for $x(\mu^k, b^k, c^k)$ to converge to the analytic center of a polytope, a result that relies on Lemmas 2.7 and 2.8.

LEMMA 2.7 (see Caron, Greenberg, and Holder [2]). *If $\mathcal{P}_b$ is bounded, $\bigcup_k \mathcal{P}_{b^k}$ is bounded.*

From Lemma 2.7 we have that a bounded polytope remains bounded under right-hand side perturbation. We now introduce the concept of *set convergence* [8] (typically called Hausdorff convergence), an idea that we use now to establish the existence of a particular sequence and use later to show that the central path converges as a set. We say that a sequence of sets H^k converges to the set H if the following two conditions hold:

1. If $h^k \in H^k$ and $h^k \to h$, then h must be in H.
2. If $h \in H$, then there exists $h^k \in H^k$ such that $h^k \to h$.

From [8] we know that $b^k \to b$ implies $\mathcal{P}_{b^k} \to \mathcal{P}_b$, which is important because we require that elements within the strict interior of the feasible set may be approached by strictly positive elements. To see that this is true, let $x \in \mathcal{P}^o_b$. Then, since $\mathcal{P}_{b^k} \to \mathcal{P}_b$, there is a sequence $x^k \in \mathcal{P}_{b^k}$ such that $x^k \to x$, and because x is positive, we have that x^k is positive for sufficiently large k. We state this fact in Lemma 2.8.

LEMMA 2.8. *If x is in $\mathcal{P}^o_b$, there exists a sequence $x^k \in \mathcal{P}^o_{b^k}$ such that $x^k \to x$.*

The next theorem provides sufficient conditions for $x(\mu^k, b^k, c^k)$ to converge to the analytic center of a polytope.

THEOREM 2.9. *Let $\mathcal{P}_b$ be bounded. Then, if the vector sequence c^k/μ^k is bounded and has the property that every cluster point is in* $\mathrm{row}(A)$*, we have that $x(\mu^k, b^k, c^k) \to \bar{x}(b)$.*

Proof. From Lemma 2.7 we have that $x(\mu^k, b^k, c^k)$ is bounded. So, there exists a subsequence such that

$$x(\mu^{k_i}, b^{k_i}, c^{k_i}) \to \hat{x} \qquad \text{and} \qquad \frac{c^{k_i}}{\mu^{k_i}} \to \hat{c}.$$

Let $x^i = x(\mu^{k_i}, b^{k_i}, c^{k_i})$, $y^i = y(\mu^{k_i}, b^{k_i}, c^{k_i})$, and $s^i = s(\mu^{k_i}, b^{k_i}, c^{k_i})$. Similar to (2.3), we have that $\hat{x} \in \mathcal{P}_b$. For any i, the necessary and sufficient conditions describing (x^i, y^i, s^i) are

$$Ax = b^{k_i}, \quad x > 0, \quad yA + s = c^{k_i}, \quad s > 0, \quad \text{and} \quad Sx = \mu^{k_i} e,$$

which means that

$$\begin{aligned} Ax^i &= b^{k_i}, \\ -\frac{y^i}{\mu^{k_i}} A &= e^T (X^i)^{-1} - \frac{c^{k_i}}{\mu^{k_i}}, \\ x^i &> 0. \end{aligned} \tag{2.4}$$

From the full row rank of A, we have that

$$-\frac{y^i}{\mu^{k_i}} = \left(e^T (X^i)^{-1} - \frac{c^{k_i}}{\mu^{k_i}} \right) A^T \left(AA^T \right)^{-1}.$$

We prove that $\hat{x}$ is positive so that this last equality implies that the sequence y^i/μ^{k_i} has a limit. Then, since $\hat{c}$ is in row(A), (2.4) implies that $e^T \hat{X}^{-1}$ is in row(A). Subsequently, we have that there is a $\hat{y}$ such that

$$A\hat{x} = b, \quad \hat{y}A = e^T \hat{X}^{-1}, \quad \hat{x} > 0,$$

and because these are the necessary and sufficient conditions describing $\bar{x}(b)$, the result is established once we show that $\hat{x}$ is positive.

From Lemma 2.8 there is a sequence, $\tilde{x}^i \in \mathcal{P}^o_{b^{k_i}}$, such that $\tilde{x}^i \to \tilde{x} \in \mathcal{P}^o_b$. The optimality of x^i implies that

$$\frac{c^{k_i}}{\mu^{k_i}} x^i - \sum_{j=1}^n \ln(x^i_j) \leq \frac{c^{k_i}}{\mu^{k_i}} \tilde{x}^i - \sum_{j=1}^n \ln(\tilde{x}^i_j),$$

which is equivalent to

$$\sum_{j=1}^n \ln(\tilde{x}^i_j) \leq \frac{c^{k_i}}{\mu^{k_i}} (\tilde{x}^i - x^i) + \sum_{j=1}^n \ln(x^i_j). \tag{2.5}$$

Since $\tilde{x}^i$ is $\Omega(1)$, the left-hand side of this last inequality is bounded from below. Suppose, for the sake of contradiction, that as $i \to \infty$, $x^i_j \to 0$ for some j. The boundedness of x^i implies that $\sum_{j=1}^n \ln(x^i_j) \to -\infty$. Hence, the inequality in (2.5) implies that $(c^{k_i}/\mu^{k_i})(\tilde{x}^i - x^i) \to \infty$. However, since $\hat{c} \in$ row(A) and $(\tilde{x} - \hat{x}) \in$ null(A), we have that

$$\frac{c^{k_i}}{\mu^{k_i}} (\tilde{x}^i - x^i) \to \hat{c}(\tilde{x} - \hat{x}) = 0.$$

Thus, no such j exists, and $\hat{x} > 0$. ∎

COROLLARY 2.10. *If $\mathcal{P}_b$ is bounded, $c^k \to c$, and $\mu^k \to \infty$, then $x(\mu^k, b^k, c^k) \to \bar{x}(b)$.*

Proof. The proof follows immediately from Theorem 2.9 because $c^k/\mu^k \to 0 \in$ row(A). ∎

While only providing sufficient conditions for the convergence of $x(\mu^k, b^k, c^k)$, Theorem 2.9 is used in the next section to develop necessary and sufficient conditions. We point out that neither μ^k, c^k, nor c^k/μ^k had to converge for $x(\mu^k, b^k, c^k)$ to converge. Because of this, Theorem 2.9 highlights the difficulty of allowing simultaneous perturbations in μ, b, and c.

3. Characterizing the convergence of the central path under simultaneous parameterization. The goal of this section is to develop necessary and sufficient conditions on $(\mu^k, b^k, c(\tau^k))$ so that $x(\mu^k, b^k, c(\tau^k))$ converges as $\mu^k \downarrow 0$ and $\tau^k \downarrow 0$. We assume throughout this section that $\tau^k \downarrow 0$. These conditions are stated in Theorem 3.8 and they completely characterize the convergence of the perturbed central path followed by an infeasible-path-following-interior-point algorithm. In this section, we allow arbitrary perturbations in b and linear changes in c. The case of independent, arbitrary, nonlinear changes in both b and c is addressed in section 5. Our first goal is to show that the objective function is constant on "cuts" of the feasible region, which are defined for any k and positive μ as

$$\mathcal{C}(\mu, k) = \{x_B : A_B x_B = b^k - A_N x_N(\mu, b^k, c(\tau^k)),\ x_B \geq 0\}.$$

$\mathcal{C}(\mu, k)$ is the subpolyhedron of $\mathcal{P}_{b^k}$ formed by fixing x_N to be $x_N(\mu, b^k, c(\tau^k))$. Lemma 3.1 shows that $c_B x_B$ is constant on each $\mathcal{C}(\mu, k)$.

LEMMA 3.1. *For any k and positive μ, $c_B x_B$ is constant on $\mathcal{C}(\mu, k)$. Consequently, $x_B(\mu, b^k, c(\tau^k))$ is the unique solution to*

$$\min\left\{\tau^k \delta c_B x_B - \mu \sum_{i \in B} \ln(x_i) : A_B x_B = b^k - A_N x_N(\mu, b^k, c(\tau^k)),\ x_B > 0\right\}. \tag{3.1}$$

Proof. By definition, $x(\mu, b^k, c(\tau^k))$ is the unique solution to

$$\min\left\{cx + \tau^k \delta c x - \mu \sum_{i=1}^{n} \ln(x_i) : x \in (\mathcal{P}_{b^k})^o\right\}.$$

Holding the components of $x_N(\mu, b^k, c(\tau^k))$ constant, we have that $x_B(\mu, b^k, c(\tau^k))$ is the unique solution to

$$\min\left\{c_B x_B + \tau^k \delta c_B x_B - \mu \sum_{i \in B} \ln(x_i) : A_B x_B = b^k - A_N x_N(\mu, b^k, c(\tau^k)),\ x_B > 0\right\}.$$

So, the result follows once we show that $c_B x_B$ is constant on $\mathcal{C}(\mu, k)$. If the columns of A_B are linearly independent, the result is immediate because $\mathcal{C}(\mu, k)$ contains a single element. Otherwise, let x_B^1 and x_B^2 be in $\mathcal{C}(\mu, k)$. Then, since $x_B^1 - x_B^2 \in \text{null}(A_B)$ and $c_B \in \text{row}(A_B)$, we have that $c_B x_B^1 = c_B x_B^2$. ∎

The fact that $x_B(\mu, b^k, c(\tau^k))$ is the unique optimal solution to the math program in (3.1) is paramount in our analysis. To aid our development, for any positive η we define $z_B(\eta, b, \delta c_B)$ to be the unique solution to

$$\min\left\{\delta c_B z_B - \eta \sum_{i \in B} \ln(z_i) : A_B z_B = b,\ z_B > 0\right\}, \tag{3.2}$$

which means that $\{z_B(\eta, b, \delta c_B) : \eta > 0\}$ is the central path for the linear program

$$\min\{\delta c_B z_B : A_B z_B = b,\ z_B \geq 0\}. \tag{3.3}$$

Because $\{z_B(\eta, b, \delta c_B) : \eta > 0\}$ is a central path for fixed b and δc_B, $z_B(\eta, b, \delta c_B)$ has a limit as $\eta \downarrow 0$, which is denoted by $z_B^*(b, \delta c_B)$. The feasible region of the math program in (3.3) is equipotent to $\mathcal{P}^*_{(b,c)}$ (just remove x_N). Since (b, c) in $\mathcal{G}$ implies that $\mathcal{P}^*_{(b,c)}$ is

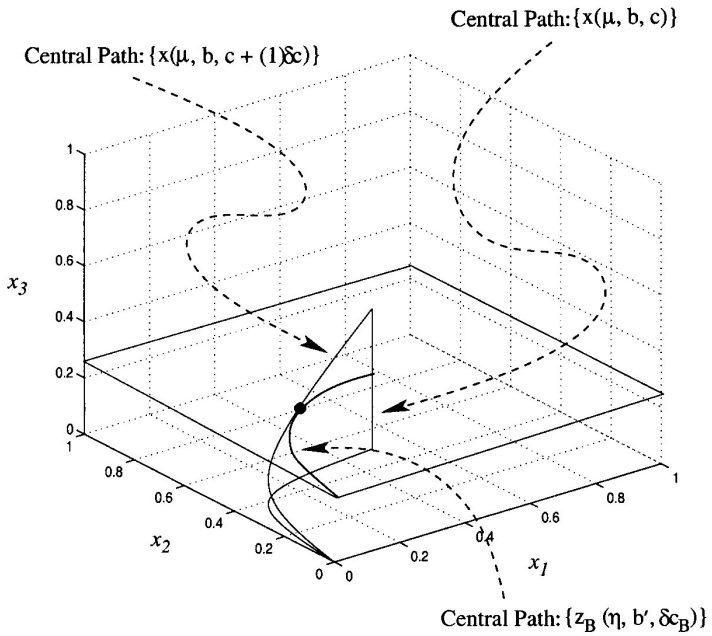


FIG. 3.1. *Four central paths associated with* (LP^1) *and how they intersect (note that* b' *is* $b - A_N x_N(\mu, b, c(1))$*).*

bounded, we have that the feasible region of (3.3) is bounded and, subsequently, that $z_B(\eta, b, \delta c_B)$ converges as $\eta \to \infty$ to the analytic center of $\{z_B : A_B z_B = b, z_B \geq 0\}$. Since $x_B^*(b, c)$ is this analytic center, we have that $\lim_{\eta \to \infty} z_B(\eta, b, \delta c_B) = x_B^*(b, c)$. In addition to the convergence properties of $z_B(\eta, b, \delta c_B)$, we have from Lemma 3.1 that

$$x_B(\mu, b^k, c(\tau^k)) = z_B(\mu/\tau^k, b^k - A_N x_N(\mu, b^k, c(\tau^k)), \delta c_B). \tag{3.4}$$

The following example illustrates the relationship between $x(\mu, b, c(\tau))$ and $z_B(\eta, b, \delta c_B)$.

Example 3.1. Consider the linear program

$$(LP^1) \quad \min\{x_3 : 0 \leq x_1 \leq 1,\ 0 \leq x_2 \leq 1,\ 0 \leq x_3 \leq 1\}.$$

Allowing x_4, x_5, and x_6 to be the slack variables, we have that the optimal partition is $(\{1, 2, 4, 5, 6\}|\{3\})$. Let $b^k = b$, so there is no right-hand side perturbation, and $\delta c = (1, 1/10, 0, 0, 0, 0)$, so $c^k = c + \tau^k \delta c = (\tau^k, \tau^k/10, 1, 0, 0, 0)$. Figure 3.1 illustrates four central paths associated with perturbations of (LP^1). The vertical line is the unperturbed central path for (LP^1), and the curve in the x_1, x_2-plane is the central path for

$$\begin{aligned}(LP^2) \quad & \min\{\delta c_B x_B : x \in \mathcal{P}^*_{(b,c)}\} \\ & = \min\left\{x_1 + \frac{1}{10} x_2 : 0 \leq x_1 \leq 1,\ 0 \leq x_2 \leq 1,\ x_3 = 0\right\}.\end{aligned}$$

The curve from $(1/2, 1/2, 1/2)$ to $(0, 0, 0)$ is the perturbed central path for $\tau^k = 1$,

and hence corresponds to the linear program

$$(LP^3) \quad \min\left\{x_1 + \left(\frac{1}{10}\right)x_2 + x_3 : 0 \le x_1 \le 1,\ 0 \le x_2 \le 1,\ 0 \le x_3 \le 1\right\}.$$

The plane passing through the feasible region is $\mathcal{C}(1,k)$, where τ^k is 1, and the curve on this subpolyhedron is the central path of

$$(LP^4) \quad \min\left\{x_1 + \left(\frac{1}{10}\right)x_2 : 0 \le x_1 \le 1,\ 0 \le x_2 \le 1,\ x_3 = x_3(1,b,c(1))\right\}.$$

The x_1, x_2, x_4, x_5, and x_6 values of this central path form the z variables defined by (3.2). Notice that the only difference between (LP^2) and (LP^4) is the value of x_3. This means that the central paths for (LP^2) and (LP^4) are the same except for the shift in x_3. Equation (3.4) shows how the shifted central path of (LP^4) intersects the perturbed central path of (LP^3).

The equality in (3.4) is important because z has the perturbations in both b and c modeled as right-hand side perturbations; i.e., there is no perturbation of the cost vector δc_B. This observation indicates that we need to understand the convergence properties of a central path under right-hand side perturbation. Lemma 3.2 states that the central solution is continuous with respect to b, and Lemma 3.4 shows that a perturbed central path converges to the analytic center of the unperturbed optimal set as long as there is no movement in c. We note that Lemma 3.4 is similar to Theorem 4.1 in [10], with the difference being that our result allows arbitrary perturbations in b.

LEMMA 3.2 (see Caron, Greenberg, and Holder [2]). *The analytic center of a bounded polyhedron is a continuous function of the right-hand side. That is, if $b^k \to b$ and $\mathcal{P}_b$ is bounded, $\lim_{k\to\infty} \bar{x}(b^k) = \bar{x}(b)$. (Note that this result is true for bounded polyhedrons that are not fully dimensional.)*

We note that since the central solution $x^*(b,c)$ is the analytic center of the polytope $\mathcal{P}^*_{(b,c)}$, we have that $x^*(b,c)$ is a continuous function of b. This is stated in the following corollary for future reference.

COROLLARY 3.3. *The central solution $x^*(b,c)$ is continuous with respect to the right-hand side b.*

LEMMA 3.4. *If $\mu^k \downarrow 0$, we have that $x(\mu^k, b^k, c) \to x^*(b,c)$.*

Proof. From Lemma 2.6 we have that $x_N(\mu^k, b^k, c) \to 0$, and from Lemma 3.1 we have that $x_B(\mu^k, b^k, c)$ is the unique solution to

$$\max\left\{\sum_{i\in B} \ln(x_i) : A_B x_B = b^k - A_N x_N(\mu^k, b^k, c),\ x_B > 0\right\}.$$

This means that $x_B(\mu^k, b^k, c)$ is the analytic center of $\{x_B : A_B x_B = b^k - A_N x_N(\mu^k, b^k, c), x_B \ge 0\}$, and from Lemma 3.2 we have that this analytic center is a continuous function of $b^k - A_N x_N(\mu^k, b^k, c)$. Since $b^k - A_N x_N(\mu^k, b^k, c) \to b$, we have that $x_B(\mu^k, b^k, c)$ converges to the analytic center of $\mathcal{P}^*_b = \{x : A_B x_B = b, x_B \ge 0\}$. ∎

We take a moment to summarize what we have. If μ^k has a positive limit, we have from (2.1) that $x(\mu^k, b^k, c(\tau^k))$ converges. The more difficult situation is if μ^k decreases to 0. From Lemma 2.6 we have that $x_N(\mu^k, b^k, c(\tau^k))$ decreases to zero as well. So, what is left to know is whether or not $x_B(\mu^k, b^k, c(\tau^k))$ converges. Since

$$x_B(\mu^k, b^k, c(\tau^k)) = z_B(\mu^k/\tau^k, b^k - A_N x_N(\mu^k, b^k, c(\tau^k)), \delta c_B),$$

we have from Lemma 3.4 that x_B converges as long as μ^k/τ^k and $b^k - A_N x_N(\mu^k, b^k, c(\tau^k))$ converge. Again, since $x_N(\mu^k, b^k, c(\tau^k))$ decreases to zero, we have that $b^k - A_N x_N(\mu^k, b^k, c(\tau^k)) \to b$. This means that $x_B(\mu^k, b^k, c(\tau^k))$ converges as long as μ^k/τ^k converges, a result that is stated in Theorem 3.5. This sufficient condition is "nearly" necessary for the sequence $x(\mu^k, b^k, c(\tau^k))$ to converge, with the problem being that if δc is in $\mathcal{H}^2(b, c)$, then $x(\mu^k, b^k, c(\tau^k))$ may converge even though μ^k/τ^k does not converge.

THEOREM 3.5. *Let $\tau^k \downarrow 0$ and $\mu^k > 0$ be such that $\mu^k \to \mu^0$. Then,*

$$\lim_{k\to\infty} x(\mu^k, b^k, c(\tau^k)) = \begin{cases} x(\mu^0, b, c) & \text{if } \mu^0 > 0, \\ x^*(b, c) & \text{if } \mu^0 = 0 \text{ and } \mu^k/\tau^k \to \infty, \\ (z_B(\eta, b, \delta c_B), 0) & \text{if } \mu^0 = 0 \text{ and } \mu^k/\tau^k \to \eta > 0, \\ (z_B^*(b, \delta c_B), 0) & \text{if } \mu^0 = 0 \text{ and } \mu^k/\tau^k \to 0. \end{cases}$$

Proof. The case of μ^0 being positive is an immediate consequence of (2.1). Assume $\mu^0 = 0$. From Lemma 2.6 we have that $x_N(\mu^k, b^k, c(\tau^k)) \to 0$. Consider the situation of $\mu^k/\tau^k \to \eta > 0$. Since μ^k/τ^k is bounded away from zero, we have from (2.1) that

$$x_B(\mu^k, b^k, c(\tau^k)) = z_B(\mu^k/\tau^k, b^k - A_N x_N(\mu^k, b^k, c(\tau^k)), \delta c_B) \to z_B(\eta, b, \delta c_B),$$

which establishes the third case. Suppose that $\mu^k/\tau^k \to 0$. We have from Lemma 3.4 that

$$x_B(\mu^k, b^k, c(\tau^k)) = z_B(\mu^k/\tau^k, b^k - A_N x_N(\mu^k, b^k, c(\tau^k)), \delta c_B) \to z_B^*(b, \delta c_B).$$

So, the fourth case is established. Finally, suppose that $\mu^k/\tau^k \to \infty$. Then, $\delta c_B/(\mu^k/\tau^k) = \tau^k \delta c_B/\mu^k \to 0 \in \text{row}(A_B)$, and since $\mathcal{P}^*_{(b,c)}$ is bounded, we have from Theorem 2.9 that

$$x_B(\mu^k, b^k, c(\tau^k)) = z_B(\mu^k/\tau^k, b^k - A_N x_N(\mu^k, b^k, c(\tau^k)), \delta c_B) \to x_B^*(b, c). \qquad \square$$

As previously stated, the reason why the conditions in Theorem 3.5 are not necessary is because if δc is in $\mathcal{H}^2(b, c)$, then $x(\mu^k, b^k, c(\tau^k))$ may converge even if μ^k/τ^k does not. Lemmas 3.6 and 3.7 address this issue.

LEMMA 3.6. *Let $\mu^k \downarrow 0$ and $\delta c \notin \mathcal{H}^2(b, c)$. Suppose that the sequence μ^k/τ^k does not converge. Then, if μ^{k_i}/τ^{k_i} and μ^{k_j}/τ^{k_j} are two convergent subsequences, we have that*

$$\lim_{i\to\infty} \mu^{k_i}/\tau^{k_i} \neq \lim_{j\to\infty} \mu^{k_j}/\tau^{k_j} \Rightarrow \lim_{i\to\infty} x(\mu^{k_i}, b^{k_i}, c(\tau^{k_i})) \neq \lim_{j\to\infty} x(\mu^{k_j}, b^{k_j}, c(\tau^{k_j})).$$

Proof. Without loss of generality, we assume that

$$\lim_{i\to\infty} \mu^{k_i}/\tau^{k_i} < \lim_{j\to\infty} \mu^{k_j}/\tau^{k_j}.$$

From Theorem 3.5 we have that

$$\lim_{i\to\infty} x(\mu^{k_i}, b^{k_i}, c(\tau^{k_i})) = \begin{cases} (z_B^*(b, \delta c_B), 0) & \text{if } \mu^{k_i}/\tau^{k_i} \to 0, \\ (z_B(\eta^1, b, \delta c_B), 0) & \text{if } \mu^{k_i}/\tau^{k_i} \to \eta^1 > 0, \end{cases}$$

and

$$\lim_{j\to\infty} x(\mu^{k_j}, b^{k_j}, c(\tau^{k_j})) = \begin{cases} (z_B(\eta^2, b, \delta c_B), 0) & \text{if } \mu^{k_j}/\tau^{k_j} \to \eta^2 < \infty, \\ x^*(b, c) & \text{if } \mu^{k_j}/\tau^{k_j} \to \infty. \end{cases}$$

Since $\delta c \notin \mathcal{H}^2$, we have from Lemma 2.4 that $\delta c_B \notin \text{col}(A_B)$. The result follows because from Theorem 2.2 we have that for any $\eta^1 < \eta^2$,

$$\delta c_B z_B^*(b, \delta c_B) < \delta c_B z_B(\eta^1, b, \delta c_B) < \delta c_B z_B(\eta^2, b, \delta c_B) < \delta c_B x_B^*(b, c). \qquad \square$$

LEMMA 3.7. *If $\delta c \in \mathcal{H}^2(b, c)$, we have for all positive η that*

$$x_B^*(b, c) = z_B(\eta, b, \delta c_B) = z_B^*(b, \delta c_B).$$

Proof. From Lemma 2.4 we have that $\delta c_B \in \text{row}(A_B)$, and from Theorem 2.3 we have that $z_B(\eta^1, b, \delta c_B) = z_B(\eta^2, b, \delta c_B)$ for all positive η^1 and η^2. Hence, for any positive η^0,

$$z_B^*(b, \delta c_B) = \lim_{\eta \downarrow 0} z_B(\eta, b, \delta c_B) = z_B(\eta^0, b, \delta c_B) = \lim_{\eta \to \infty} z_B(\eta, b, \delta c_B) = x_B^*(b, c). \qquad \square$$

Theorem 3.8 states the necessary and sufficient conditions for the convergence of $x(\mu^k, b^k, c(\tau^k))$.

THEOREM 3.8. *Let $\tau^k \downarrow 0$ and $\mu^k \downarrow 0$. If $\delta c \in \mathcal{H}^2(b, c)$, then $x(\mu^k, b^k, c(\tau^k)) \to x^*(b, c)$. Otherwise, $\delta c \notin \mathcal{H}^2(b, c)$, and $x(\mu^k, b^k, c(\tau^k))$ converges if and only if μ^k/τ^k converges.*

Proof. Suppose that $\delta c \in \mathcal{H}^2(b, c)$. From Lemma 2.6 we have that $x_N(\mu^k, b^k, c(\tau^k)) \to 0$. Also, from Lemma 3.7 we have that

$$\begin{aligned} x_B(\mu^k, b^k, c(\tau^k)) &= z_B(\mu^k/\tau^k, b^k - A_N x_N(\mu^k, b^k, c(\tau^k)), \delta c_B) \\ &= z_B^*(b^k - A_N x_N(\mu^k, b^k, c(\tau^k)), \delta c_B). \end{aligned}$$

From Lemma 3.2 we know that z_B^* is a continuous function of the right-hand side $b^k - A_N x_N(\mu^k, b^k, c(\tau^k))$. So, from Lemma 3.7 we have that

$$\begin{aligned} \lim_{k \to \infty} x(\mu^k, b^k, c(\tau^k)) &= \lim_{k \to \infty} \left(z_B^*(b^k - A_N x_N(\mu^k, b^k, c(\tau^k)), \delta c_B), x_N(\mu^k, b^k, c(\tau^k)) \right) \\ &= (z_B^*(b, \delta c_B), 0) \\ &= x^*(b, c). \end{aligned}$$

Assume that $\delta c \notin \mathcal{H}^2(b, c)$. If μ^k/τ^k converges, then Theorem 3.5 shows that $x(\mu^k, b^k, c(\tau^k))$ converges (and provides the limit). If μ^k/τ^k does not converge, this sequence has at least two cluster points, and hence there are two convergent subsequences, say, μ^{k_i}/τ^{k_i} and μ^{k_j}/τ^{k_j}, such that $\lim_{i \to \infty} \mu^{k_i}/\tau^{k_i} \neq \lim_{j \to \infty} \mu^{k_j}/\tau^{k_j}$. Theorem 3.5 implies that both

$$\lim_{i \to \infty} x(\mu^{k_i}, b^{k_i}, c(\tau^{k_i})) \qquad \text{and} \qquad \lim_{j \to \infty} x(\mu^{k_j}, b^{k_j}, c(\tau^{k_j}))$$

exist, and Lemma 3.6 implies that these limits are different. Hence, $x(\mu^k, b^k, c(\tau^k))$ does not converge. $\square$

We conclude this section by classifying the convergence of the perturbed central path followed by infeasible-path-following-interior-point algorithms. We require the dual counterpart of Theorem 3.8, which we state without proof.

THEOREM 3.9. *Let $\mu^k \downarrow 0$, $\rho^k \downarrow 0$, and $c^k \to c$. If $\delta b \in \mathcal{H}^1(b, c)$, then $(y(\mu^k, b(\rho^k), c^k), s(\mu^k, b(\rho^k), c^k)) \to (y^*(b, c), s^*(b, c))$. Otherwise, $\delta b \notin \mathcal{H}^1(b, c)$, and $(y(\mu^k, b(\rho^k), c^k), s(\mu^k, b(\rho^k), c^k))$ converges if and only if μ^k/ρ^k converges.*

As mentioned in section 1, the perturbed central path followed by an infeasible-path-following-interior-point algorithm has linear perturbations in b and c, with the directions of change defined by residuals. Table 3.1 shows the sequences whose convergence characterizes the convergence of the perturbed central path.

TABLE 3.1

Let δb and δc be defined by the residuals in (1.4). Depending on whether or not δb is in $\mathcal{H}^1(b,c)$ and δc is in $\mathcal{H}^2(b,c)$, we have that the convergence of the indicated sequences is required for, and guarantees, the convergence of $(x(\mu^k, b(\rho^k), c(\tau^k)), y(\mu^k, b(\rho^k), c(\tau^k)), s(\mu^k, b(\rho^k), c(\tau^k)))$.

	Cost perturbation	
	$\delta c \notin \mathcal{H}^2(b,c)$	$\delta c \in \mathcal{H}^2(b,c)$
	$\Updownarrow$	$\Updownarrow$
Right-hand side perturbation	$\delta c_B \notin \text{row}(A_B)$	$\delta c_B \in \text{row}(A_B)$
$\delta b \notin \mathcal{H}^1(b,c) \Leftrightarrow \delta b \notin \text{col}(A_B)$	μ^k/ρ^k & μ^k/τ^k	μ^k/ρ^k
$\delta b \in \mathcal{H}^2(b,c) \Leftrightarrow \delta b \in \text{col}(A_B)$	μ^k/τ^k	
	Must converge	

4. Set convergence. The objective of this section is to establish a set (Hausdorff) convergence property for the perturbed central path. Theorem 4.1 shows how the central path behaves as a set under simultaneous changes in b and c, provided that the change in c is linear. We illustrate the set convergence result with the following example.

Example 4.1. As in Example 3.1, consider the linear program

$$\min\{x_3 : 0 \le x_1 \le 1,\ 0 \le x_2 \le 1,\ 0 \le x_3 \le 1\}.$$

Let x_4, x_5, and x_6 be the slack variables, $b^k = b$ (so there is no right-hand side perturbation), and $\delta c = (1/4, 1/2000, 0, 0, 0, 0)$. The central paths corresponding to b and $c(\tau^k)$, for $\tau^k = 1, 0.8, 0.6, 0.4, 0.2$, are shown in Figure 4.1. The vertical line is the central path of the unperturbed problem, i.e., the vertical line is $PCP_{(b,c)}$. The curve in the x_1, x_2-plane is the central path for the linear program

$$\min\{1/4x_1 + 1/2000x_2 : 0 \le x_1 \le 1,\ 0 \le x_2 \le 1,\ x_3 = 0\} = \min\{\delta c_B x_B : x \in \mathcal{P}^*\}.$$

Observe that the perturbed central paths *converge* to these two central paths.

Example 4.1 indicates, and Theorem 4.1 proves, that the perturbed central paths converge to the union of two central paths. The first of these paths is $PCP_{(b,c)}$, i.e., the central path of the unperturbed linear program. The second of these paths is denoted by $PCP^*_{(b,c,\delta c)}$ and corresponds to minimizing $\delta c x$ over the optimal face. Hence, $PCP^*_{(b,c,\delta c)}$ is defined by the linear program

$$\min\{\delta c_B x_B : A_B x_B = b,\ x_B \ge 0,\ x_N = 0\}.$$

The elements of $PCP^*_{(b,c,\delta c)}$ have the form of $(z_B(\eta, b, \delta c_B), 0)$, and hence $PCP^*_{(b,c,\delta c)}$ is equipotent to $\{z_B(\eta, b, \delta c_B) : \eta > 0\}$. The closure of $PCP_{(b,c)}$ is $\overline{PCP}_{(b,c)}$ and is either $PCP_{(b,c)} \cup \{x^*(b,c)\} \cup \{\bar{x}(b)\}$ or $PCP_{(b,c)} \cup \{x^*(b,c)\}$, depending on whether or not the feasible region is bounded. The closure of $PCP^*_{(b,c,\delta c)}$ is $\overline{PCP^*}_{(b,c,\delta c)} = PCP^*_{(b,c,\delta c)} \cup \{(z^*_B(b, c_B), 0)\} \cup \{x^*(b,c)\}$.

THEOREM 4.1. *If $\tau^k \downarrow 0$, we have that $\overline{PCP}_{(b^k, c(\tau^k))} \to \overline{PCP}_{(b,c)} \cup \overline{PCP^*}_{(b,c,\delta c)}$.*

Proof. We begin by establishing that

$$PCP_{(b^k, c(\tau^k))} \to \overline{PCP}_{(b,c)} \bigcup \overline{PCP^*}_{(b,c,\delta c)}.$$

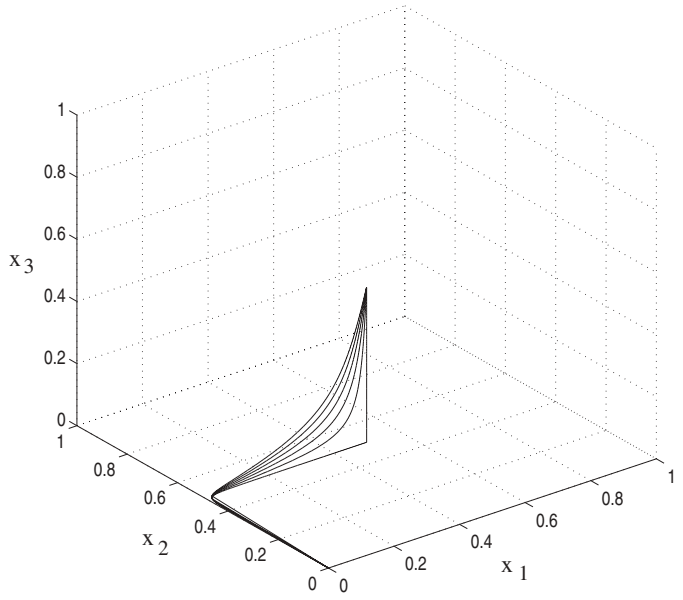


FIG. 4.1. *The central paths of the perturbed data converge to the union of two central paths.*

Let $x^k \in PCP_{(b^k,c(\tau^k))}$ be such that $x^k \to \hat{x}$. Then, for each k there is a μ^k such that $x^k = x(\mu^k, b^k, c(\tau^k))$. Let μ^{k_i} be a convergent subsequence of μ^k (remember that ∞ is a possible cluster point). We consider three cases to show that $\hat{x} \in \overline{PCP}_{(b,c)} \cup \overline{PCP^*}_{(b,c,\delta c)}$.

Case 1. If $\mu^{k_i} \to \hat{\mu} > 0$, we have from (2.1) that

$$x(\mu^{k_i}, b^{k_i}, c(\tau^{k_i})) \to x(\hat{\mu}, b, c) = \hat{x} \in \overline{PCP}_{(b,c)}.$$

Case 2. Suppose that $\mu^{k_i} \downarrow 0$. If $\delta c \in \mathcal{H}^2$, Theorem 3.8 shows that

$$x(\mu^{k_i}, b^{k_i}, c(\tau^{k_i})) \to x^*(b, c) = \hat{x} \in \overline{PCP}_{(b,c)}.$$

Otherwise, $\delta c \notin \mathcal{H}^2$, and Theorem 3.8 shows that μ^{k_i}/τ^{k_i} must converge. From Theorem 3.5 we have that

$$x(\mu^{k_i}, b^{k_i}, c(\tau^{k_i})) \to \hat{x} = \begin{cases} x^*(b,c) & \text{if } \mu^{k_i}/\tau^{k_i} \to \infty, \\ (z_B(\eta, b, \delta c_B), 0) & \text{if } \mu^{k_i}/\tau^{k_i} \to \eta > 0, \\ (z_B^*(b, \delta c_B), 0) & \text{if } \mu^{k_i}/\tau^{k_i} \to 0. \end{cases}$$

Since $x^*(b,c) \in \overline{PCP}_{(b,c)}$, and both $(z_B(\eta,b,\delta c_B),0)$ and $(z_B^*(b,\delta c_B),0)$ are in $\overline{PCP^*}_{(b,c,\delta c)}$, we have that $\hat{x}$ is in $\overline{PCP}_{(b,c)} \cup \overline{PCP^*}_{(b,c,\delta c)}$.

Case 3. Suppose that $\mu^{k_i} \to \infty$. Then, $c(\tau^{k_i})/\mu^{k_i} \to 0 \in \text{row}(A)$. If we knew that $\mathcal{P}_b$ was bounded, we would have from Theorem 2.9 that

$$x(\mu^{k_i}, b^{k_i}, c(\tau^{k_i})) \to \hat{x} = \bar{x}(b) \in \overline{PCP}_{(b,c)}.$$

So, our goal in this case becomes to use the fact that $x(\mu^{k_i}, b^{k_i}, c(\tau^{k_i}))$ converges as $\mu^{k_i} \to \infty$ to show that $\mathcal{P}_b$ is bounded. Let $x^i = x(\mu^{k_i}, b^{k_i}, c(\tau^{k_i}))$, $y^i = y(\mu^{k_i}, b^{k_i}, c(\tau^{k_i}))$,

and $s^i = s(\mu^{k_i}, b^{k_i}, c(\tau^{k_i}))$. From Gordon's theorem of the alternative, we have that $\mathcal{P}_b$ is bounded if and only if there is a row vector y such that $yA > 0$. For $j = 1, 2, \ldots, n$, we have that $x_j^i s_j^i = \mu^{k_i}$, $x_j^i \to \hat{x}_j$, and $\mu^{k_i} \to \infty$. Consequently, we have that $s_j^i \to \infty$. From the dual constraints we have that each component of $(-y^i)A = s^i - c(\tau^{k_i})$ approaches infinity, and hence the system $yA > 0$ is consistent. So, $\mathcal{P}_b$ is bounded.

At this point we have established that if $x^k \in \mathit{PCP}_{(b^k, c(\tau^k))}$ converges, then the limit of this sequence is in $\overline{\mathit{PCP}}_{(b,c)} \cup \overline{\mathit{PCP}^*}_{(b,c,\delta c)}$. We now show that any element in $\overline{\mathit{PCP}}_{(b,c)} \cup \overline{\mathit{PCP}^*}_{(b,c,\delta c)}$ is the limit of a sequence in $\mathit{PCP}_{(b^k, c(\tau^k))}$. Let x be in $\overline{\mathit{PCP}}_{(b,c)} \cup \overline{\mathit{PCP}^*}_{(b,c,\delta c)}$. Then, x is one of $\bar{x}(b)$ (if $\mathcal{P}_b$ is bounded), $x(\hat{\mu}, b, c)$ (for some positive $\hat{\mu}$), $x^*(b, c)$, $(z_B(\eta, b, \delta c_B), 0)$ (for some positive η), or $(z_B^*(b, \delta c_B), 0)$. From Theorems 2.9 and 3.5 we have for $\tau^k = 1/k$ that

$$
\begin{aligned}
&x(\hat{\mu} + 1/k, b^k, c(\tau^k)) \to x(\hat{\mu}, b, c), \quad x\big(\sqrt{\tau^k}, b^k, c(\tau^k)\big) \to x^*(b, c),\\
&x(\eta\tau^k, b^k, c(\tau^k)) \to (z_B(\eta, b, \delta c_B), 0), \quad x((\tau^k)^2, b^k, c(\tau^k)) \to (z_B^*(b, \delta c_B), 0),\\
&x(k, b^k, c(\tau^k)) \to \bar{x}(b) \quad \text{(if } \mathcal{P}_b \text{ is bounded)}.
\end{aligned}
$$

Since all four of these sequences are in $\mathit{PCP}_{(b^k, c(\tau^k))}$, we have that

$$\mathit{PCP}_{(b^k, c(\tau))} \to \overline{\mathit{PCP}}_{(b,c)} \cup \overline{\mathit{PCP}^*}_{(b,c,\delta c)}.$$

What remains to be shown is that if the sequence $x^k \in \overline{\mathit{PCP}}_{(b^k, c(\tau^k))}$ converges and contains either $x^*(b^k, c(\tau^k))$ or, in the case that $\mathcal{P}_b$ is bounded, $\bar{x}(b^k)$ infinitely many times, the limit of this sequence is in $\overline{\mathit{PCP}}_{(b,c)} \cup \overline{\mathit{PCP}^*}_{(b,c,\delta c)}$. If $\mathcal{P}_b$ is bounded, we have from Lemma 2.7 that $\mathcal{P}_{b^k}$ is bounded for sufficiently large k. Furthermore, Lemma 3.2 shows that $\bar{x}(b)$ is a continuous function of b. So, if $x^k \in \overline{\mathit{PCP}}_{(b^k, c(\tau^k))}$ contains $\bar{x}(b^k)$ infinitely many times and converges to $\hat{x}$, we have that $\hat{x} = \bar{x}(b) \in \overline{\mathit{PCP}}_{(b,c)}$. Suppose that $x^k \in \overline{\mathit{PCP}}_{(b^k, c(\tau^k))}$ converges to $\hat{x}$ and that this sequence contains $x^*(b^k, c(\tau^k))$ infinitely many times. Without loss of generality, we assume that $x^k = x^*(b^k, c(\tau^k))$. First, because $(B|N)$ need not be the same as $(B(b^k, c(\tau^k))|N(b^k, c(\tau^k)))$, we do not automatically know that $x_N^k = x_N^*(b^k, c(\tau^k)) \to 0$ (Lemma 2.6 does not apply). However, x_N^k does converge to 0 as the following argument shows. Let $\varepsilon > 0$. For each k we have that

$$x_N^k = x_N^*(b^k, c(\tau^k)) = \lim_{\mu \downarrow 0} x_N(\mu, b^k, c(\tau^k)).$$

So, there is a positive $\hat{\mu}^k$ such that $\mu \in (0, \hat{\mu}^k)$ implies that $\|x_N(\mu, b^k, c(\tau^k)) - x_N^*(b^k, c(\tau^k))\| < \varepsilon/2$. Choose $\mu^k \in (0, \hat{\mu}^k)$ so that $\mu^k \downarrow 0$. From Lemma 2.6 we have that $x_N(\mu^k, b^k, c(\tau^k)) \to 0$. Hence, there exists a natural number K such that for $k \geq K$, we have $\|x_N(\mu^k, b^k, c(\tau^k))\| < \varepsilon/2$. Hence, for $k \geq K$,

$$\|x_N^*(b^k, c(\tau^k))\| \leq \|x_N^*(b^k, c(\tau^k)) - x_N(\mu^k, b^k, c(\tau^k))\| + \|x_N(\mu^k, b^k, c(\tau^k))\| < \varepsilon.$$

So, $x_N^k = x_N^*(b^k, c(\tau^k)) \to 0$. Using this fact, Lemma 3.2 to establish the fifth equality, and Lemma 3.4 to establish the fourth equality, we have that

$$
\begin{aligned}
\hat{x}_B = \lim_{k\to\infty} x_B^*(b^k, c(\tau^k)) &= \lim_{k\to\infty} \Big(\lim_{\mu \downarrow 0} x_B(\mu, b^k, c(\tau^k)) \Big)\\
&= \lim_{k\to\infty} \Big(\lim_{\mu \downarrow 0} z_B(\mu/\tau^k, b^k - A_N x_N(\mu, b^k, c(\tau^k)), \delta c_B) \Big)\\
&= \lim_{k\to\infty} z_B^*(b^k - A_N x_N^*(b^k, c(\tau^k)), \delta c_B)\\
&= z_B^*(b, \delta c_B).
\end{aligned}
$$

Hence, we have that $x^k = x^*(b^k, c(\tau^k)) \to (z_B^*(b, \delta c_B), 0) \in \overline{PCP^*}_{(b,c,\delta c)}$, which completes the proof. ∎

A corollary to Theorem 4.1 is that the perturbed central path is continuous over $\mathcal{H}^2(b,c)$, meaning that as long as $\delta c \in \mathcal{H}^2(b,c)$, $\overline{PCP}_{(b^k, c(\tau^k))} \to \overline{PCP}_{(b,c)}$. This follows because if $\delta c \in \mathcal{H}^2(b,c)$, we have from Lemma 3.7 that

$$\overline{PCP^*}_{(b,c,\delta c)} = \{x^*(b,c)\} \subset \overline{PCP}_{(b,c)}.$$

This result is stated in the following corollary.

COROLLARY 4.2. *We have that if $\delta c \in \mathcal{H}^2(b,c)$ and $\tau^k \downarrow 0$, then $\overline{PCP}_{(b^k, c(\tau^k))} \to \overline{PCP}_{(b,c)}$.*

We conclude this section by showing why our results are stated from the primal perspective. This is because it is possible for b^k, $c(\tau^k)$, and $x(\mu^k, b^k, c(\tau^k))$ to converge, while the dual elements diverge. For example, suppose that $c \in \text{row}(A)$, which implies that

- $\mathcal{P}_b$ is bounded;
- $(B|N) = (\{1, 2, \ldots, n\}|\emptyset)$;
- $x(\mu^k, b^k, c(\tau^k)) = x_B(\mu^k, b^k, c(\tau^k)) = z_B(\mu^k/\tau^k, b^k, \delta c_B)$;
- $x^*(b,c) = \bar{x}(b)$.

Let $\tau^k \downarrow 0$ and μ^k be the sequence $1, 2, 1, 2, 1, 2, \ldots$. Then, $\mu^k/\tau^k \to \infty$, and we have from Corollary 2.10 that $x(\mu^k, b^k, c(\tau^k)) = z_B(\mu^k/\tau^k, b^k, \delta c_B) \to x^*(b,c) = \bar{x}(b)$. However, Theorem 2.3 implies that the corresponding dual sequence $s(\mu^k, b^k, c(\tau^k))$ has the two cluster points of $s(1,b,c)$ and $s(2,b,c) = 2s(1,b,c)$. The problem here is that $s_i(\mu^k, b^k, c(\tau^k)) = \mu^k / x_i(\mu^k, b^k, c(\tau^k))$, and we see that the dual elements fail to converge because the sequence μ^k does not converge. To guarantee the convergence of $s(\mu^k, b^k, c(\tau^k))$, one needs to guarantee the convergence of $\mu^k / x_i(\mu^k, b^k, c(\tau^k))$, $i = 1, 2, \ldots, n$ (which is not implied by the convergence of μ^k and $x(\mu^k, b^k, c(\tau^k))$). While Theorem 4.3 does not completely resolve this issue, it does show when the convergence of μ^k is guaranteed.

THEOREM 4.3. *Let $\tau^k \downarrow 0$. Then, the convergence of $x(\mu^k, b^k, c(\tau^k))$ implies the convergence of μ^k if and only if $c \notin \text{row}(A)$.*

Proof. Assume that $c \in \text{row}(A)$. Then, as discussed after Theorem 2.3, $\mathcal{P}_b$ is bounded. Let $\mu^k = 1, 2, 1, 2, \ldots$ and $\tau^k = 1/k$. Then, $\mu^k/\tau^k \to \infty$, and as just discussed, $x(\mu^k, b^k, c(\tau^k)) \to \bar{x}(b)$. Hence, the convergence of $x(\mu^k, b^k, c(\tau^k))$ cannot guarantee the convergence of μ^k.

Assume that $c \notin \text{row}(A)$ and suppose, for the sake of contradiction, that μ^k does not converge. Then there exist subsequences μ^{k_i} and μ^{k_j} such that

$$0 \le \lim_{i \to \infty} \mu^{k_i} < \lim_{j \to \infty} \mu^{k_j} \le \infty.$$

If $\mu^{k_i} \to \mu^1 > 0$, we have from (2.1) that $x(\mu^{k_i}, b^{k_i}, c(\tau^{k_i})) \to x(\mu^1, b, c)$. From (2.1) and Corollary 2.10 we have that

$$x(\mu^{k_j}, b^{k_j}, c(\tau^{k_j})) \to \begin{cases} x(\mu^2, b, c) & \text{if } \mu^{k_j} \to \mu^2 < \infty, \\ \bar{x}(b) & \text{if } \mu^{k_j} \to \infty. \end{cases}$$

However, Theorem 2.2 shows that $cx(\mu^1, b, c) < cx(\mu^2, b, c) < c\bar{x}(b)$, where the last inequality is included only when $\bar{x}$ exists. This is a contradiction since it implies that

$$\lim_{i \to \infty} x(\mu^{k_i}, b^{k_i}, c(\tau^{k_i})) \neq \lim_{j \to \infty} x(\mu^{k_j}, b^{k_j}, c(\tau^{k_j})).$$

The only situation left is that of $\mu^{k_i} \downarrow 0$. However, if $\mu^{k_i} \downarrow 0$, we have the contradiction from Lemma 2.6 that

$$0 = \lim_{i\to\infty} x_N(\mu^{k_i}, b^{k_i}, c(k_i)) \neq \lim_{j\to\infty} x_N(\mu^{k_j}, b^{k_j}, c(\tau^{k_j})) > 0. \qquad \square$$

In this section, we have shown that while the limit of a central path is not continuous in b and c, the perturbed central paths are well behaved if viewed as a set. Moreover, from Corollary 4.2 we have that the central path is continuous over $\mathcal{H}^2(b, c)$.

5. Independent, nonlinear perturbations. In this section we remove the restriction that the perturbation in c must be linear. The analysis increases in difficulty, and characterizing the convergence of the perturbed central path under arbitrary, simultaneous, and independent perturbations in b and c remains an open question. We provide sufficient conditions to guarantee the convergence of $x(\mu^k, b^k, c^k)$ and develop a process to find the limit. An example illustrates the difficulties of establishing exactly when $x(\mu^k, b^k, c^k)$ converges.

The sufficient conditions require that $\mathcal{G}^2$ be partitioned into equivalence classes. For any $b \in \mathcal{G}^1$, we say that c^1 and c^2 in $\mathcal{G}^2$ are "A-similar," denoted by $c^1 \overset{A}{\sim} c^2$, if $PCP_{(b,c^1)} \cap PCP_{(b,c^2)} \neq \emptyset$. The first goal of this section is to show that $\overset{A}{\sim}$ is an equivalence relation on $\mathcal{G}^2$. We begin by showing that central paths may not intersect unless they are equal. The first lemma provides sufficient conditions for two primal central paths to be equivalent.

LEMMA 5.1. *Let* $c_0^1 = \text{proj}_{\text{null}(A)}\, c^1$ *and* $c_0^2 = \text{proj}_{\text{null}(A)}\, c^2$. *Then,* $PCP_{(b,c_0^1)} = PCP_{(b,c^1)}$ *and* $PCP_{(b,c_0^2)} = PCP_{(b,c^2)}$. *Moreover, if* $c_0^1 = \alpha c_0^2$ *for some positive* α, $PCP_{(b,c^1)} = PCP_{(b,c^2)}$.

Proof. Let $c_R^1 = \text{proj}_{\text{row}(A)}\, c^1$ and $c_R^2 = \text{proj}_{\text{row}(A)}\, c^2$ so that $c^1 = c_0^1 + c_R^1$ and $c^2 = c_0^2 + c_R^2$. Let $\alpha > 0$ be such that $c_0^1 = \alpha c_0^2$. Since c_R^1 and c_R^2 are in row(A), we have from Theorem 2.3 that $c_R^1 x$ and $c_R^2 x$ are constant on $\mathcal{P}_b$. This means that $x(\mu, b, c^1)$ and $x(\mu, b, c^2)$ are, respectively, the unique solutions to

$$\min\left\{c_0^1 x - \mu \sum_{i=1}^n \ln(x_i) : x \in \mathcal{P}_b^o\right\} \qquad \text{and} \qquad \min\left\{c_0^2 x - \mu \sum_{i=1}^n \ln(x_i) : x \in \mathcal{P}_b^o\right\}.$$

Hence, $PCP_{(b,c_0^1)} = PCP_{(b,c^1)}$ and $PCP_{(b,c_0^2)} = PCP_{(b,c^2)}$. Multiplying the objective function of the second math program by α shows that $x(\alpha\mu, b, c^1) = x(\mu, b, c^2)$, which implies that $PCP_{(b,c^1)} = PCP_{(b,c^2)}$. $\square$

The following corollary is stated for future reference.

COROLLARY 5.2. *If* $\text{proj}_{\text{null}(A)}\, c^1 = \alpha\, \text{proj}_{\text{null}(A)}\, c^2$ *for some* $\alpha > 0$, *then*

$$x(\mu, b, c^1) = x(\mu, b, \text{proj}_{\text{null}(A)}\, c^1) = x(\alpha\mu, b, \text{proj}_{\text{null}(A)}\, c^2) = x(\alpha\mu, b, c^2).$$

Proof. The result is immediate from the proof of Lemma 5.1. $\square$

The next theorem establishes that the central paths within a polyhedron are either the same or disjoint. Since $PCP_{(b,c)}$ contains only those elements that correspond to a positive μ, this does not say that two different central paths may not terminate at the same point. However, it does say that two different central paths may not cross en route to either $x^*(b, c)$ or $\bar{x}(b)$.

THEOREM 5.3. *If* $PCP_{(b,c^1)} \cap PCP_{(b,c^2)} \neq \emptyset$, $PCP_{(b,c^1)} = PCP_{(b,c^2)}$.

Proof. From Corollary 5.2 we know that there is no loss of generality by assuming that c^1 and c^2 are in null(A). Let μ^1 and μ^2 be positive such that $x(\mu^1, b, c^1) =$

$x(\mu^2, b, c^2)$. Since $s(\mu^1, b, c^1)X(\mu^1, b, c^1) = \mu^1 e^T$ and $s(\mu^2, b, c^1)X(\mu^2, b, c^1) = \mu^2 e^T$, we have that $s(\mu^1, b, c^1) = \mu^1 e^T X^{-1}(\mu^1, b, c^1)$ and $s(\mu^2, b, c^1) = \mu^2 e^T X^{-1}(\mu^2, b, c^1)$. From the dual feasibility constraints we have that

$$c^1 - \mu^1 e^T X^{-1}(\mu^1, b, c^1) - y(\mu^1, b, c^1)A = 0,$$
$$c^2 - \mu^2 e^T X^{-1}(\mu^2, b, c^2) - y(\mu^2, b, c^2)A = 0.$$

Multiplying the first equation by $1/\mu^1$, the second equation by $1/\mu^2$, and subtracting yields

$$(1/\mu^1)c^1 - (1/\mu^2)c^2 = ((1/\mu^1)y(\mu^1, b, c^1) - (1/\mu^2)y(\mu^2, b, c^2))A.$$

Since the left-hand side is in the null(A) and the right-hand side is in the row(A), both must be zero. Hence, $c^1 = (\mu^1/\mu^2)c^2$, and from Lemma 5.1 we have that $PCP_{(b,c^1)} = PCP_{(b,c^2)}$. □

Two important corollaries follow.

COROLLARY 5.4. *If* $c^1 \overset{A}{\sim} c^2$, $PCP_{(b,c^1)} = PCP_{(b,c^2)}$.

COROLLARY 5.5. *We have that* $\operatorname{proj}_{\text{null}(A)} c^1 = \alpha \operatorname{proj}_{\text{null}(A)} c^2$, *for some positive* α *if and only if* $PCP_{(b,c^1)} = PCP_{(b,c^2)}$.

Proof. The sufficiency is established by Lemma 5.1. The necessity follows because if $PCP_{(b,c^1)} = PCP_{(b,c^2)}$, then there are a positive μ^1 and μ^2 such that $x(\mu^1, b, c^1) = x(\mu^2, b, c^2)$, and from the proof of Theorem 5.3 we have that $\operatorname{proj}_{\text{null}(A)} c^1 = \alpha \operatorname{proj}_{\text{null}(A)} c^2$ for some positive α. □

Theorem 5.6 states that $\overset{A}{\sim}$ is indeed an equivalence relation.

THEOREM 5.6. $\overset{A}{\sim}$ *is an equivalence relation on* $\mathcal{G}^2$. *Furthermore, the equivalence class of* c^1 *is*

$$[c^1]_A = \{c : \operatorname{proj}_{\text{null}(A)} c^1 = \alpha \operatorname{proj}_{\text{null}(A)} c \text{ for some positive } \alpha\}.$$

Proof. Clearly $c^1 \overset{A}{\sim} c^1$ and, if $c^1 \overset{A}{\sim} c^2$, then $c^2 \overset{A}{\sim} c^1$. So $\overset{A}{\sim}$ is reflexive and symmetric. From Corollary 5.4 we have that if $c^1 \overset{A}{\sim} c^2$ and $c^2 \overset{A}{\sim} c^3$, then $PCP_{(b,c^1)} = PCP_{(b,c^2)} = PCP_{(b,c^3)}$, which implies that $c^1 \overset{A}{\sim} c^3$. Hence, $\overset{A}{\sim}$ is transitive and an equivalence relation. From Theorem 5.3 and Corollary 5.5 we have that the equivalence classes hold as stated. □

Our conditions that guarantee the convergence of $x(\mu^k, b^k, c^k)$ rely on two new types of convergence. For a sequence x^k, we let $\mathbf{C}(x^k)$ be the set of cluster points of x^k. Furthermore, for any sequence c^k, we set $d^k = c^k/\|c^k\|$ as long as $c^k \neq 0$, and we define $\mathcal{F}(c^k)$ to be

$$\mathcal{F}(c^k) = \mathbf{C}(c^k) \cup \mathbf{C}(d^k).$$

In addition to the cluster points of c^k, the set $\mathcal{F}$ contains the "limiting directions" of the cost vectors. For example, if c^k is $(1/k, 1/k)$ for k even and (k, k^2) for k odd, $\mathbf{C}(c^k) = \{(0,0)\}$ and $\mathbf{C}(d^k) = \{(1/\sqrt{2}, 1/\sqrt{2}), (0,1)\}$. We say that c^k is *class convergent* if the cluster points of c^k and the limiting directions of c^k are contained in the same equivalence class.

DEFINITION 5.7. *The sequence* c^k *is class convergent to* $[c]_A$ *if* $\mathcal{F}(c^k) \subseteq [c]_A$.

DEFINITION 5.8. *The sequence* (μ^k, c^k) *is proportionately convergent if for any two subsequences, say* c^{k_i} *and* c^{k_j}, *having the property that*

$$\lim_{i\to\infty} \operatorname{proj}_{\text{null}(A)} c^{k_i}/\|c^{k_i}\| = \alpha \lim_{j\to\infty} \operatorname{proj}_{\text{null}(A)} c^{k_j}/\|c^{k_j}\|,$$

we subsequently have that

$$\lim_{i\to\infty} \mu^{k_i}/\|c^{k_i}\| = \alpha \lim_{j\to\infty} \mu^{k_j}/\|c^{k_j}\|.$$

We point out that a proportionately convergent sequence may have the property that c^k contains a subsequence of zeros; however, this subsequence is not a candidate for either c^{k_i} or c^{k_j}.

The next theorem provides sufficient conditions for $x(\mu^k, b^k, c^k)$ to converge to an element of a central path. The sequence c^k is not required to converge but is instead required to be class convergent. As Example 5.1 demonstrates, this weaker condition on c^k is still too restrictive for necessity.

THEOREM 5.9. *We have that $x(\mu^k, b^k, c^k)$ converges to an element of $PCP_{(b,c)}$ provided that*

1. *c^k is class convergent to $[c]_A$,*
2. *(μ^k, c^k) is proportionately convergent,*
3. *$c^k \neq 0$ for $k = 1, 2, 3, \ldots$, and*
4. *$\mu^k/\|c^k\| = \Theta(1)$.*

Proof. Since $\mu^k/\|c^k\| = \Theta(1)$ and $x(\mu^k, b^k, c^k) = x(\mu^k/\|c^k\|, b^k, c^k/\|c^k\|)$, we have from Lemma 2.5 that $x(\mu^k, b^k, c^k)$ is bounded. The result is established by showing that all cluster points of $x(\mu^k, b^k, c^k)$ are equal. Consider the subsequences

$$x(\mu^{k_i}, b^{k_i}, c^{k_i}) \to \hat{x}^1, \quad x(\mu^{k_j}, b^{k_j}, c^{k_j}) \to \hat{x}^2, \quad c^{k_i}/\|c^{k_i}\| \to \hat{c}^1, \quad \text{and} \quad c^{k_j}/\|c^{k_j}\| \to \hat{c}^2.$$

From the class convergence we have that there is a positive α^1 and α^2 such that

$$\begin{aligned}
\lim_{i\to\infty} \alpha^1 \operatorname{proj}_{\operatorname{null}(A)} c^{k_i}/\|c^{k_i}\| &= \alpha^1 \operatorname{proj}_{\operatorname{null}(A)} \hat{c}^1 \\
&= \operatorname{proj}_{\operatorname{null}(A)} c \\
&= \alpha^2 \operatorname{proj}_{\operatorname{null}(A)} \hat{c}^2 \\
&= \lim_{j\to\infty} \alpha^2 \operatorname{proj}_{\operatorname{null}(A)} c^{k_j}/\|c^{k_j}\|.
\end{aligned}$$

From the proportional convergence of (μ^k, c^k) and the assumption that $\mu^k/\|c^k\|$ is bounded away from zero, we have that

$$0 < \hat{\mu} = \lim_{i\to\infty} \alpha^1 \mu^{k_i}/\|c^{k_i}\| = \lim_{j\to\infty} \alpha^2 \mu^{k_j}/\|c^{k_j}\|.$$

From Corollary 5.2 we see that

$$\begin{aligned}
x(\mu^{k_i}, b^{k_i}, c^{k_i}) &= x(\alpha^1 \mu^{k_i}/\|c^{k_i}\|, b^{k_i}, \alpha^1 \operatorname{proj}_{\operatorname{null}(A)} c^{k_i}/\|c^{k_i}\|), \\
x(\mu^{k_j}, b^{k_j}, c^{k_j}) &= x(\alpha^2 \mu^{k_j}/\|c^{k_j}\|, b^{k_j}, \alpha^2 \operatorname{proj}_{\operatorname{null}(A)} c^{k_j}/\|c^{k_j}\|).
\end{aligned}$$

We now have from (2.1) that

$$\begin{aligned}
\hat{x}^1 &= \lim_{i\to\infty} x(\mu^{k_i}, b^{k_i}, c^{k_i}) \\
&= \lim_{i\to\infty} x(\alpha^1 \mu^{k_i}/\|c^{k_i}\|, b^{k_i}, \alpha^1 \operatorname{proj}_{\operatorname{null}(A)} c^{k_i}/\|c^{k_i}\|) \\
&= x(\hat{\mu}, b, \operatorname{proj}_{\operatorname{null}(A)} c) \\
&= \lim_{j\to\infty} x(\alpha^2 \mu^{k_j}/\|c^{k_j}\|, b^{k_j}, \alpha^2 \operatorname{proj}_{\operatorname{null}(A)} c^{k_j}/\|c^{k_j}\|) \\
&= \lim_{j\to\infty} x(\mu^{k_j}, b^{k_j}, c^{k_j}) \\
&= \hat{x}^2.
\end{aligned}$$

Hence, $x(\mu^k, b^k, c^k)$ converges to an element in $PCP_{(b,c)}$. ∎

We point out that Theorem 5.9 guarantees convergence only to an element of $PCP_{(b,c)}$, and hence every component of the limit is positive. This is guaranteed in the proof by the condition that $\mu^k/\|c^k\| = \Theta(1)$. The situation is more complicated if $\mu^k/\|c^k\| \downarrow 0$, and we illustrate the increased complication in the following example. This example has the desirable property that c^k converges, but even with this property the convergence of $x(\mu^k, b^k, c^k)$ requires the analysis of several nested linear programs. The example shows how we construct the *induced* sequences of (μ^k, b^k, c^k).

Example 5.1. Consider the linear program

$$\min\{(1/k)x_1 + (1/\sqrt{k})x_2 + (1/\sqrt{k})x_3 : 0 \le x_1 \le 1,\ 0 \le x_2 \le 1,\ 0 \le x_3 \le 1\}.$$

Let $\mu^k = 1/k$, and let x_4, x_5, and x_6 be the slack vectors. We point out that Theorem 5.9 does not apply because $\mu^k/\|c^k\| = 1/\sqrt{1+2k} \downarrow 0$. We consider a sequence of linear programs to analyze the convergence of $x(\mu^k, b^k, c^k)$. The idea is to iteratively reduce the original problem by "linearizing" the cost-coefficient perturbations and then use the results from section 3 to identify a collection of variables that must be zero.

The "root" problem is defined by the limit of c^k, which is 0, and is

$$(LP^0)\quad \min\{0x_1 + 0x_2 + 0x_3 : 0 \le x_1 \le 1,\ 0 \le x_2 \le 1,\ 0 \le x_3 \le 1\}.$$

The optimal partition for (LP^0) is $(B^1|N^1) = (\{1,2,3,4,5,6\}|\emptyset)$. We linearize c^k by rewriting it as

$$\begin{aligned} c^k &= 0 + \|c^k - 0\| \left(\frac{c^k - 0}{\|c^k - 0\|} \right) \\ &= 0 + \frac{\sqrt{2k+1}}{k} \begin{pmatrix} & & 1 \\ & & \sqrt{k} \\ & \frac{1}{\sqrt{2k+1}} & \sqrt{k} \\ & & 0 \\ & & 0 \\ & & 0 \end{pmatrix}. \end{aligned}$$

We let $\hat{c} = 0$, $\tau^k = \|c^k - 0\|$, and $\delta c^k = (c^k - 0)/\|c^k - 0\|$ so that

$$x(\mu^k, b^k, c^k) = x(\mu^k, b^k, \hat{c} + \tau^k \delta c^k).$$

The constant term $\hat{c}$ is used in the root problem to identify the optimal partition $(B^1|N^1)$, and from Lemma 2.6 we know that the variables indexed by N^1 are zero for every cluster point of $x(\mu^k, b^k, c^k)$. Unfortunately, these may or may not be the only variables that are zero (and in this example none of the zero variables are indexed by N^1 because it is empty). The variables in N^1 are essentially removed from the problem by redefining the right-hand side to be $b^k - A_{N^1} x_{N^1}(\mu^k, b^k, c^k)$, which in this case is simply b^k. If N^1 had not been empty, this would have reduced the number of variables in the problem. The first *induced* subsequence of (μ^k, b^k, c^k) is $(\mu^k/\tau^k, b^k - A_{N^1} x_{N^1}(\mu^k, b^k, c^k), \delta c^k_{B^1}) = (\mu^{(1,k)}, b^{(1,k)}, \delta c^{(1,k)}_{B^1})$, where the first superscript indicates that this is the first induced subsequence. From (3.4) we have that

$$x_{B^1}(\mu^k, b^k, \hat{c} + \tau^k \delta c^k) = x_{B^1}(\mu^k/\tau^k, b^k - A_{N^1} x_{N^1}(\mu^k, b^k, c^k), \delta c^k_{B^1}),$$

where the function on the right-hand side is the z function. We now have that

$$\left.\begin{aligned} x_{B^1}(\mu^k, b^k, c^k) &= x_{B^1}(\mu^k, b^k, \hat{c} + \tau^k \delta c^k) \\ &= x_{B^1}(\mu^k/\tau^k, b^k - A_{N^1} x_{N^1}(\mu^k, b^k, c^k), \delta c^k_{B^1}) \\ &= x_{B^1}\big(\mu^{(1,k)}, b^{(1,k)}, \delta c^{(1,k)}_{B^1}\big). \end{aligned}\right\} \tag{5.1}$$

If $\delta c^k_{B^1}$ had been constant, we could have established the limit of $x(\mu^k, b^k, c^k)$ from Theorems 3.5 and 3.8, and this limit would have been on the central path defined by minimizing $\delta c^k_{B^1} x_{B^1}$ over the optimal face of the root problem. However, $\delta c^k_{B^1}$ is not constant, and we repeat the process by linearizing the new cost coefficients $\delta c^{(1,k)}_{B^1}$.

Notice that the sequence $\delta c^{(1,k)}_{B^1}$ does not converge to zero, but rather $\delta c^{(1,k)}_{B^1} \to (0, 1/\sqrt{2}, 1\sqrt{2}, 0, 0, 0)^T = \hat{\delta c}_{B^1}$. The first subproblem is defined by this limit and is

$$(LP^1) \quad \min\{(1/\sqrt{2})x_2 + (1/\sqrt{2})x_3 : 0 \le x_i \le 1,\ i = 1, 2, 3\}.$$

The optimal partition for LP^1 is $(B^2|N^2) = (\{1, 4, 5, 6\}|\{2, 3\})$, which partitions B^1. As before, we linearize $\delta c^{(1,k)}_{B^1}$ by rewriting it as

$$\delta c^{(1,k)}_{B^1} = \hat{\delta c}_{B^1} + \|\delta c^{(1,k)}_{B^1} - \hat{\delta c}_{B^1}\| \left(\frac{\delta c^{(1,k)}_{B^1} - \hat{\delta c}_{B^1}}{\|\delta c^{(1,k)}_{B^1} - \hat{\delta c}_{B^1}\|} \right).$$

If we let $\tau^{(2,k)} = \|\delta c^{(1,k)}_{B^1} - \hat{\delta c}_{B^1}\|$ and $\delta c^{(2,k)}_{B^1} = (\delta c^{(1,k)}_{B^1} - \hat{\delta c}_{B^1})/\|\delta c^{(1,k)}_{B^1} - \hat{\delta c}_{B^1}\|$, then similar to (5.1) we have that

$$\begin{aligned} x_{B^2}(\mu^k, b^k, c^k) &= x_{B^2}\big(\mu^{(1,k)}, b^{(1,k)}, \delta c^{(1,k)}_{B^1}\big) \\ &= x_{B^2}\big(\mu^{(1,k)}, b^{(1,k)}, \hat{\delta c}_{B^1} + \tau^{(2,k)} \delta c^{(2,k)}_{B^1}\big) \\ &= x_{B^2}\big(\mu^{(1,k)}/\tau^{(2,k)}, b^{(1,k)} - A_{N^2} x_{N^2}(\mu^k, b^k, c^k), \delta c^{(2,k)}_{B^2}\big). \end{aligned}$$

From Lemma 2.6 we have that the components indexed by N^2 are zero in every cluster point of $x(\mu^k, b^k, c^k)$, and we have moved these variables to the right-hand side in the last equality (this is the first reduction for this example because N^1 was empty). The remaining components are indexed by $B^2 \subseteq B^1$. The second induced sequence of (μ^k, b^k, c^k) is $(\mu^{(1,k)}/\tau^{(2,k)}, b^{(1,k)} - A_{N^2} x_{N^2}(\mu^k, b^k, c^k), \delta c^{(2,k)}_{B^2}) = (\mu^{(2,k)}, b^{(2,k)}, \delta c^{(2,k)}_{B^2})$.

It is easily checked that $\mu^{(2,k)} = \sqrt{2/(2 + 2/(\sqrt{2k} + \sqrt{2k+1})^2)} \to 1$, which is important because $\mu^{(2,k)}$ does *not* converge to zero. The second subproblem requires only the B^2 components of the limit of $\delta c^{(2,k)}_{B^1}$, and it is easy to check that $\delta c^{(2,k)}_{B^2} \to (1, 0, 0, 0)$ (the first component corresponds to x_1 and the zero elements correspond with the slack variables x_4, x_5, and x_6). The second subproblem is

$$(LP^2) \quad \min\{x_1 : 0 \le x_1 \le 1\}.$$

Since $\mu^{(2,k)} \to 1$, we have from (2.1) that

$$x_{B^2}(\mu^k, b^k, c^k) = x_{B^2}\big(\mu^{(2,k)}, b^{(2,k)}, \delta c^{(2,k)}_{B^1}\big) \to x_{B^2}(1, b, \hat{\delta c}_{B^2}).$$

We have that $x_1(1, b, \hat{\delta c}_{B^2})$ is the unique solution to $\min\{x_1 + \ln(x_1) + \ln(1 - x_1) : 0 \le x_1 \le 1\}$, and a straightforward calculation shows that $x_1(1, b, \hat{\delta c}_{B^2}) = (3 - \sqrt{5})/2$. We conclude that

$$x(\mu^k, b^k, c^k) \to ((3 - \sqrt{5})/2, 0, 0, (\sqrt{5} - 1)/2, 1, 1)^T.$$

TABLE 5.1
The process for constructing the induced sequences of (μ^k, b^k, c^k).

Step 1 Set $j = 0$, $(B^0|N^0) = (\{1, 2, \ldots, n\}|\emptyset)$ and $(\mu^k, b^k, c^k) = (\mu^{(0,k)}, b^{(0,k)}, \delta c_{B^0}^{(0,k)})$.

Step 2 Stop with exit code 0 if any of the following are true:

- $B^j = \emptyset$,
- $(\mu^{(j,k)}, b^{(j,k)}, \delta c_{B^j}^{(j,k)})$ satisfies conditions (1)–(4) of Theorem 5.9, or
- $j \geq 1$ and $\mu^{(j,k)}/\|\delta c_{B^j}^{(j,k)}\| \to \infty$.

Step 3 If we have that $\|\delta c_{B^j}^{(j,k)}\| \neq 0$, $\mu^{(j,k)}/\|\delta c_{B^j}^{(j,k)}\| \downarrow 0$, and that there exists a $\hat{c}_{B^j}^j$ such that $\delta c_{B^j}^{(j,k)}$ is class convergent to $[\hat{c}_{B^j}^j]_{A_{B^j}}$, then continue with Step 4. Otherwise, stop with exit code 1.

Step 4 Solve the linear program

$$(LP^j) \quad \min\left\{\hat{c}_{B^j}^j x_{B^j} : A_{B^j} x_{B^j} = b,\ x_{B^j} \geq 0\right\}$$

and let $(B^{j+1}|N^{j+1})$ be the optimal partition.

Step 5 Set

$$\begin{aligned}
\tau^{(j+1,k)} &= \|\delta c_{B^j}^{(j,k)} - \hat{c}_{B^j}^j\|, \\
\mu^{(j+1,k)} &= \mu^{(j,k)}/\tau^{(j+1,k)}, \\
b^{(j+1,k)} &= b^{(j,k)} - A_{N^{j+1}} x_{N^{j+1}}^j(\mu^{(j,k)}, b^{(j,k)}, \delta c_{B^j}^{(j,k)}), \\
\delta c_{B^j}^{(j+1,k)} &= (1/\tau^{(j+1,k)})(\delta c_{B^j}^{(j,k)} - \hat{c}_{B^j}^j).
\end{aligned}$$

Step 7 Let $j = j + 1$ and go to Step 2.

The technique used in Example 5.1 suggests an algorithmic manner for calculating the limit of $x(\mu^k, b^k, c^k)$. Instead of trying to calculate this limit directly, we instead calculate the limit of c^k and use this limit to form the root problem. The N set of the corresponding optimal partition indexes a collection of variables that must decrease to zero, and in fact, this is the entire collection of zero variables if $\mu^{(1,k)}$ has a positive limit. However, if $\mu^{(1,k)}$ decreases to zero, the variables whose value must be zero are moved to the right-hand side, and the limit of $\delta c_{B^1}^{(1,k)}$ is calculated to form the first subproblem. Again, we know that any variables listed in the corresponding N set of the optimal partition are zero in the limit. The process repeats until either all variables are found to be zero or until $\mu^{(j,k)}$ does not converge to zero for some j.

Example 5.1 has the property that the cost coefficients converge at each step, but the proof of Theorem 5.9 shows that this need not be the case. Instead, at each step of the procedure we need the cost coefficients to be class convergent. As long as this is true, we continue to form the induced sequences until we have a criterion that guarantees either convergence or divergence. The process in Table 5.1 describes how to construct the induced sequences, and Theorem 5.11 shows that $x(\mu^k, b^k, c^k)$ converges if this process terminates with an exit code of 0. In support of this result, Lemma 5.10 extends Lemma 2.6 to allow the class convergence of c^k.

LEMMA 5.10. *Let* $(B|N)$ *be the optimal partition for* $\min\{cx : Ax = b, x \geq 0\}$. *If* c^k *is a nonzero sequence that is class convergent to* $[c]_A$ *and* $\mu^k/\|c^k\| \downarrow 0$, *then* $x_N(\mu^k, b^k, c^k) \downarrow 0$.

Proof. We have from Theorem 5.6 that there is no loss of generality by assuming that c is in null(A). Since $x(\mu^k, b^k, c^k) = x(\mu^k/\|c^k\|, b^k, c^k/\|c^k\|)$ and $c^k/\|c^k\|$ is bounded, we have from Lemma 2.5 that $x(\mu^k, b^k, c^k)$, $y(\mu^k, b^k, c^k)$, and $s(\mu^k, b^k, c^k)$

are bounded. So, there is a subsequence $(\mu^{k_i}, b^{k_i}, c^{k_i})$ such that

$$\begin{aligned} x(\mu^{k_i}, b^{k_i}, c^{k_i}) = x(\mu^{k_i}/\|c^{k_i}\|, b^{k_i}, c^{k_i}/\|c^{k_i}\|) &\to \hat{x}, \\ y(\mu^{k_i}/\|c^{k_i}\|, b^{k_i}, c^{k_i}/\|c^{k_i}\|) &\to \hat{y}, \\ s(\mu^{k_i}/\|c^{k_i}\|, b^{k_i}, c^{k_i}/\|c^{k_i}\|) &\to \hat{s}, \\ c^{k_i}/\|c^{k_i}\| &\to \hat{c}. \end{aligned}$$

For notational ease, we let

$$x^i = x(\mu^{k_i}/\|c^{k_i}\|, b^{k_i}, c^{k_i}/\|c^{k_i}\|), \qquad y^i = y(\mu^{k_i}/\|c^{k_i}\|, b^{k_i}, c^{k_i}/\|c^{k_i}\|),$$
$$\text{and} \qquad s^i = s(\mu^{k_i}/\|c^{k_i}\|, b^{k_i}, c^{k_i}/\|c^{k_i}\|).$$

From the assumption that c^k is class convergent to $[c]_A$, we have that there is a positive α such that $\alpha \operatorname{proj}_{\text{null}(A)} \hat{c} = c$. Since

$$\left.\begin{aligned} Ax^i &= b^{k_i}, \\ y^iA + s^i &= c^{k_i}/\|c^{k_i}\|, \\ s^ix^i &= n\mu^{k_i}/\|c^{k_i}\| \end{aligned}\right\} \Rightarrow \left\{\begin{aligned} A\hat{x} &= b, \\ \hat{y}A + \hat{s} &= \hat{c}, \\ \hat{s}\hat{x} &= 0, \end{aligned}\right.$$

we have that $\hat{x}$ is an optimal solution to $\min\{\hat{c}x : Ax = b, x \geq 0\}$. Let $\tilde{y}$ be such that $\tilde{y}A = \operatorname{proj}_{\text{row}(A)} \hat{c}$, from which we have that $\hat{c} = \operatorname{proj}_{\text{null}(A)} \hat{c} + \operatorname{proj}_{\text{row}(A)} \hat{c} = \operatorname{proj}_{\text{null}(A)} \hat{c} + \tilde{y}A$. Substituting this into $\hat{y}A + \hat{s} = \hat{c}$, we have that

$$A\hat{x} = b, \quad \hat{x} \geq 0, \quad \alpha(\hat{y} - \tilde{y})A + \alpha\hat{s} = \alpha \operatorname{proj}_{\text{null}(A)} \hat{c} = c, \quad \alpha\hat{s} \geq 0, \quad \text{and} \quad \hat{s}\hat{x} = 0.$$

Hence, $\hat{x}$ is also an optimal solution to $\min\{cx : Ax = b, x \geq 0\}$, which implies that $x_N(\mu^k, b^k, c^k) \to \hat{x}_N = 0$. ∎

THEOREM 5.11. *If the process in Table* 5.1 *stops with an exit code of* 0*, then* $x(\mu^k, b^k, c^k)$ *converges.*

Proof. If the process terminates with $j = 0$ and an exit code of 0, then we have that $(\mu^k, b^k, c^k) = (\mu^{(0,k)}, b^{(0,k)}, \delta c_{B^0}^{(0,k)})$ satisfies conditions 1–4 of Theorem 5.9, which implies that $x(\mu^k, b^k, c^k)$ converges. Suppose that the process in Table 5.1 terminates with an exit code of 0 and that the induced sequences are $(\mu^{(j,k)}, b^{(j,k)}, \delta c_{B^{j-1}}^{(j,k)})$ for $j = 1, 2, \ldots, J$. The proof follows with a careful inspection of how the sequence $x(\mu^k, b^k, c^k)$ partitions itself as the process continues. From the definition of the first induced sequence, we have that

$$x(\mu^k, b^k, c^k) = x_{B^0}\left(\mu^{(0,k)}, b^{(0,k)}, \delta c_{B^0}^{(0,k)}\right) = \left(\begin{array}{c} x_{B^1}(\mu^{(0,k)}, b^{(0,k)}, \hat{c}_{B^0}^0 + \tau^{(1,k)}\delta c_{B^0}^{(1,k)}) \\ \hline x_{N^1}(\mu^{(0,k)}, b^{(0,k)}, \delta c_{B^0}^{(0,k)}) \end{array}\right).$$

From (3.4) we have that

$$x_{B^1}(\mu^{(0,k)}, b^{(0,k)}, \hat{c}_{B^0}^0 + \tau^{(1,k)}\delta c_{B^0}^{(1,k)}) = x_{B^1}(\mu^{(1,k)}, b^{(1,k)}, \delta c_{B^1}^{(1,k)}).$$

Using the second induced sequence, we have that

$$x_{B^1}\left(\mu^{(1,k)}, b^{(1,k)}, \delta c_{B^1}^{(1,k)}\right) = \left(\begin{array}{c} x_{B^2}(\mu^{(1,k)}, b^{(1,k)}, \hat{c}_{B^1}^1 + \tau^{(2,k)}\delta c_{B^1}^{(2,k)}) \\ \hline x_{N^2}(\mu^{(1,k)}, b^{(1,k)}, \delta c_{B^1}^{(1,k)}) \end{array}\right),$$

which implies that

$$x(\mu^k, b^k, c^k) = \begin{pmatrix} x_{B^2}(\mu^{(1,k)}, b^{(1,k)}, \hat{c}^1_{B^1} + \tau^{(2,k)}\delta c^{(2,k)}_{B^1}) \\ \hline x_{N^2}(\mu^{(1,k)}, b^{(1,k)}, \delta c^{(1,k)}_{B^1}) \\ \hline x_{N^1}(\mu^{(0,k)}, b^{(0,k)}, \delta c^{(0,k)}_{B^0}) \end{pmatrix}.$$

Again, from (3.4) and the definition the third induced sequence, we have that

$$x_{B^2}\big(\mu^{(1,k)}, b^{(1,k)}, \hat{c}^1_{B^1} + \tau^{(2,k)}\delta c^{(2,k)}_{B^1}\big) = x_{B^2}\big(\mu^{(2,k)}, b^{(2,k)}, \delta c^{(2,k)}_{B^2}\big)$$
$$= \begin{pmatrix} x_{B^3}(\mu^{(2,k)}, b^{(2,k)}, \hat{c}^2_{B^2} + \tau^{(3,k)}\delta c^{(3,k)}_{B^2}) \\ \hline x_{N^3}(\mu^{(2,k)}, b^{(2,k)}, \delta c^{(2,k)}_{B^2}) \end{pmatrix}.$$

The process continues until

$$x(\mu^k, b^k, c^k) = \begin{pmatrix} x_{B^J}(\mu^{(J,k)}, b^{(J,k)}, \delta c^{(J,k)}_{B^J}) \\ \hline x_{N^J}(\mu^{(J-1,k)}, b^{(J-1,k)}, \delta c^{(J-1,k)}_{B^{J-1}}) \\ \hline \vdots \\ \hline x_{N^2}(\mu^{(1,k)}, b^{(1,k)}, \delta c^{(1,k)}_{B^1}) \\ \hline x_{N^1}(\mu^{(0,k)}, b^{(0,k)}, \delta c^{(0,k)}_{B^0}) \end{pmatrix}.$$

The fact that the first induced sequence was created implies that $\delta c^{(0,k)}_{B^0} \neq 0$, $\mu^{(0,k)}/\|\delta c_{B^0}\| \downarrow 0$, and $\delta c^{(0,k)}_{B^0}$ is class convergent to $[\hat{c}^0_{B^0}]_{A_{B^0}}$. By assumption we have that $b^{(0,k)} = b^k \to b$. So, from Lemma 5.10 we have that $x^0_{N^1}(\mu^{(0,k)}, b^{(0,k)}, \delta c^{(0,k)}_{B^0}) \downarrow 0$, which subsequently implies that $b^{(1,k)} = b^{(0,k)} - A_{N^1}x^0_{N^1}(\mu^{(0,k)}, b^{(0,k)}, \delta c^{(0,k)}_{B^0}) \to b$. By the same logic, and repeated applications of Lemma 5.10, we find that

$$\begin{pmatrix} x_{N^J}(\mu^{(J-1,k)}, b^{(J-1,k)}, \delta c^{(J-1,k)}_{B^{J-1}}) \\ \hline \vdots \\ \hline x_{N^2}(\mu^{(1,k)}, b^{(1,k)}, \delta c^{(1,k)}_{B^1}) \\ \hline x_{N^1}(\mu^{(0,k)}, b^{(0,k)}, \delta c^{(0,k)}_{B^0}) \end{pmatrix} \downarrow 0,$$

which subsequently implies that $b^{(j,k)} \to b$ for $j = 1, 2, \ldots, J$. At this point we have that if the process terminated because $B^J = \emptyset$, then $x(\mu^k, b^k, c^k) \downarrow 0$. Suppose that $(\mu^{(J,k)}, b^{(J,k)}, \delta c_{B^J}^{(J,k)})$ satisfies conditions 1–4 of Theorem 5.9; then we have that $x_{B^J}^J(\mu^{(J,k)}, b^{(J,k)}, \delta c_{B^J}^{(J,k)})$ converges, and hence so does $x(\mu^k, b^k, c^k)$.

Suppose that $\mu^{(J,k)}/\|\delta c_{B^J}^{(J,k)}\| \to \infty$, which subsequently implies that $\delta c_{B^J}^{(J,k)}/\mu^{(J,k)} \to 0 \in \text{row}(A)$. The set $\{x_{B^{J-1}} : A_{B^J} x_{B^J} = b, x_{B^J} \geq 0, x_{N^J} = 0\}$ is bounded because it is the optimal set of (LP^{J-1}). So, from Theorem 2.9 we have that $x_{B^J}(\mu^{(J,k)}, b^{(J,k)}, \delta c_{B^J}^{(J,k)})$ converges. □

We conclude by pointing out that $x(\mu^k, b^k, c^k)$ can converge if the process in Table 5.1 terminates with an exit code of 1. As an example, let $b^k = 1$, $\mu^k = 1/k$, $A = [1, 1]$, and c^k be $(1, 1)$ if k is even and $(1/k, 1/k)$ if k is odd. Then, for all k we have that $c^k \in \text{row}(A)$, and from Theorem 2.3 we know that $x(\mu^k, b^k, c^k) = \bar{x}(b) = (1/2, 1/2)^T$. However, $\mu^k/\|c^k\|$ is $1/(k\sqrt{2})$ if k is even and is $1/\sqrt{2}$ if k is odd. Hence, the sequence $\mu^k/\|c^k\|$ does not decrease to zero and is not $\Theta(1)$, and the process terminates with an exit code of 1.

6. Conclusions and future research. We have accomplished three goals in this paper. First, we have completely characterized the convergence of the perturbed central path followed by many infeasible-interior-point methods. This result is succinctly depicted in Table 3.1. Second, we have shown that the perturbed central path converges as a set as long as the cost vector perturbation is linear. In fact, the central path is continuous over the set of cost directions for which the optimal partition is invariant. Third, we provided sufficient conditions for the perturbed central path to converge under arbitrary, simultaneous changes in b and c. These are the first results in the literature that deal with this complicated situation; however, characterizing the convergence under such data perturbations remains an open question.

Acknowledgments. The author is grateful to an anonymous referee for meticulously reading an earlier draft of this work.

REFERENCES

[1] J. Bonnans and F. Potra, *On the convergence of the iteration sequence of infeasible path following algorithms for linear complementarity problems*, Math. Oper. Res., 22 (1997), pp. 378–407.

[2] R. Caron, H. Greenberg, and A. Holder, *Analytic centers and repelling inequalities*, European J. Oper. Res., 143 (2002), pp. 268–290.

[3] A. Fiacco and G. McCormick, *Nonlinear Programming: Sequential Unconstrained Minimization Techniques*, John Wiley, New York, 1968.

[4] K. Frisch, *The Logarithmic Potential Method of Convex Programming*, Technical report, University Institute of Economics, Oslo, Norway, 1955.

[5] H. Greenberg, *Mathematical Programming Glossary*, http://carbon.cudenver.edu/~hgreenbe/glossary/glossary.index.php (1996–2001).

[6] H. J. Greenberg, *Simultaneous primal-dual right-hand-side sensitivity analysis from a strictly complementary solution of a linear program*, SIAM J. Optim., 10 (2000), pp. 427–442.

[7] H. J. Greenberg, A. G. Holder, K. Roos, and T. Terlaky, *On the dimension of the set of rim perturbations for optimal partition invariance*, SIAM J. Optim., 9 (1998), pp. 207–216.

[8] B. Grunbaum, *Convex Polytopes*, Wiley-Interscience, New York, 1967.

[9] M. Halická, E. de Klerk, and C. Roos, *On the convergence of the central path in semidefinite optimization*, SIAM J. Optim., 12 (2002), pp. 1090–1099.

[10] A. Holder, J. Sturm, and S. Zhang, *Marginal and parametric analysis of the central optimal solution*, Information Syst. Operational Res., 39 (2001), pp. 394–415.

[11] P. Huard, *Resolution of mathematical programming with nonlinear constraints by the method of centres*, in Nonlinear Programming, J. Abadie, ed., John Wiley, New York, 1967, pp. 209–219.

[12] L. McLinden, *An analogue of Moreau's approximation theorem, with applications to the nonlinear complementarity problem*, Pacific J. Math., 88 (1980), pp. 101–161.
[13] N. Megiddo, *Pathways to the optimal set in linear programming*, in Progress in Mathematical Programming: Interior-Point Algorithms and Related Methods, N. Megiddo, ed., Springer-Verlag, New York, 1989, pp. 131–158.
[14] S. Mizuno, M. Todd, and Y. Ye, *A surface of analytic centers and primal-dual infeasible-interior point algorithms for linear programming*, Math. Oper. Res., 20 (1995), pp. 135–162.
[15] R. Monteiro and T. Tsuchiya, *Limiting behavior of the derivatives of certain trajectories associated with a monotone horizontal linear complementarity problem*, Math. Oper. Res., 21 (1996), pp. 793–814.
[16] Y. Nesterov and A. Nemirovski, *Multi-parameter surfaces of analytic centers and long-step surface-following interior point methods*, Math. Oper. Res., 23 (1998), pp. 1–38.
[17] C. Papadimitriou and K. Steiglitz, *Combinatorial Optimization: Algorithms and Complexity*, Prentice–Hall, Englewood Cliffs, NJ, 1982.
[18] C. Roos, T. Terlaky, and J.-P. Vial, *Theory and Algorithms for Linear Optimization: An Interior Point Approach*, John Wiley, New York, 1997.
[19] G. Sonnevend, *An "analytic centre" for polyhedrons and new classes of global algorithms for linear (smooth, convex) programming*, in System Modelling and Optimization, Lecture Notes in Control and Inform. Sci. 84, A. Prekopa, J. Szelezsan, and B. Straziсky, eds., Springer-Verlag, Heidelberg, 1986, pp. 866–875.
[20] G. Sonnevend, *An implementation of the method of analytic centers*, in Analysis and Optimization Systems, Lecture Notes in Control and Inform. Sci. 111, A. Bensoussan and J. Lions, eds., Springer-Verlag, Heidelberg, 1988, pp. 297–308.
[21] G. Sonnevend, *New algorithms in convex programming based on a notion of "centre" (for systems of analytic inequalities) and on rational extrapolation*, in Trends in Mathematical Optimization, Internat. Schriftenreihe Numer. Math. 84, K. Hoffman et al., eds., Birkhauser-Verlag, Basel, 1988, pp. 311–326.
[22] G. Sonnevend, *Applications of the notion of analytic center in approximation (estimation) problems*, J. Comput. Appl. Math., 28 (1989), pp. 349–358.
[23] G. Sonnevend and J. Stoer, *Global ellipsoidal approximations and homotopy methods for solving convex analytic programs*, Appl. Math. Optim., 21 (1990), pp. 139–165.
[24] G. Sonnevend, J. Stoer, and G. Zhao, *On the complexity of following the central path of linear programs by linear extrapolation* II, Math. Programming, 52 (1991), pp. 527–553.
[25] M. Wright, *The Interior-Point Revolution in Constrained Optimization*, Appl. Optim. 24, Kluwer, Dordrecht, The Netherlands, 1998, pp. 359–381.
[26] S. J. Wright, *Primal-Dual Interior-Point Methods*, SIAM, Philadelphia, 1997.
[27] Y. Ye, *Interior Point Algorithms Theory and Analysis*, John Wiley, New York, 1997.
[28] A. Yildirim, *An interior-point perspective on sensitivity analysis in semidefinite programming*, Math. Oper. Res., 28 (2003), pp. 649–676.
[29] A. Yildirim and M. Todd, *Sensitivity analysis in linear programming and semidefinite programming using interior-point methods*, Math. Program., 90 (2001), pp. 229–261.
[30] E. A. Yildirim and M. J. Todd, *An interior-point approach to sensitivity analysis in degenerate linear programs*, SIAM J. Optim., 12 (2002), pp. 692–714.

SIAM J. OPTIM.
Vol. 14, No. 3, pp. 869–893

A BUNDLE METHOD FOR SOLVING VARIATIONAL INEQUALITIES*

GENEVIÈVE SALMON†, JEAN-JACQUES STRODIOT†, AND VAN HIEN NGUYEN†

Abstract. In this paper, we present a bundle method for solving a generalized variational inequality problem. This problem consists of finding a zero of the sum of two multivalued operators defined on a real Hilbert space. The first one, F, is monotone and the second is the subdifferential of a lower semicontinuous proper convex function. Our method is based on the auxiliary problem principle due to Cohen, and our strategy is to approximate, in the subproblems, the nonsmooth convex function by a sequence of convex piecewise linear functions, as in the bundle method for nonsmooth optimization. This makes the subproblems more tractable. First, we explain how to build, step by step, suitable piecewise linear approximations by means of a bundle strategy, and we present a new stopping criterion to determine whether the current approximation is good enough. This criterion is the same as that commonly used in the special case of nonsmooth optimization. Second, we study the convergence of the algorithm for the case when the stepsizes are chosen going to zero and for the case bounded away from zero. In the first case, the convergence can be proved under rather mild assumptions: the operator F is paramonotone and possibly multivalued. In the second case, the convergence needs a stronger assumption: F is single-valued and satisfies a Dunn property. Finally, we illustrate the behavior of the proposed algorithm by some numerical tests.

Key words. generalized variational inequality, multivalued mapping, auxiliary problem principle, bundle method, gap functions, paramonotone operator, Dunn property

AMS subject classifications. 65K05, 90C25

DOI. 10.1137/S1052623401384096

1. Introduction. Let F be a monotone multivalued operator defined on a real Hilbert space H with inner product $\langle\cdot,\cdot\rangle$, let C be a nonempty closed convex subset of H, and let $\varphi : H \to \mathbb{R} \cup \{+\infty\}$ be a lower semicontinuous (l.s.c.) proper convex function. We consider the following general variational inequality problem:

$$(P) \qquad \begin{cases} \text{Find } x^* \in C \text{ and } r(x^*) \in F(x^*) \text{ such that, for all } x \in C, \\ \\ \langle r(x^*), x - x^* \rangle + \varphi(x) - \varphi(x^*) \geq 0. \end{cases}$$

In this paper, we assume that $C \subseteq \text{int}(\text{dom } \varphi)$. Moreover, we suppose that there exists at least one solution to this problem. Existence results for problem (P) can be found, for example, in [3, 13, 15].

This problem can also be expressed in an inclusion form as follows: Find x^* such that $0 \in F(x^*) + \partial(\varphi + \psi_C)(x^*)$, where ψ_C denotes the indicator function associated with C (i.e., $\psi_C(x) = 0$ if $x \in C$ and $+\infty$ otherwise) and $\partial(\varphi + \psi_C)(x^*)$ denotes the subdifferential of the convex function $\varphi + \psi_C$ at x^*. So, problem (P) is a particular case of the problem that consists of finding a zero of the sum of two operators.

A large variety of problems can be seen as special instances of problem (P). For example, when F is the subdifferential of a finite-valued convex continuous function

*Received by the editors January 24, 2001; accepted for publication (in revised form) July 9, 2003; published electronically March 23, 2004. This work was supported in part by a grant from the Belgian Fonds National de la Recherche Scientifique (FNRS: B8/5-CB/MF-4.515 and B8/5-CB/SP-9.579).

http://www.siam.org/journals/siopt/14-3/38409.html

†Unité d'Optimisation, Département de Mathématique, Facultés Universitaires Notre Dame de la Paix, Namur, Belgium (genevieve.salmon@fundp.ac.be, jean-jacques.strodiot@fundp.ac.be, vhnguyen@undp.ac.be).

f defined on H, problem (P) reduces to the nondifferentiable convex optimization problem

$$(OP) \qquad \min_{x\in C}\{f(x)+\varphi(x)\}.$$

On the other hand, in the particular case where F is single-valued and $\varphi = 0$, problem (P) reduces to the following classical variational inequality problem:

$$(VIP) \qquad \begin{cases} \text{Find } x^* \in C \text{ such that, for all } x \in C, \\ \\ \langle F(x^*), x - x^*\rangle \geq 0. \end{cases}$$

Important research has been devoted to finding the solution to problem (VIP) (see, for example, [12, 15, 16, 17, 19, 20, 24, 30]). However, variational inequalities with a multivalued mapping F and a function $\varphi \neq 0$ are encountered in many applications, in particular, in mechanical problems (see, e.g., [29]) and equilibrium problems (see, e.g., [9, 21, 28]). So it is worth studying implementable methods for solving such problems. That is the purpose of this paper.

Algorithms that can be applied for solving problem (P) or one of its variants are very numerous. For the case when F is maximal monotone, the most famous method is the proximal method (see, e.g., [14, 26, 34, 35]) which consists of finding a zero of the operator $F + \partial(\varphi + \psi_C)$ by using the scheme

$$x^{k+1} = [I + \mu_k(F + \partial(\varphi + \psi_C))]^{-1}(x^k), \tag{1.1}$$

where $\{\mu_k\}_{k\in\mathbb{N}}$ is a sequence of positive real numbers. Splitting methods have also been studied to solve problem (P). Here the multivalued operators F and $\partial(\varphi + \psi_C)$ play separate roles. The simplest splitting method is the forward–backward scheme (see, e.g., [40]), whose iteration is given by

$$x^{k+1} \in [I + \mu_k\partial(\varphi + \psi_C)]^{-1}[I - \mu_k F](x^k), \tag{1.2}$$

where $\{\mu_k\}_{k\in\mathbb{N}}$ is a sequence of positive real numbers. When $\varphi = 0$, we obtain a projection method in the following sense: First, one element $r(x^k)$ is computed in $F(x^k)$ and then the vector $x^k - \mu_k r(x^k)$ is projected onto the closed convex set C.

Cohen developed in [10] a general algorithmic framework for solving problem (P), based on the so-called auxiliary problem principle. The corresponding method is a generalization of the forward–backward method. More precisely, let Ω be a strongly monotone and Lipschitz continuous auxiliary operator on H, and let $\{\mu_k\}_{k\in\mathbb{N}}$ be a sequence of positive real numbers. The problem considered at iteration k is the following:

$$x^{k+1} \in [\Omega + \mu_k\partial(\varphi + \psi_C)]^{-1}[\Omega - \mu_k F](x^k),$$

i.e.,

$$\begin{cases} \text{choose } r(x^k) \in F(x^k) \text{ and find } x^{k+1} \in C \text{ such that, for all } x \in C, \\ \\ \langle r(x^k) + \mu_k^{-1}[\Omega(x^{k+1}) - \Omega(x^k)], x - x^{k+1}\rangle + \varphi(x) - \varphi(x^{k+1}) \geq 0. \end{cases}$$

In this paper, Ω is chosen as the gradient of some continuously differentiable and strongly convex function h with Lipschitz continuous gradient. In that case, the

subproblem can also be equivalently written in the following minimization form:

$$(AP^k) \quad \begin{cases} x^{k+1} = \operatorname{argmin}_{x\in C}\{\varphi(x) + \langle r(x^k), x - x^k\rangle \\ \qquad\qquad\qquad\qquad +\mu_k^{-1}[h(x) - h(x^k) - \langle \nabla h(x^k), x - x^k\rangle]\}, \\ \text{with } r(x^k) \in F(x^k). \end{cases}$$

The assumptions imposed on the function h ensure that this problem has one and only one solution.

The convergence of the sequence $\{x^k\}_{k\in\mathbb{N}}$ generated by solving subproblems (AP^k) was first studied in the literature for the case when the sequence of stepsizes $\{\mu_k\}_{k\in\mathbb{N}}$ is bounded away from zero and later for the case when this sequence converges to zero. In the first situation, the convergence was obtained for the case when F is single-valued and F is required either to be strongly monotone (see, e.g., [10]) or to satisfy the (pseudo) Dunn property (see, e.g., [27, 36, 42]). In the second situation for the case when the sequence $\{\mu_k\}_{k\in\mathbb{N}}$ converges to zero Cohen proved in [10] the strong convergence of the scheme for the case when F is multivalued and strongly monotone. More recently, in [41] Zhu obtained convergence results under weaker monotonicity assumptions. He proves weak convergence under a condition satisfied, for example, if the operator is either paramonotone and compact-valued or is the subdifferential of an l.s.c. proper convex function.

When φ is a nonsmooth convex function, subproblems (AP^k) may be very hard to solve. Several authors proposed approximating the function φ by a sequence of more tractable convex functions; see, e.g., [22, 25, 37, 38, 39].

When $F = 0$ and $C = H$, problem (P) reduces to minimizing the nondifferentiable convex function φ on H. This problem can be solved by the so-called bundle method introduced in the 1980s by Correa and Lemaréchal [11]. In this method, the effective domain of φ is supposed to be the whole space H, and the strategy is to approximate the function φ, at the proximal iteration k, by a piecewise linear convex function, built step by step, and to move to the next iterate only when the approximation is suitable. This gives the following algorithm.

Bundle algorithm to minimize φ on H. Let an initial point x^0 be given, together with a tolerance $m \in]0, 1[$ and a positive sequence $\{\mu_k\}_{k\in\mathbb{N}}$. Set $y^0 = x^0$ and $k = 0, i = 1$.

Step 1. Choose a piecewise linear convex function $\theta^i \le \varphi$ and solve

$$\min_{x\in H}\{\ \theta^i(x) + (2\mu_k)^{-1}\|x - x^k\|^2\} \tag{1.3}$$

to obtain the unique optimal solution y^i.

Step 2. If the decrease is sufficient, i.e., if

$$\varphi(x^k) - \varphi(y^i) \ge m\,[\varphi(x^k) - \theta^i(y^i)], \tag{1.4}$$

then set $x^{k+1} = y^i$ and increase k by 1.

Step 3. Increase i by 1 and go to Step 1.

Let x^k be the current outer iterate. The solution y^i of subproblem (1.3) with the current approximation θ^i is called a trial point. If the decrease between $\varphi(x^k)$ and $\varphi(y^i)$ is sufficient in the sense that the stopping test (1.4) is verified, then the current outer iterate x^k is updated and the resulting step is called a *serious-step*. Otherwise, the iterate x^k is kept fixed for the next inner iteration. We say that a *null-step* has

been made, and the new trial point y^i will be used to improve the next approximation θ^{i+1} of φ. As proven in [11], this method can be seen as a practical implementation of the classical proximal method in convex optimization. Our purpose in this paper is to use these ideas to solve problem (P). In our case, the subproblem to solve at Step 1 will be (AP^k) with φ replaced by some approximation θ^i, and the stopping test in Step 2 will be adapted to take into account the contribution of the operator F.

So, we will first show how to build, step by step, suitable piecewise linear approximations $\theta^i \leq \varphi$ by means of a bundle strategy and how to adapt the stopping criterion. Then we study the convergence of the resulting algorithm by separating the case where the stepsizes go to zero from the case where they are bounded away from zero. When the stepsizes go to zero (but not too fast), we prove that the sequence $\{x^k\}_{k\in\mathbb{N}}$ generated by the algorithm is bounded and that each weak limit point of this sequence is a solution of problem (P) if F is paramonotone, weakly closed on C, and Lipschitz continuous on bounded subsets of C; or if F is the subdifferential of a convex continuous function and is bounded on bounded subsets of C; or if F is strongly monotone on C and bounded on bounded subsets of C. It results in a very general convergence theorem, not only for the weak limit points of $\{x^k\}_{k\in\mathbb{N}}$ but also for the weak convergence (where, in addition, ∇h is weakly continuous) and the strong convergence (where F is strongly monotone) of the sequence $\{x^k\}_{k\in\mathbb{N}}$. When $\varphi = 0$, our results generalize those obtained by Zhu in [41]. Note that if we take $F = 0$ and $C = H$, our scheme reduces to the classical bundle method for optimization. This method is known to have a slow convergence rate when the stepsizes converge to zero. For this reason, we study separately the case where the stepsizes are bounded away from zero. In that case, we have to impose stronger assumptions on F to get convergence: F is restricted to be single-valued and to satisfy the (pseudo) Dunn property. Consequently, these last results can be applied for $F = 0$, and the classical convergence results for the optimization case can be recovered.

Other contributions to the construction of bundle methods for monotone variational inequalities (or for the equivalent problem of finding zeroes of monotone point-to-set operators) have appeared in the literature. For instance, in [6], a bundle method is presented for finding a zero of a maximal monotone operator T defined on H. This method is based on the paper [5] by the same authors, where an ε-enlargement of the operator T is defined. The main difference between the two methods is that our method takes into account the special structure of $T = F + \partial(\varphi + \psi_C)$ by using the bundle technique not on the operator T but directly on the function φ.

The paper is organized as follows. In section 2, we specify the bundle scheme proposed for solving problem (P) and we prove that if only null-steps are made after some x^k has been reached, then x^k actually solves problem (P). In sections 3 and 4, we suppose that the bundle algorithm generates an infinite sequence $\{x^k\}_{k\in\mathbb{N}}$ and we prove the boundedness of $\{x^k\}_{k\in\mathbb{N}}$ and the weak and strong convergence of this sequence to a solution of problem (P). Section 3 is devoted to the case where the stepsizes go to zero and the operator F is possibly multivalued, while section 4 deals with stepsizes bounded away from zero and a single-valued operator F. Finally, in section 5 we present the results of some numerical tests designed to illustrate the behavior of the bundle algorithm. Throughout this paper, we denote by $\Gamma_0(H)$ the set of l.s.c. proper convex functions from H into $\mathbb{R} \cup \{+\infty\}$. Any other undefined term or usage should be taken as in the books [3] and [33].

2. Bundle strategy. The bundle algorithm designed to minimize φ on H can be adapted for solving problem (P) in the following way.

BUNDLE ALGORITHM FOR SOLVING PROBLEM (P). Let an initial point x^0 be given, together with a tolerance $m \in]0,1[$ and a positive sequence $\{\mu_k\}_{k\in\mathbb{N}}$. Compute $r(x^0) \in F(x^0)$. Set $y^0 = x^0, k = 0, i = 1$.

Step 1. Choose a piecewise linear convex function $\theta^i \leq \varphi$ and solve

$$(P_i^k) \quad \min_{x\in C}\{\theta^i(x) + \langle r(x^k), x - x^k\rangle + \mu_k^{-1}[h(x) - h(x^k) - \langle \nabla h(x^k), x - x^k\rangle]\},$$

to obtain the unique optimal solution $y^i \in C$.

Step 2. If the trial point y^i is suitable, i.e., if

$$\varphi(x^k) - \varphi(y^i) \geq m\,[\varphi(x^k) - \theta^i(y^i)] + (1-m)\,\langle r(x^k), y^i - x^k\rangle, \tag{2.1}$$

then set $x^{k+1} = y^i$, compute $r(x^{k+1}) \in F(x^{k+1})$, and increase k by 1.

Step 3. Increase i by 1 and go to Step 1.

Each trial point $y^i \in C$ is obtained by solving the approximate auxiliary subproblem (P_i^k), namely, subproblem (AP^k), with the function φ replaced by the approximation $\theta^i \leq \varphi$. This approximation is suitable if the stopping criterion (2.1) is satisfied. This criterion is obtained from (1.4) by comparing the optimization case with the variational inequality case. In the optimization case, the proximal iteration is approximated to obtain subproblem (1.3), while in the variational inequality case, subproblem (AP^k) is approximated to obtain (P_i^k). By comparing these two situations, we observe that we pass from the optimization case to the variational inequality case by replacing the functions φ and θ^i with $\varphi + \langle r(x^k), \cdot - x^k\rangle$ and $\theta^i + \langle r(x^k), \cdot - x^k\rangle$, respectively. If these updates are set in criterion (1.4), we obtain the new criterion (2.1). So, when the stopping criterion holds, the *outer* iterate x^k is updated and we say that a *serious-step* is made. Otherwise, x^k is kept fixed for the next *inner* iteration, which will be performed with an improvement of the approximation θ^i. This step is called a *null-step*. In what follows, we call i_k the *inner* iteration that has produced x^k (with $i_0 = 0$).

In order to prove the convergence of this algorithm, we have to impose conditions on the functions $\theta^i, i = 1, 2, \ldots$. Before presenting these conditions, first we observe that, by optimality of $y^i \in C$, we have

$$\gamma^i \equiv \mu_k^{-1}[\nabla h(x^k) - \nabla h(y^i)] - r(x^k) \in \partial[\theta^i + \psi_C](y^i). \tag{2.2}$$

Then we define the aggregate affine function l^i by

$$l^i(y) = \theta^i(y^i) + \langle \gamma^i, y - y^i\rangle, \quad y \in C. \tag{2.3}$$

We have $l^i(y^i) = \theta^i(y^i)$ and, using (2.2) and (2.3),

$$l^i(y) \leq \theta^i(y) \quad \text{for all } y \in C. \tag{2.4}$$

Now we require the following conditions on the functions θ^i:

(C1) $\theta^i \leq \varphi$ on C for all $i = 1, 2, \ldots$,

(C2) $l^i \leq \theta^{i+1}$ on C for all $i \in]i_k, i_{k+1}[$,

(C3) $\varphi(y^i) + \langle s(y^i), \cdot - y^i\rangle \leq \theta^{i+1}$ for all $i \in]i_k, i_{k+1}[$,

(C3) $\varphi(y^{i_k}) + \langle s(y^{i_k}), \cdot - y^{i_k}\rangle \leq \theta^i$ for all $i \in]i_k, i_{k+1}]$,

where $s(y^i)$ denotes a subgradient of φ at y^i. Here we suppose that, at each point of C, one subgradient of φ is available. Note that φ admits a subgradient at each point of C since $C \subseteq \operatorname{int}(\operatorname{dom}\varphi)$ (see [3, Chapter 4, section 3, Theorem 17]).

The first three conditions are similar to those introduced in [11] within the framework of nonsmooth convex optimization. Condition (C4) will be used in the next section to show the weak convergence of the sequence $\{x^k\}_{k\in\mathbb{N}}$ generated by the bundle algorithm when the operator F is multivalued.

Let us now mention a few examples of functions θ^i satisfying (C1)–(C4). For the first function, we can take $\theta^1 = \varphi(y^0) + \langle s(y^0), \cdot - y^0\rangle$, and for $i = 1, 2, \ldots$, we can choose

$$\theta^{i+1} = \max_{0\le j\le i}\{\varphi(y^j) + \langle s(y^j), \cdot - y^j\rangle\}. \tag{2.5}$$

It is easy to see that (C1), (C3), and (C4) are satisfied. Since $\theta^i \le \theta^{i+1}$, (C2) follows from (2.4). Other choices are possible, e.g., for all $i \ge i_k$,

$$\theta^{i+1} = \max_{j\in\{i_k,i\}}\{\theta^i(y^i) + \langle \gamma^i, \cdot - y^i\rangle, \varphi(y^j) + \langle s(y^j), \cdot - y^j\rangle\}. \tag{2.6}$$

Indeed, (C2), (C3), and (C4) are obvious, and (C1) is satisfied because $\gamma^i \in \partial(\theta^i + \Psi_C)(y^i)$ and $s(y^i) \in \partial\varphi(y^i)$.

In what follows we will also need to consider the following functions:

$$\tilde{l}^i(y) = l^i(y) + \langle r(x^k), y - x^k\rangle + \mu_k^{-1}\ [\ h(y) - h(x^k) - \langle \nabla h(x^k), y - x^k\rangle\ ],$$
$$\tilde{\theta}^i(y) = \theta^i(y) + \langle r(x^k), y - x^k\rangle + \mu_k^{-1}\ [\ h(y) - h(x^k) - \langle \nabla h(x^k), y - x^k\rangle\ ].$$

Using (2.2) and (2.3), it is easy to see that, for all $y \in C$,

$$\tilde{l}^i(y) = \tilde{l}^i(y^i) + \mu_k^{-1}\ [\ h(y) - h(y^i) - \langle \nabla h(y^i), y - y^i\rangle\ ]. \tag{2.7}$$

Moreover, we have

$$\tilde{\theta}^i(x^k) = \theta^i(x^k) \quad \text{and} \quad \tilde{l}^i(y^i) = \tilde{\theta}^i(y^i), \tag{2.8}$$

and, by condition (C2),

$$\tilde{l}^i \le \tilde{\theta}^{i+1} \text{ on } C. \tag{2.9}$$

We can now study the convergence of the bundle algorithm. In what follows, we will assume that the following conditions hold.

Assumption A.

- Problem (P) admits at least one solution;
- F is a monotone operator defined on H;
- $\varphi \in \Gamma_0(H)$;
- C is a nonempty closed convex subset of H such that $C \subseteq \text{int}(\text{dom } \varphi)$;
- $\partial\varphi$ is bounded on bounded subsets of C;
- $h : H \to \mathbb{R}$ is continuously differentiable and strongly convex over C with modulus $\beta > 0$, and its gradient ∇h is Lipschitz continuous over C with modulus $\Lambda > 0$;
- the sequence $\{\theta^i\}_{i\in\mathbb{N}_0}$ satisfies conditions (C1)–(C3).

Remark 1. Since $\varphi \in \Gamma_0(H)$, φ is also weakly l.s.c. over H (see [15, Chapter 1, Corollary 2.2]) and continuous over int(dom φ) (see [15, Chapter 1, Corollary 2.5]).

Remark 2. We know that a monotone mapping is locally bounded at interior points of its domain (see [31, Chapter 3, section 2.2]). Since int(dom φ) = int(dom $\partial\varphi$), we deduce that $\partial\varphi$ is locally bounded at any point of int(dom φ) (see [31,

Chapter 1, section 2.6]). Hence, when H is finite dimensional, $\partial\varphi$ is always bounded on bounded subsets of C. This is not necessarily the case in a general Hilbert space. However, a sufficient condition for $\partial\varphi$ to be bounded on bounded subsets of C is that $|\varphi|$ be bounded on bounded subsets of C (see [1, p. 3]).

PROPOSITION 2.1. *Suppose that Assumption* A *holds. If the stopping test is suppressed in the bundle algorithm after some outer iterate x^k has been reached, then $[\varphi(y^i) - \theta^i(y^i)] \to 0$ and $y^i \to z(x^k)$, where $z(x^k)$ denotes the unique solution of problem (AP^k), i.e., $z(x^k) = \operatorname{argmin}_{x\in C}\{\varphi(x) + \langle r(x^k), x - x^k\rangle + \mu_k^{-1}[\, h(x) - h(x^k) - \langle \nabla h(x^k), x - x^k\rangle\,]\}$.*

Proof. Since i_k denotes the *inner* iteration that has produced x^k, and only null-steps are made after reaching x^k, all the inequalities below have to be understood for $i > i_k$, i.e., for i large enough.

First, in order to show that $\varphi(y^i) - \theta^i(y^i) \to 0$, we proceed in three steps.

1. The sequence $\{\tilde{l}^i(y^i)\}_{i\in\mathbb{N}}$ is convergent and $[y^{i+1} - y^i] \to 0$. For all i we have

$$\begin{aligned}
\varphi(x^k) &\geq \theta^{i+1}(x^k) && \text{(by } (C1)\text{)}\\
&= \tilde{\theta}^{i+1}(x^k) && \text{(by (2.8))}\\
&\geq \tilde{\theta}^{i+1}(y^{i+1}) && \text{(by definition of } y^{i+1}\text{)}\\
&= \tilde{l}^{i+1}(y^{i+1}) && \text{(by (2.8))}\\
&\geq \tilde{l}^{i}(y^{i+1}) && \text{(by (2.9))}\\
&= \tilde{l}^{i}(y^{i}) + \mu_k^{-1}\, D_h(y^{i+1}, y^i) && \text{(by (2.7))}\\
&\geq \tilde{l}^{i}(y^{i}) + (2\mu_k)^{-1}\, \beta\, \|y^{i+1} - y^i\|^2 && \text{(since } h \text{ is strongly convex } (\beta)\text{)}\\
&\geq \tilde{l}^{i}(y^{i}), &&
\end{aligned}$$

where $D_h(y, z) = h(y) - h(z) - \langle \nabla h(z), y - z\rangle$.

From these relations, we deduce that the sequence $\{\tilde{l}^i(y^i)\}_i$ is nondecreasing and bounded above by $\varphi(x^k)$. So it is convergent. Moreover, we also obtain that

$$\tilde{l}^{i+1}(y^{i+1}) - \tilde{l}^{i}(y^{i}) \geq (2\mu_k)^{-1}\beta\, \|y^{i+1} - y^i\|^2 \geq 0,$$

and then $[y^{i+1} - y^i] \to 0$ (strongly) because the left-hand side tends to zero.

2. The sequence $\{y^i\}_{i\in\mathbb{N}}$ is bounded.

Let $y \in C$ be fixed. Using successively (C1), the definition of $\tilde{\theta}^{i+1}$, (2.9), (2.7), and the strong convexity of h, we have

$$\begin{aligned}
\varphi(y) + \langle r(x^k), y - x^k\rangle + \mu_k^{-1}\, [\, h(y) - h(x^k) - \langle \nabla h(x^k), y - x^k\rangle\,] &\geq \tilde{\theta}^{i+1}(y)\\
&\geq \tilde{l}^{i}(y^i) + \mu_k^{-1} D_h(y, y^i) \geq \tilde{l}^{i}(y^i) + (2\mu_k)^{-1}\beta\, \|y - y^i\|^2.
\end{aligned}$$

Since the sequence $\{\tilde{l}^i(y^i)\}_i$ is convergent, the sequence $\{y - y^i\}_i$ must be bounded and thus also the sequence $\{y^i\}_{i\in\mathbb{N}}$.

3. $[\varphi(y^{i+1}) - \theta^{i+1}(y^{i+1})] \to 0$.

Using successively (C3), (C1), and the definition of the subgradient $s(y^{i+1})$, we obtain

$$\langle s(y^i), y^{i+1} - y^i\rangle \leq \theta^{i+1}(y^{i+1}) - \varphi(y^i) \leq \varphi(y^{i+1}) - \varphi(y^i) \leq \langle s(y^{i+1}), y^{i+1} - y^i\rangle.$$

Since the subdifferential $\partial\varphi$ is bounded on the bounded sequence $\{y^i\}_{i\in\mathbb{N}}$, the sequence $\{s(y^i)\}_{i\in\mathbb{N}}$ is bounded and, as $\|y^{i+1} - y^i\| \to 0$, the opposite sides of the previous inequalities tend to zero. Hence

$$[\theta^{i+1}(y^{i+1}) - \varphi(y^i)] \to 0 \quad \text{and} \quad [\varphi(y^{i+1}) - \varphi(y^i)] \to 0,$$

and thus $\varphi(y^{i+1}) - \theta^{i+1}(y^{i+1}) = \varphi(y^{i+1}) - \varphi(y^i) + \varphi(y^i) - \theta^{i+1}(y^{i+1}) \to 0$. This establishes that $[\varphi(y^i) - \theta^i(y^i)] \to 0$.

Second, we show that $y^i \to z(x^k)$. Using successively (C1), (2.4), (2.3), and (2.2), we have for all $y \in C$

$$\begin{aligned} \varphi(y) &\geq \theta^i(y) \geq l^i(y) = \theta^i(y^i) + \langle \gamma^i, y - y^i \rangle \\ &= \theta^i(y^i) + \mu_k^{-1} \langle \nabla h(x^k) - \nabla h(y^i), y - y^i \rangle - \langle r(x^k), y - y^i \rangle. \end{aligned} \tag{2.10}$$

Since the sequence $\{y^i\}_i$ is bounded, we can extract a subsequence that weakly converges in C. Without loss of generality, let us suppose that $y^i \rightharpoonup \bar{y} \in C$. If we take $y = \bar{y}$ in (2.10) and if we use the strong monotonicity of ∇h, we obtain

$$\begin{aligned} \mu_k(\varphi(\bar{y}) - \theta^i(y^i)) &\geq \langle \nabla h(x^k) - \nabla h(y^i), \bar{y} - y^i \rangle - \mu_k \langle r(x^k), \bar{y} - y^i \rangle \\ &= -\mu_k \langle r(x^k), \bar{y} - y^i \rangle + \langle \nabla h(x^k) - \nabla h(\bar{y}), \bar{y} - y^i \rangle \\ &\quad + \langle \nabla h(\bar{y}) - \nabla h(y^i), \bar{y} - y^i \rangle \\ &\geq -\mu_k \langle r(x^k), \bar{y} - y^i \rangle + \langle \nabla h(x^k) - \nabla h(\bar{y}), \bar{y} - y^i \rangle \\ &\quad + \beta \|\bar{y} - y^i\|^2. \end{aligned} \tag{2.11}$$

Since φ is weakly l.s.c. and $[\varphi(y^i) - \theta^i(y^i)] \to 0$, we have directly that $\overline{\lim}_i [\varphi(\bar{y}) - \theta^i(y^i)] \leq 0$. Then passing to the superior limit in (2.11) gives that $\overline{\lim}_i \|y^i - \bar{y}\|^2 = 0$, and thus $y^i \to \bar{y}$. Now, from (2.10), we have, for all $y \in C$,

$$\begin{aligned} \varphi(y) &\geq [\theta^i(y^i) - \varphi(y^i)] + [\varphi(y^i) - \varphi(\bar{y})] \\ &\quad + \varphi(\bar{y}) + \mu_k^{-1} \langle \nabla h(x^k) - \nabla h(y^i), y - y^i \rangle - \langle r(x^k), y - y^i \rangle. \end{aligned}$$

If we take the limit in this last inequality and if we use the facts that $[\varphi(y^i) - \theta^i(y^i)] \to 0$, $y^i \to \bar{y}$, φ, and ∇h are continuous on C, we obtain that, for all $y \in C$,

$$\varphi(y) \geq \varphi(\bar{y}) + \mu_k^{-1} \langle \nabla h(x^k) - \nabla h(\bar{y}), y - \bar{y} \rangle - \langle r(x^k), y - \bar{y} \rangle.$$

This means that

$$\mu_k^{-1}(\nabla h(x^k) - \nabla h(\bar{y})) - r(x^k) \in \partial(\varphi + \psi_C)(\bar{y}),$$

and consequently that $\bar{y} = z(x^k)$. This completes the proof. ∎

This basic result gives information on what happens in the bundle algorithm if only null-steps are made after some *outer* iterate has been reached. So, it will be used to prove the following first convergence property of the bundle algorithm.

THEOREM 2.2. *Consider the bundle algorithm for solving problem* (P). *Suppose that Assumption* A *holds. If some iterate* x^k *is reached and, from then on,* k *remains fixed, i.e., only null-steps are performed, then* x^k *actually solves problem* (P).

Proof. The iteration that has produced x^k is denoted by i_k. From x^k, we make only null-steps. Thus, for all $i > i_k$,

$$\varphi(x^k) - \varphi(y^i) < m\,[\varphi(x^k) - \theta^i(y^i)] + (1 - m)\,\langle r(x^k), y^i - x^k \rangle.$$

If we pass to the limit on i in this inequality, and if we use the facts that $[\varphi(y^i) - \theta^i(y^i)] \to 0$, $y^i \to z(x^k)$, and φ is continuous on C, we obtain

$$\varphi(x^k) - \varphi(z(x^k)) \leq m\,[\varphi(x^k) - \varphi(z(x^k))] + (1 - m)\,\langle r(x^k), z(x^k) - x^k \rangle.$$

Since $1-m>0$, this means that

$$\varphi(x^k) \le \varphi(z(x^k)) + \langle r(x^k), z(x^k)-x^k\rangle. \tag{2.12}$$

On the other hand, by definition of $z(x^k)$, we have for all $y \in C$

$$\varphi(z(x^k)) + \langle r(x^k), z(x^k)-x^k\rangle + \mu_k^{-1}[h(z(x^k)) - h(x^k) - \langle \nabla h(x^k), z(x^k)-x^k\rangle]$$
$$\le \varphi(y) + \langle r(x^k), y-x^k\rangle + \mu_k^{-1}[h(y) - h(x^k) - \langle \nabla h(x^k), y-x^k\rangle].$$

If we take $y=x^k$ in this last inequality, we deduce that

$$\varphi(z(x^k)) + \langle r(x^k), z(x^k)-x^k\rangle$$
$$\le \varphi(x^k) + \mu_k^{-1}[h(x^k) - h(z(x^k)) + \langle \nabla h(x^k), z(x^k)-x^k\rangle].$$

Since h is strongly convex with modulus $\beta>0$, we have

$$h(x^k) - h(z(x^k)) + \langle \nabla h(x^k), z(x^k)-x^k\rangle \le -(\beta/2)\|z(x^k)-x^k\|^2.$$

Hence,

$$\varphi(z(x^k)) + \langle r(x^k), z(x^k)-x^k\rangle$$
$$\le \varphi(x^k) - \beta\,(2\mu_k)^{-1}\|z(x^k)-x^k\|^2 \le \varphi(x^k). \tag{2.13}$$

Combining (2.12) and (2.13), we deduce easily that $z(x^k)=x^k$. From the definition of $z(x^k)$, this means that $0 \in r(x^k) + \partial(\varphi+\psi_C)(x^k)$, i.e., x^k solves problem (P). This completes the proof. □

When the case considered in this theorem does not occur, k tends to $+\infty$ and the bundle algorithm generates an infinite sequence $\{x^k\}_{k\in\mathbb{N}}$. We now study the convergence of this sequence. That is the purpose of sections 3 and 4.

3. Convergence when stepsizes go to zero. In this section, we suppose that the bundle algorithm generates an infinite sequence $\{x^k\}_{k\in\mathbb{N}}$. The operator F is multivalued and the sequence $\{\mu_k\}_{k\in\mathbb{N}}$ is chosen to be of the following form:

$$\begin{cases} \mu_k = \lambda_k/\eta_k \text{ for all } k\in\mathbb{N}, \text{ with } \{\lambda_k\}_{k\in\mathbb{N}} \text{ a sequence of positive numbers;} \\ \\ \eta_k = \begin{cases} \max\{1, \|r(x^0)\|\} \text{ if } k=0; \\ \\ \max\{\eta_{k-1}, \|r(x^k)\|\} \text{ if } k\ge 1. \end{cases} \end{cases}$$

The introduction of the sequence $\{\eta_k\}_{k\in\mathbb{N}}$ allows us to prove that the sequence $\{x^k\}_{k\in\mathbb{N}}$ is bounded without any additional assumption on the mapping F. Moreover, as it is classically assumed in the multivalued case (see, e.g., [10]), the positive sequence $\{\lambda_k\}_{k\in\mathbb{N}}$ will be such that $\sum_{k=0}^{+\infty}\lambda_k^2 < +\infty$ and $\sum_{k=0}^{+\infty}\lambda_k = +\infty$. This rule is also considered in the literature for nonsmooth minimization problems; see, e.g., [1, 4, 8, 32]. We proceed in three steps to prove the convergence of the algorithm. First, we study the boundedness of the sequence $\{x^k\}_{k\in\mathbb{N}}$, then its weak convergence, and finally its strong convergence.

In the convergence proofs, we consider the sequence $\{\Gamma^k(x^*,\cdot)\}_{k\in\mathbb{N}}$ of Lyapunov functions defined on C by

$$\Gamma^k(x^*,x) = h(x^*) - h(x) - \langle \nabla h(x), x^*-x\rangle$$
$$+ \lambda_k\,(m\,\eta_k)^{-1}[\langle r(x^*), x-x^*\rangle + \varphi(x) - \varphi(x^*)], \tag{3.1}$$

where $x^* \in C$ denotes a solution of problem (P) and $r(x^*)$ is the element in $F(x^*)$ such that $\langle r(x^*), x - x^* \rangle + \varphi(x) - \varphi(x^*) \geq 0$ for all x in C. Since h is strongly convex with modulus $\beta > 0$, we have immediately that, for all $x \in C$,

$$\Gamma^k(x^*, x) \geq (\beta/2)\|x - x^*\|^2. \tag{3.2}$$

The next lemma gives an upper bound on $\Gamma^{k+1}(x^*, x^{k+1}) - \Gamma^k(x^*, x^k)$, which will be often used in what follows.

LEMMA 3.1. *Suppose that Assumption* A *holds and that* $\{\lambda_k\}_{k\in\mathbb{N}}$ *is a nonincreasing sequence of positive numbers. Then we have for all* $k \in \mathbb{N}$,

$$\begin{aligned}\Gamma^{k+1}(x^*, x^{k+1}) - \Gamma^k(x^*, x^k) \leq -c\|x^{k+1} - x^k\|^2 + \lambda_k^2 u \\ + (\lambda_k/\eta_k)[\langle r(x^k), x^* - x^k\rangle + \varphi(x^*) - \varphi(x^k)],\end{aligned} \tag{3.3}$$

with $c, u > 0$.

Proof. First observe that the optimality conditions satisfied by $x^{k+1} \in C$ are

$$\begin{aligned}\langle \eta_k^{-1} r(x^k) + \lambda_k^{-1}(\nabla h(x^{k+1}) - \nabla h(x^k)), x - x^{k+1}\rangle \\ + \eta_k^{-1}(\theta^{i_{k+1}}(x) - \theta^{i_{k+1}}(x^{k+1})) \geq 0 \quad \text{for all } x \in C,\end{aligned} \tag{3.4}$$

where $r(x^k) \in F(x^k)$. Using the definition of the Lyapunov function and noticing that $\lambda_{k+1} \leq \lambda_k$, and $\eta_{k+1} \geq \eta_k$ for all $k \in \mathbb{N}$, we can write

$$\Gamma^{k+1}(x^*, x^{k+1}) - \Gamma^k(x^*, x^k) \leq \Gamma^k(x^*, x^{k+1}) - \Gamma^k(x^*, x^k) = s_1 + s_2 + s_3, \tag{3.5}$$

with

$$\begin{aligned} s_1 &= h(x^k) - h(x^{k+1}) + \langle \nabla h(x^k), x^{k+1} - x^k\rangle, \\ s_2 &= \langle \nabla h(x^k) - \nabla h(x^{k+1}), x^* - x^{k+1}\rangle, \\ s_3 &= \lambda_k\,(m\,\eta_k)^{-1}[\langle r(x^*), x^{k+1} - x^k\rangle + \varphi(x^{k+1}) - \varphi(x^k)]. \end{aligned}$$

For s_1, we derive easily from the strong convexity of h that

$$s_1 \leq -(\beta/2)\|x^{k+1} - x^k\|^2. \tag{3.6}$$

Using (3.4) with $x = x^*$, we obtain

$$\begin{aligned} s_2 &\leq (\lambda_k/\eta_k)[\langle r(x^k), x^* - x^{k+1}\rangle + \theta^{i_{k+1}}(x^*) - \theta^{i_{k+1}}(x^{k+1})] \\ &= (\lambda_k/\eta_k)[\langle r(x^k), x^* - x^k\rangle + \varphi(x^*) - \varphi(x^k) \\ &\qquad + \langle r(x^k), x^k - x^{k+1}\rangle \\ &\qquad + \theta^{i_{k+1}}(x^*) - \varphi(x^*) + \varphi(x^k) - \theta^{i_{k+1}}(x^{k+1})]. \end{aligned} \tag{3.7}$$

From the stopping test (2.1), we deduce that

$$\varphi(x^k) - \theta^{i_{k+1}}(x^{k+1}) \leq \frac{1}{m}\,[\varphi(x^k) - \varphi(x^{k+1})] - \frac{1-m}{m}\,\langle r(x^k), x^{k+1} - x^k\rangle. \tag{3.8}$$

Combining the fact that $\theta^{i_{k+1}} \leq \varphi$ with (3.7) and (3.8), we derive that

$$\begin{aligned} s_2 + s_3 &\leq (\lambda_k/\eta_k)[\langle r(x^k), x^* - x^k\rangle + \varphi(x^*) - \varphi(x^k) \\ &\qquad + (1/m)\langle r(x^k) - r(x^*), x^k - x^{k+1}\rangle] \\ &\leq (\lambda_k/\eta_k)[\langle r(x^k), x^* - x^k\rangle + \varphi(x^*) - \varphi(x^k)] \\ &\qquad + (1/m)\,[(1/2\tau)\,(\lambda_k^2/\eta_k^2)\|r(x^k) - r(x^*)\|^2 + (\tau/2)\|x^k - x^{k+1}\|^2], \end{aligned} \tag{3.9}$$

where τ is any positive constant.

From the definition of the sequence $\{\eta_k\}_{k\in\mathbb{N}}$, we have

$$\begin{aligned}(1/\eta_k^2)\|r(x^k)-r(x^*)\|^2 &\leq (1/\eta_k^2)\left[\|r(x^k)\|^2+\|r(x^*)\|^2+2\|r(x^k)\|\,\|r(x^*)\|\right]\\ &\leq 1+\|r(x^*)\|^2+2\|r(x^*)\|\\ &= [1+\|r(x^*)\|]^2.\end{aligned}\tag{3.10}$$

Combining inequalities (3.5), (3.6), (3.9), (3.10), we obtain

$$\begin{aligned}&\Gamma^{k+1}(x^*,x^{k+1})-\Gamma^k(x^*,x^k)\\ &\quad\leq -(1/2)(\beta-\tau/m)\,\|x^{k+1}-x^k\|^2+(2m\tau)^{-1}[1+\|r(x^*)\|]^2\,\lambda_k^2\\ &\qquad+(\lambda_k/\eta_k)[\langle r(x^k),x^*-x^k\rangle+\varphi(x^*)-\varphi(x^k)].\end{aligned}\tag{3.11}$$

If we choose τ such that $0<\tau<\beta m$, then we obtain that inequality (3.3) holds with

$$\begin{aligned}c &= (1/2)(\beta-\tau/m)>0,\\ u &= (2m\tau)^{-1}[1+\|r(x^*)\|]^2>0.\end{aligned}$$ □

The next theorem gives conditions to ensure boundedness of the sequence $\{x^k\}_{k\in\mathbb{N}}$.

THEOREM 3.2. *Assume that the assumptions of Lemma* 3.1 *hold. If* $\sum_{k=0}^{+\infty}\lambda_k^2<+\infty$*, then the sequence* $\{\Gamma^k(x^*,x^k)\}_{k\in\mathbb{N}}$ *is convergent, the sequence* $\{x^k\}_{k\in\mathbb{N}}$ *is bounded,* $\sum_{k=0}^{+\infty}\|x^{k+1}-x^k\|^2<+\infty$*, and*

$$\sum_{k=0}^{+\infty}(\lambda_k/\eta_k)[\langle r(x^k),x^k-x^*\rangle+\varphi(x^k)-\varphi(x^*)]<+\infty.\tag{3.12}$$

Proof. Since x^* is a solution of problem (P), $r(x^k)\in F(x^k)$ for all k, and F is monotone, we have that

$$(\lambda_k/\eta_k)[\langle r(x^k),x^*-x^k\rangle+\varphi(x^*)-\varphi(x^k)]\leq 0.$$

So, we derive from (3.3) that

$$\Gamma^{k+1}(x^*,x^{k+1})-\Gamma^k(x^*,x^k)\leq\lambda_k^2\,u.$$

Since the series $\sum_{k=0}^{+\infty}\lambda_k^2$ is convergent, it follows that $\{\Gamma^k(x^*,x^k)\}_{k\in\mathbb{N}}$ is a convergent sequence. Using inequality (3.2), we conclude that the sequence $\{x^k\}_{k\in\mathbb{N}}$ is bounded. Then, rearranging the terms of inequality (3.3) as

$$\begin{aligned}&c\|x^{k+1}-x^k\|^2+(\lambda_k/\eta_k)[\,\langle r(x^k),x^k-x^*\rangle+\varphi(x^k)-\varphi(x^*)\,]\\ &\quad\leq\Gamma^k(x^*,x^k)-\Gamma^{k+1}(x^*,x^{k+1})+\lambda_k^2u,\end{aligned}$$

we obtain, using the convergence of the sequence $\{\Gamma^k(x^*,x^k)\}_{k\in\mathbb{N}}$ and of the series $\sum_{k=0}^{+\infty}\lambda_k^2$, that $\sum_{k=0}^{+\infty}\|x^{k+1}-x^k\|^2<+\infty$ and that (3.12) holds. □

To prove that any weak limit point of the sequence $\{x^k\}_{k\in\mathbb{N}}$ is a solution of problem (P), we will use the concept of gap function (see, e.g., [2]). We recall that a function $l:C\longrightarrow\mathbb{R}\cup\{+\infty\}$ is a gap function with respect to problem (P) if

$$\text{for all } x\in C, l(x)\geq 0 \text{ and } l(\bar{x})=0 \text{ if and only if } \bar{x} \text{ is a solution of } (P).$$

In our context, the usefulness of the gap functions appears in the next proposition.

PROPOSITION 3.3. *Let l be a gap function with respect to problem (P). If l is a weakly l.s.c. function on C and if $l(x^k) \to 0$, then any weak limit point of the sequence $\{x^k\}_{k\in\mathbb{N}}$ generated by the algorithm is a solution of (P).*

Proof. First, notice that the sequence $\{x^k\}_{k\in\mathbb{N}}$ is contained in C. Then, let $\bar{x}$ be a weak limit point of this sequence. We have $x^{k_p} \rightharpoonup \bar{x}$ and, by assumption, that

$$0 = \lim_{k\to+\infty} l(x^k) = \liminf_{p\to+\infty} l(x^{k_p}) \geq l(\bar{x}) \geq 0,$$

i.e., $l(\bar{x}) = 0$ and $\bar{x}$ is a solution of (P). □

To prove that $l(x^k) \to 0$, we will use the following lemma due to Cohen and Zhu [8, Lemma 4].

LEMMA 3.4. *If l is a Lipschitz continuous function on $\{x^k \mid k \in \mathbb{N}\}$, and if $\{\lambda_k\}$ is a sequence of positive numbers such that*

(a) $\sum \lambda_k = +\infty$;

(b) $\sum \lambda_k l(x^k) < +\infty$;

(c) $\exists\, \delta > 0$ *such that for all* $k \in \mathbb{N}$, $\|x^{k+1} - x^k\| \leq \delta\lambda_k$,

then $l(x^k) \to 0$.

First we give three existence results of gap functions weakly l.s.c. on C and Lipschitz continuous on bounded subsets of C. Then we prove that assumptions (b) and (c) of Lemma 3.4 are satisfied for our algorithm. However, before giving these results, we need to recall some definitions and properties concerning multivalued operators. A multivalued operator F is said to be Lipschitz continuous on a subset B of C if

$$\exists\, L > 0 \text{ such that for all } x, y \in B, \quad e(F(x), F(y)) \leq L\,\|x - y\|,$$

where $e(F(x), F(y)) = \sup_{r\in F(x)} \inf_{s\in F(y)} \|r - s\|$. The next lemma will be used in what follows.

LEMMA 3.5. *Let B be a bounded subset of C. If F is Lipschitz continuous on B, and if there exists $\bar{y} \in B$ such that $F(\bar{y})$ is bounded, then F is bounded on B, i.e., there exists $\alpha > 0$ such that $\|r(x)\| \leq \alpha$ for all $x \in B$ and $r(x) \in F(x)$.*

Proof. Let $\epsilon > 0$. Then, by assumption, $e(F(x), F(\bar{y})) \leq L\|x - \bar{y}\|$ for all $x \in B$, i.e.,

$$\text{for all } x \in B, \text{ for all } r(x) \in F(x),\ \exists\, r(\bar{y}) \in F(\bar{y}) \text{ such that } \|r(x) - r(\bar{y})\| \leq L\|x - \bar{y}\| + \epsilon.$$

Since B and $F(\bar{y})$ are bounded, there exist $\alpha_1 > 0$ and $\alpha_2 > 0$ such that $\|x\| \leq \alpha_1$ for all $x \in B$ and $\|r(\bar{y})\| \leq \alpha_2$ for all $r(\bar{y}) \in F(\bar{y})$. Then, for all $x \in B$ and $r(x) \in F(x)$, we have successively

$$\begin{aligned} \|r(x)\| &\leq \|\, r(x) - r(\bar{y})\,\| + \|\, r(\bar{y})\,\| \\ &\leq L\,[\,\|x\| + \|\,\bar{y}\,\|\,] + \epsilon + \alpha_2 \\ &\leq L\,[\,\alpha_1 + \|\,\bar{y}\,\|\,] + \epsilon + \alpha_2, \end{aligned}$$

i.e., what we have to prove. □

A multivalued operator F is said to be weakly closed on C if

$$z^k \rightharpoonup \bar{z},\ z^k \in C \text{ and } \ r^k \rightharpoonup \bar{r}, r^k \in F(z^k) \implies \bar{r} \in F(\bar{z}).$$

In particular, when F is weakly closed on C, $F(z)$ is a weakly closed subset of H for each $z \in C$.

A multivalued operator F is paramonotone on C if F is monotone on C and, for all $x, y \in C$, and, $r(x) \in F(x), r(y) \in F(y)$,

$$\langle r(x) - r(y), x - y\rangle = 0 \implies r(y) \in F(x) \text{ and } r(x) \in F(y).$$

This notion was introduced by Bruck [7] and further studied in [18]. Let us mention the following result due to Iusem [18]: If F is paramonotone, and if x^* is a solution of (P), then $\bar{x}$ is a solution of (P) if and only if

$$\bar{x} \in C \text{ and } \exists\, \bar{r} \in F(\bar{x}) \text{ such that } \langle \bar{r}, x^* - \bar{x}\rangle + \varphi(x^*) - \varphi(\bar{x}) \geq 0. \tag{3.13}$$

PROPOSITION 3.6. *Let x^* denote any solution of problem (P).*

(a) *If F is paramonotone on C, and $F(x)$ is a bounded and weakly closed subset of H for all $x \in C$, then $l(x) = \inf_{r(x)\in F(x)}\langle r(x), x - x^*\rangle + \varphi(x) - \varphi(x^*)$ is a gap function.*

(b) *If, in addition, F and φ are Lipschitz continuous on bounded subsets of C, then l is Lipschitz continuous on bounded subsets of C.*

(c) *If, in addition, F is weakly closed on C, then l is weakly l.s.c. on C.*

Proof. (a) Since F is monotone and x^* is a solution of (P), for each $x \in C$ and $r(x) \in F(x)$, we have

$$\begin{aligned}\langle r(x), x - x^*\rangle + \varphi(x) - \varphi(x^*) &= \langle r(x) - r(x^*), x - x^*\rangle \\ &\quad + \langle r(x^*), x - x^*\rangle + \varphi(x) - \varphi(x^*) \geq 0.\end{aligned}$$

So, using the definition of l, we obtain that $l(x) \geq 0$. Now if $\bar{x}$ is a solution of (P), then we have immediately that

$$l(\bar{x}) \leq \langle r(\bar{x}), \bar{x} - x^*\rangle + \varphi(\bar{x}) - \varphi(x^*) \leq 0 \leq l(\bar{x}).$$

So, $l(\bar{x}) = 0$. Conversely, suppose that $l(\bar{x}) = 0$. Then, by definition of the infimum, there exists a sequence $\{r_k\}_{k\in\mathbb{N}}$ contained in $F(\bar{x})$ such that, for all $k \geq 1$,

$$0 \leq \langle r_k, \bar{x} - x^*\rangle + \varphi(\bar{x}) - \varphi(x^*) < 1/k.$$

Since the subset $F(\bar{x})$ is bounded and weakly closed, there exists a subsequence of $\{r_k\}_{k\in\mathbb{N}}$ that weakly converges to some $r \in F(\bar{x})$. Then $0 \leq \langle r, \bar{x} - x^*\rangle + \varphi(\bar{x}) - \varphi(x^*) \leq 0$, and by (3.13), $\bar{x}$ is a solution of (P) because F is paramonotone.

(b) Let B be a bounded subset of C and $\alpha_1 > 0$ be such that $\|x\| \leq \alpha_1$ for all $x \in B$. Since φ is Lipschitz continuous on B, it is sufficient to prove that there exists $L_1 > 0$ such that, for all $x, y \in B$,

$$\inf_{r\in F(x)} \langle r, x - x^*\rangle + \sup_{s\in F(y)} \langle s, x^* - y\rangle \leq L_1\|x - y\|. \tag{3.14}$$

Let $x, y \in B, \epsilon > 0$, and $s \in F(y)$. Since $e(F(y), F(x)) \leq L\|x - y\|$, we have

$$\inf_{r\in F(x)} \|r - s\| \leq L\|x - y\|.$$

So, there exists $r \in F(x)$ such that $\|r - s\| \leq L\|x - y\| + \epsilon/(\alpha_1 + \|x^*\|)$. Then

$$\begin{aligned}\langle r, x - x^*\rangle + \langle s, x^* - y\rangle &= \langle r, x - y\rangle + \langle r - s, y - x^*\rangle \\ &\leq \|r\|\|x - y\| + \|r - s\|\|y - x^*\| \\ &\leq \|r\|\|x - y\| + L\|x - y\|(\alpha_1 + \|x^*\|) + \epsilon.\end{aligned} \tag{3.15}$$

Moreover, by Lemma 3.5, F is bounded on B and, consequently, there exists $\alpha > 0$ such that $\|r\| \leq \alpha$ for all $x \in B$ and $r \in F(x)$. Then, from (3.15), we deduce that

$$\inf_{r \in F(x)} \langle r, x - x^* \rangle + \langle s, x^* - y \rangle \leq L_1 \|x - y\| + \epsilon,$$

where $L_1 = \alpha + L(\alpha_1 + \|x^*\|)$. Since this inequality is satisfied for all $s \in F(y)$ and $\epsilon > 0$, we obtain (3.14).

(c) Suppose that F is weakly closed on C. Since φ is weakly l.s.c. on C, we have only to prove that

$$l^1(x) \equiv \inf_{r(x) \in F(x)} \langle r(x), x - x^* \rangle$$

is weakly l.s.c. on C. Let $x^k \rightharpoonup \bar{x}$ with $x^k \in C$, and let $\bar{l}^1$ be a limit point of the sequence $\{l^1(x^k)\}_{k \in \mathbb{N}}$. We have to prove that $\bar{l}^1 \geq l^1(\bar{x})$. Without loss of generality, we can assume that $l^1(x^k) \to \bar{l}^1$. Let $\epsilon > 0$. By definition of the infimum, for each k, there exists $r(x^k) \in F(x^k)$ such that

$$\langle r(x^k), x^k - x^* \rangle \leq l^1(x^k) + \epsilon. \tag{3.16}$$

Since the sequence $\{x^k\}_{k \in \mathbb{N}}$ is bounded and contained in C, and since F is bounded on bounded subsets of C, the sequence $\{r(x^k)\}_{k \in \mathbb{N}}$ is bounded, and thus there exists a subsequence $\{r(x^{k'})\}_{k' \in K}$ weakly converging to some $\bar{r}$. Since F is weakly closed, it follows that $\bar{r} \in F(\bar{x})$. Now, F being monotone, we have that $\langle r(x^{k'}) - r(\bar{x}), x^{k'} - \bar{x} \rangle \geq 0$, and thus that

$$\langle r(x^{k'}), x^{k'} - x^* \rangle \geq \langle r(\bar{x}), x^{k'} - \bar{x} \rangle + \langle r(x^{k'}), \bar{x} - x^* \rangle. \tag{3.17}$$

Combining (3.16) and (3.17), we obtain

$$l^1(x^{k'}) + \epsilon \geq \langle r(\bar{x}), x^{k'} - \bar{x} \rangle + \langle r(x^{k'}), \bar{x} - x^* \rangle. \tag{3.18}$$

Passing to the limit in (3.18) and noticing that $\langle \bar{r}, \bar{x} - x^* \rangle \geq l^1(\bar{x})$, we have that $\bar{l}^1 + \epsilon \geq l^1(\bar{x})$. Since ϵ is arbitrary, we have that $\bar{l}^1 \geq l^1(\bar{x})$ and, consequently, l is weakly l.s.c. on C. □

PROPOSITION 3.7. *Let x^* denote any solution of problem (P). If $F = \partial f, f \in \Gamma_0(H)$, and $C \subseteq \operatorname{int}(\mathrm{dom} f)$, then $l(x) = f(x) + \varphi(x) - f(x^*) - \varphi(x^*)$ is a gap function such that, for all $x \in C$ and $r(x) \in F(x)$,*

$$\langle r(x), x - x^* \rangle + \varphi(x) - \varphi(x^*) \geq l(x).$$

The function l is convex and weakly l.s.c. on C and, if in addition f and φ are Lipschitz continuous on bounded subsets of C, then l is also Lipschitz continuous on bounded subsets of C.

Proof. For all $x \in C, r(x) \in F(x) = \partial f(x)$, we have $f(x^*) \geq f(x) + \langle r(x), x^* - x \rangle$. So, we obtain

$$\langle r(x), x - x^* \rangle + \varphi(x) - \varphi(x^*) \geq f(x) - f(x^*) + \varphi(x) - \varphi(x^*) = l(x).$$

The remainder of the proof is obvious. □

PROPOSITION 3.8. *If F is strongly monotone of modulus $\alpha > 0$ on C, then $l(x) = \|x - x^*\|^2$ is a gap function such that, for all $x \in C$ and $r(x) \in F(x)$,*

$$\langle r(x), x - x^* \rangle + \varphi(x) - \varphi(x^*) \geq \alpha l(x),$$

where x^* denotes the unique solution of (P). Moreover, l is strongly convex, weakly l.s.c. on H, and Lipschitz continuous on bounded subsets of C.

Proof. Since x^* is the unique solution of problem (P), it is obvious that l is a gap function and that l is strongly convex and weakly l.s.c. on H. Moreover, for all $x \in C$, we have

$$\langle r(x), x - x^* \rangle + \varphi(x) - \varphi(x^*) = \langle r(x) - r(x^*), x - x^* \rangle + \langle r(x^*), x - x^* \rangle + \varphi(x) - \varphi(x^*).$$

Since F is strongly monotone of modulus α and x^* is the solution of (P), we obtain immediately that the right-hand side of the previous equality is not less than $\alpha l(x)$. Finally, let B be a bounded subset of C. Then there exists $\alpha_1 > 0$ such that $\|z\| \leq \alpha_1$ for all $z \in B$. So, for $x, y \in B$, we have successively

$$\begin{aligned} \|x - x^*\|^2 - \|y - x^*\|^2 &= \|x - y\|^2 + 2\langle x - y, y - x^* \rangle \\ &\leq \|x - y\|[\ \|x - y\| + 2\|y - x^*\|\ ] \\ &\leq \|x - y\|[\ 4\alpha_1 + 2\|x^*\|\ ]; \end{aligned}$$

i.e., l is Lipschitz continuous on B. □

In order to get a more general convergence result, we put together, in the same assumption, the properties requested on the gap function. These properties are satisfied in the three situations described in Propositions 3.6, 3.7, and 3.8.

Assumption I.

(i) $\exists\, \alpha > 0,\ \exists\, l : C \to \mathbb{R} \cup \{+\infty\}$ such that

$$\text{for all } x \in C,\ \text{for all } r(x) \in F(x), \quad \langle r(x), x - x^* \rangle + \varphi(x) - \varphi(x^*) \geq \alpha\ l(x);$$

(ii) for all $x \in C$, $l(x) \geq 0$ and $l(\bar{x}) = 0 \Leftrightarrow \bar{x}$ is a solution of (P);

(iii) l is weakly l.s.c. on C and Lipschitz continuous on bounded subsets of C.

The purpose of the next proposition is to prove that conditions (b) and (c) of Lemma 3.4 are satisfied.

PROPOSITION 3.9. (a) *Assume that the assumptions of Theorem* 3.2 *and Assumption* I(i), (ii) *are satisfied. If F is bounded on bounded subsets of C, then* $\sum \lambda_k l(x^k) < +\infty$.

(b) *Assume that Assumption* A *holds and that the sequence $\{\theta^i\}_{i \in \mathbb{N}_0}$ satisfies condition* (C4). *Then there exists $\delta > 0$ such that, for all $k \geq 1$, $\|x^{k+1} - x^k\| \leq \delta \lambda_k$.*

Proof. (a) Since the sequence $\{x^k\}_{k \in \mathbb{N}}$ is bounded and F is bounded on bounded subsets of C, the sequences $\{r(x^k)\}_{k \in \mathbb{N}}$ and also $\{\eta_k\}_{k \in \mathbb{N}}$ are bounded. Then, using successively Theorem 3.2 and Assumption I(i), (ii), we have

$$\sum_{k=1}^{+\infty} \lambda_k [\langle r(x^k), x^k - x^* \rangle + \varphi(x^k) - \varphi(x^*)] < +\infty \text{ and } \sum_{k=1}^{+\infty} \lambda_k l(x^k) < +\infty.$$

(b) From the optimality conditions (3.4) applied to $x = x^k$, we obtain

$$\begin{aligned} &\langle \nabla h(x^{k+1}) - \nabla h(x^k), x^{k+1} - x^k \rangle \\ &\quad \leq (\lambda_k / \eta_k)[\langle r(x^k), x^k - x^{k+1} \rangle + \theta^{i_{k+1}}(x^k) - \theta^{i_{k+1}}(x^{k+1})]. \end{aligned} \tag{3.19}$$

Since h is strongly convex and $\|r(x^k)\| \leq \eta_k$, we derive from (3.19) that

$$\beta \|x^{k+1} - x^k\|^2 \leq \lambda_k \|x^{k+1} - x^k\| + (\lambda_k / \eta_k)[\theta^{i_{k+1}}(x^k) - \theta^{i_{k+1}}(x^{k+1})]. \tag{3.20}$$

Now since $\theta^{i_{k+1}} \leq \varphi$ and, by construction (see condition (C4)),

$$\theta^{i_{k+1}}(x) \geq \varphi(x^k) + \langle s(x^k), x - x^k \rangle \text{ for all } x \in C,$$

we have

$$\begin{aligned} \theta^{i_{k+1}}(x^k) - \theta^{i_{k+1}}(x^{k+1}) &\leq \varphi(x^k) - \varphi(x^k) - \langle s(x^k), x^{k+1} - x^k \rangle \\ &= \langle s(x^k), x^k - x^{k+1} \rangle \\ &\leq \|s(x^k)\| \, \|x^{k+1} - x^k\|. \end{aligned}$$

Hence, since $\partial\varphi$ is bounded on bounded subsets of C, we have that the sequence $\{\|s(x^k)\|\}_{k\in\mathbb{N}}$ is bounded and there exists $\delta_\varphi > 0$ such that, for all k,

$$\theta^{i_{k+1}}(x^k) - \theta^{i_{k+1}}(x^{k+1}) \leq \delta_\varphi \, \|x^{k+1} - x^k\|. \tag{3.21}$$

Finally, from (3.20), (3.21), and since $\eta_k \geq 1$, we deduce that $\|x^{k+1} - x^k\| \leq \delta\lambda_k$ for all k, with $\delta = (1/\beta)[1 + \delta_\varphi]$. □

We are now ready to state our main convergence result.

THEOREM 3.10. *Suppose that the following conditions are satisfied:*

- *Assumptions* A *and* I *hold.*
- *F is bounded on bounded subsets of C.*
- *$\{\lambda_k\}_{k\in\mathbb{N}}$ is nonincreasing and $\sum \lambda_k = +\infty$, $\sum \lambda_k^2 < +\infty$.*

Then the sequence $\{x_k\}_{k\in\mathbb{N}}$ is bounded, $l(x^k) \to 0$, and any weak limit point of $\{x_k\}_k$ is a solution of problem (P). If, in addition, ∇h is weakly continuous on C, then $\{x_k\}_k$ weakly converges to a solution of (P). If, in addition, the gap function l is strongly convex on an open set containing C, then $x^k \to x^*$, the unique solution of (P).

Proof. The first part of the theorem follows immediately from Lemma 3.1, Theorem 3.2, Proposition 3.9, Lemma 3.4, and Proposition 3.3. Suppose now that ∇h is weakly continuous on C and that the sequence $\{x^k\}_{k\in\mathbb{N}}$ has two different weak limit points x^1 and x^2. Let $\{x^{m(k)}\}_{k\in\mathbb{N}}$ be the subsequence of $\{x^k\}_{k\in\mathbb{N}}$ weakly converging to x^1 and $\{x^{n(k)}\}_{k\in\mathbb{N}}$ be the subsequence weakly converging to x^2. By the first part of the theorem, x^1 and x^2 are solutions of problem (P). Then, by Theorem 3.2, the sequences of Lyapunov functions $\{\Gamma^k(x^1, x^k)\}_{k\in\mathbb{N}}$ and $\{\Gamma^k(x^2, x^k)\}_{k\in\mathbb{N}}$ are convergent in $\mathbb{R}$. We denote their limits by Γ_1 and Γ_2, respectively. By definition of the Lyapunov function, we have

$$\begin{aligned} &\Gamma^{n(k)}(x^1, x^{n(k)}) - \Gamma^{n(k)}(x^2, x^{n(k)}) \\ &\quad = h(x^1) - h(x^2) - \langle \nabla h(x^{n(k)}), x^1 - x^2 \rangle \\ &\qquad + \lambda_{n(k)} \, (m \, \eta_{n(k)})^{-1} [\langle r(x^1), x^{n(k)} - x^1 \rangle - \langle r(x^2), x^{n(k)} - x^2 \rangle + \varphi(x^2) - \varphi(x^1)]. \end{aligned}$$

Since ∇h is weakly continuous on C, since $\eta_k \geq 1$ for all k, and since $\lambda_k \to 0$, we obtain, taking the limit on k in the last equality, that

$$\Gamma_1 - \Gamma_2 = h(x^1) - h(x^2) - \langle \nabla h(x^2), x^1 - x^2 \rangle. \tag{3.22}$$

Since the roles of x^1 and x^2 can be reversed, we also have that

$$\Gamma_1 - \Gamma_2 = h(x^1) - h(x^2) - \langle \nabla h(x^1), x^1 - x^2 \rangle. \tag{3.23}$$

Comparing (3.22) and (3.23), we obtain $\langle \nabla h(x^1) - \nabla h(x^2), x^1 - x^2 \rangle = 0$. Since ∇h is strongly monotone, this inequality implies that $x^1 = x^2$. So the sequence $\{x^k\}_{k\in\mathbb{N}}$ weakly converges to a solution of (P).

If the gap function l is strongly convex with constant $s > 0$ on an open convex set containing C, then x^* is the unique solution of problem (P), $\partial l(x^*)$ is nonempty, and for any $e^* \in \partial l(x^*)$,

$$l(x^k) - l(x^*) - \langle e^*, x^k - x^* \rangle \geq (s/2)\|x^k - x^*\|^2. \tag{3.24}$$

Since $l(x^k) \to 0$, $l(x^*) = 0$, and $x^k \rightharpoonup x^*$, we obtain, passing to the limit in (3.24), that $\|x^k - x^*\| \to 0$, i.e., $x^k \to x^*$ strongly. This completes the proof. ∎

Remark. For example, if $h(x) = (1/2)x^T x$ for all $x \in H$, then ∇h is weakly continuous on H. Moreover, when H is a finite dimensional space, ∇h is continuous in the strong topology and thus in the weak topology.

Using Propositions 3.6, 3.7, and 3.8, which give sufficient conditions to ensure that Assumption I is satisfied, we can particularize our main result (Theorem 3.10) to get two more precise convergence theorems. However, before presenting them, we prove a preliminary lemma.

LEMMA 3.11. *Let $g \in \Gamma_0(H)$ and let B be a bounded subset of* int*(dom g). If ∂g is bounded on B, then g is Lipschitz continuous on B.*

Proof. Let $x, y \in B$. Since $B \subseteq \text{int}(\text{dom } g)$, the subdifferentials $\partial g(x)$ and $\partial g(y)$ are nonempty. Let $s(x) \in \partial g(x)$ and $s(y) \in \partial g(y)$. Then

$$\begin{aligned} g(x) - g(y) &\leq \langle s(x), x - y \rangle \leq \|s(x)\| \, \|x - y\|, \\ g(y) - g(x) &\leq \langle s(y), y - x \rangle \leq \|s(y)\| \, \|y - x\|. \end{aligned}$$

So $|g(x) - g(y)| \leq L\|x - y\|$, where $L = \sup\{\|s(z)\| \mid z \in B, s(z) \in \partial g(z)\}$. Since ∂g is bounded on B, this constant L is finite, and thus g is Lipschitz continuous on B. ∎

THEOREM 3.12. *Suppose that Assumption* A *holds, ∇h is weakly continuous on C, $\{\lambda_k\}_{k \in \mathbb{N}}$ is nonincreasing, and $\sum \lambda_k = +\infty$, $\sum \lambda_k^2 < +\infty$.*

(a) *If F is paramonotone, weakly closed on C, and Lipschitz continuous on bounded subsets of C, and if $F(x)$ is a bounded subset of H for all $x \in C$, then the whole sequence $x^k \rightharpoonup \bar{x}$, where $\bar{x}$ is a solution of (P).*

(b) *If $F = \partial f$ with $f \in \Gamma_0(H)$ and $C \subseteq$ int*(dom f), and if ∂f is bounded on bounded subsets of C, then the whole sequence $x^k \rightharpoonup \bar{x}$, where $\bar{x}$ is a solution of (P).*

When H is a finite dimensional space, the assumption on ∂f is always true.

Proof. By Theorem 3.10, it is sufficient to prove that Assumption I holds and that F is bounded on bounded subsets of C.

(a) Since $\partial\varphi$ is bounded on bounded subsets of C, it follows from Lemma 3.11 that φ is Lipschitz continuous on bounded subsets of C. All the assumptions of Proposition 3.6 are then satisfied, and thus Assumption I is satisfied. Finally, using Lemma 3.5, F is bounded on bounded subsets of C.

(b) By Lemma 3.11, f and φ are Lipschitz continuous on bounded subsets of C. So, using Proposition 3.7, Assumption I is satisfied. The conclusion follows because $F = \partial f$ is bounded on bounded subsets of C. ∎

THEOREM 3.13. *Suppose that Assumption* A *holds, ∇h is weakly continuous on C, $\{\lambda_k\}_{k \in \mathbb{N}}$ is nonincreasing, and $\sum \lambda_k = +\infty$, $\sum \lambda_k^2 < +\infty$. If F is strongly monotone on C and bounded on bounded subsets of C, then the whole sequence x^k strongly converges to x^*, the unique solution of (P).*

Proof. From Proposition 3.8, we have that Assumption I is satisfied. Then the conclusion follows from Theorem 3.10 because F is bounded on bounded subsets of C and the gap function $l(x) = \|x - x^*\|^2$ is strongly convex on H. ∎

Remark. When $\varphi = 0$, these results generalize those obtained by Zhu in [41]. When $F = 0$ and $C = H$, our scheme reduces to the bundle algorithm to minimize φ on H. If we particularize our convergence results to that case, we do not recover the classical results for the proximal method because of the assumption on the sequence $\{\mu_k\}_{k\in\mathbb{N}}$. The fact that the stepsizes converge to 0 entails that the convergence rate will be slow (at best sublinear) (see [35]). In consequence, it is important to present convergence results when the stepsizes are bounded away from 0 even if we have to impose some restrictions on the mapping F.

4. Convergence when stepsizes are bounded away from zero. Here, contrary to the last section, the stepsizes are assumed to be bounded away from zero. This allows us to take, for example, constant stepsizes. In order to prove the convergence of the algorithm in this case, we have to impose stronger conditions on the mapping F. The first restriction is that the operator F be single-valued so that the subproblem (P_i^k) can be written

$$(P_i^k) \quad \min_{x\in C}\{\theta^i(x) + \langle F(x^k), x - x^k\rangle + \mu_k^{-1}[\, h(x) - h(x^k) - \langle \nabla h(x^k), x - x^k\rangle \,]\,\}.$$

The next theorem is a first step toward the convergence study of the general iterative scheme. It highlights the minimal assumptions under which each weak limit point of the sequence $\{x^k\}_{k\in\mathbb{N}}$, if it exists, is a solution of problem (P).

THEOREM 4.1. *Suppose that Assumption* A *holds. Moreover, assume that the following conditions are satisfied:*

- $F : H \to H$ *is single-valued and weakly continuous on* C*;*
- $\mu_k \geq \underline{\mu} > 0$ *for all* $k \in \mathbb{N}$*;*
- *the sequence* $\{x^k\}_{k\in\mathbb{N}}$ *is bounded and is such that the sequence* $\{\|x^{k+1} - x^k\|\}_{k\in\mathbb{N}}$ *converges to zero.*

Then every weak limit point of the sequence $\{x^k\}_{k\in\mathbb{N}}$ *is a solution of problem* (P).

Proof. Let x^* be a weak limit point of $\{x^k\}_{k\in\mathbb{N}}$ and let $\{x^k\}_{k\in K\subset\mathbb{N}}$ be a subsequence weakly converging to x^*. Since $\{\|x^{k+1} - x^k\|\}_{k\in\mathbb{N}} \to 0$, we have that $\{x^{k+1}\}_{k\in K} \rightharpoonup x^*$.

From the monotonicity of F, we deduce that for all j,

$$\begin{aligned}\langle F(x^k), y - x^{k+1}\rangle &= \langle F(x^k), y - x^*\rangle + \langle F(x^k), x^k - x^{k+1}\rangle + \langle F(x^k), x^* - x^k\rangle \\ &\leq \langle F(x^k), y - x^*\rangle + \langle F(x^k), x^k - x^{k+1}\rangle + \langle F(x^*), x^* - x^k\rangle.\end{aligned}$$

Since F is weakly continuous on C and $\|x^{k+1} - x^k\| \to 0$, we obtain that

$$\limsup_{k\in K}\langle F(x^k), y - x^{k+1}\rangle \leq \langle F(x^*), y - x^*\rangle. \tag{4.1}$$

Since $\partial\varphi$ is bounded on bounded subsets, $\{x^k\}_{k\in\mathbb{N}}$ is bounded, and $\|x^{k+1} - x^k\| \to 0$, we deduce that $[\varphi(x^k) - \varphi(x^{k+1})] \to 0$. Moreover, from the stopping test (2.1), we have that

$$\begin{aligned} 0 &\leq \varphi(x^{k+1}) - \theta^{i_{k+1}}(x^{k+1}) \\ &\leq \tfrac{1-m}{m}\,[\varphi(x^k) - \varphi(x^{k+1}) - \langle F(x^k), x^{k+1} - x^k\rangle].\end{aligned}$$

If we pass to the limit in these last inequalities on $k \in K$, and if we use the facts that $[\varphi(x^k) - \varphi(x^{k+1})] \to 0$, $\|x^{k+1} - x^k\| \to 0$, and F is weakly continuous, we deduce that $\lim_{k\in K}[\varphi(x^{k+1}) - \theta^{i_{k+1}}(x^{k+1})] = 0$. Therefore, since φ is weakly l.s.c., we derive that

$$\liminf_{k\in K} \theta^{i_{k+1}}(x^{k+1}) = \liminf_{k\in K}[\theta^{i_{k+1}}(x^{k+1}) - \varphi(x^{k+1}) + \varphi(x^{k+1})] \geq \varphi(x^*). \tag{4.2}$$

Now, by definition of $\{x^k\}_{k\in\mathbb{N}}$, we have that, for all $k \in \mathbb{N}$,

$$0 \leq \langle F(x^k) + \mu_k^{-1}(\nabla h(x^{k+1}) - \nabla h(x^k)), y - x^{k+1}\rangle + \theta^{i_{k+1}}(y) - \theta^{i_{k+1}}(x^{k+1}).$$

Passing then to the superior limit on $k \in K$ in the above inequality and using the Lipschitz continuity of ∇h together with relations (4.1), (4.2) and the facts that $\mu_k \geq \underline{\mu} > 0$ and $\theta^{i_{k+1}} \leq \varphi$, we obtain the following inequality:

$$0 \leq \langle F(x^*), y - x^*\rangle + \varphi(y) - \varphi(x^*),$$

which means that x^* is a solution of (P). □

We now study under which conditions the sequence $\{x^k\}_{k\in\mathbb{N}}$ is bounded and the sequence $\{\|x^{k+1} - x^k\|\}_{k\in\mathbb{N}}$ converges to zero.

THEOREM 4.2. *Suppose that Assumption* A *holds and that the following conditions are satisfied:*

- *There exist $\underline{\mu}$ and $\overline{\mu}$ such that, for all $k \in \mathbb{N}$,*

$$0 < \underline{\mu} \leq \mu_{k+1} \leq \mu_k \leq \overline{\mu}.$$

- *F satisfies the following pseudo-Dunn property with some modulus $\gamma > \overline{\mu}/(2\beta m^2)$, i.e., for all $x, y \in C$,*

If $\langle F(x), y - x\rangle + \varphi(y) - \varphi(x) \geq 0$ holds, then

$$\langle F(y), y - x\rangle + \varphi(y) - \varphi(x) \geq \gamma\|F(y) - F(x)\|^2.$$

Then, the sequence $\{x^k\}_{k\in\mathbb{N}}$ is bounded. Moreover, $\lim_{k\to+\infty}\|x^{k+1} - x^k\| = 0$ and $\lim_{k\to+\infty}\|F(x^k) - F(x^*)\| = 0$.

Proof. Let x^* be the solution of problem (P). We consider the sequence of Lyapunov functions $\{\Gamma^k(x^*, \cdot)\}_{k\in\mathbb{N}}$ defined on H by

$$\begin{aligned}\Gamma^k(x^*, x) = {} & h(x^*) - h(x) - \langle \nabla h(x), x^* - x\rangle \\ & + (\mu_k/m)[\langle F(x^*), x - x^*\rangle + \varphi(x) - \varphi(x^*)].\end{aligned} \tag{4.3}$$

By the same process as in the proof of Lemma 3.1, we obtain (3.5), (3.6), and (3.9) such that

$$\begin{aligned}\Gamma^{k+1}(x^*, x^{k+1}) - \Gamma^k(x^*, x^k) \leq {} & \mu_k[\langle F(x^k), x^* - x^k\rangle + \varphi(x^*) - \varphi(x^k)] \\ & + (1/m)\,(1/2\tau)\,\mu_k^2\,\|F(x^k) - F(x^*)\|^2 - (1/2)(\beta - \tau/m)\,\|x^{k+1} - x^k\|^2,\end{aligned} \tag{4.4}$$

where τ is any positive constant.

Since $\langle F(x^*), x^k - x^*\rangle + \varphi(x^k) - \varphi(x^*) \geq 0$, we have that

$$\langle F(x^k), x^k - x^*\rangle + \varphi(x^k) - \varphi(x^*) \geq \gamma\|F(x^k) - F(x^*)\|^2.$$

Hence, since $\mu_k \leq \bar{\mu}$ for all $k \in \mathbb{N}$,

$$\begin{aligned}&\mu_k[\langle F(x^k), x^* - x^k\rangle + \varphi(x^*) - \varphi(x^k)] + (2\tau m)^{-1}\,\mu_k^2\,\|F(x^k) - F(x^*)\|^2 \\ &\quad \leq -\mu_k\gamma\|F(x^k) - F(x^*)\|^2 + (2\tau m)^{-1}\,\mu_k^2\,\|F(x^k) - F(x^*)\|^2 \\ &\quad \leq -\mu_k\,(\gamma - \bar{\mu}/2\tau m)\,\|F(x^k) - F(x^*)\|^2.\end{aligned}$$

Since $\gamma > \overline{\mu}/2m^2\beta$, we can choose τ such that $(\gamma - \bar{\mu}/2\tau m) > 0$ and $(\beta - \tau/m) > 0$. Consequently, (4.4) becomes

$$\Gamma^{k+1}(x^*, x^{k+1}) - \Gamma^k(x^*, x^k) \leq -c_1 \underline{\mu} \|F(x^k) - F(x^*)\|^2 - c_2 \|x^{k+1} - x^k\|^2, \tag{4.5}$$

with $c_1 = \gamma - \bar{\mu}/2\tau m > 0$ and $c_2 = (1/2)(\beta - \tau/m) > 0$.

It follows from (4.5) that $\{\Gamma^k(x^*, x^k)\}_{k\in\mathbb{N}}$ is a Cauchy sequence. Hence, it is convergent in H. Then using (3.2), we deduce that the sequence $\{x^k\}_{k\in\mathbb{N}}$ is bounded and, passing to the limit in (4.5), that the sequences $\{\|x^{k+1} - x^k\|\}_{k\in\mathbb{N}}$ and $\{\|F(x^k) - F(x^*)\|\}_{k\in\mathbb{N}}$ converge to zero. □

We can now state the main convergence result.

THEOREM 4.3. *Let us suppose that all assumptions of Theorem* 4.2 *are fulfilled such that the sequence* $\{x^k\}_{k\in\mathbb{N}}$ *is bounded. The following conclusions can be derived:*

1. *If* F *is weakly continuous on* C*, then each weak limit point of the sequence* $\{x^k\}_{k\in\mathbb{N}}$ *is a solution of problem* (P)*.*
2. *If* ∇h *is weakly continuous on* C*, then the whole sequence* $\{x^k\}_{k\in\mathbb{N}}$ *weakly converges to some solution of problem* (P)*.*
3. *If, moreover,* F *is strongly monotone on* C*, then* $\{x^k\}_{k\in\mathbb{N}}$ *strongly converges to the unique solution* x^* *of problem* (P)*.*

Proof. Conclusion 1 follows directly from Theorems 4.2 and 4.1. To prove conclusion 2, we have to show that the sequence $\{x^k\}_{k\in\mathbb{N}}$ has a unique weak limit point. Assume that $\{x^k\}_{k\in\mathbb{N}}$ has two weak limit points $\bar{x}$ and $\tilde{x}$. By conclusion 1, these two points are solutions of problem (P) and, from the proof of Theorem 4.2, the sequences $\{\Gamma^k(\tilde{x}, x^k)\}_{k\in\mathbb{N}}$ and $\{\Gamma^k(\bar{x}, x^k)\}_{k\in\mathbb{N}}$ are convergent. Let $\tilde{\Gamma}$ and $\bar{\Gamma}$ be their respective limits.

On the other hand, by the definition of the Lyapunov function, we have, for all $k \in \mathbb{N}$ and $x \in C$,

$$\begin{aligned}
&\Gamma^k(\tilde{x}, x) - \Gamma^k(\bar{x}, x) \\
&\quad = (h(\tilde{x}) - h(\bar{x}) - \langle \nabla h(x), \tilde{x} - \bar{x}\rangle) \\
&\qquad + (\mu_k/m)\,(\langle F(\tilde{x}), \bar{x} - \tilde{x}\rangle + \varphi(\bar{x}) - \varphi(\tilde{x}) + \langle F(\tilde{x}) - F(\bar{x}), x - \bar{x}\rangle).
\end{aligned}$$

Let $\{x^k\}_{k\in K\subset\mathbb{N}}$ be a subsequence of $\{x^k\}_{k\in\mathbb{N}}$ converging to $\bar{x}$. If we set $x = x^k$ in the above inequality, we can write

$$\begin{aligned}
&\Gamma^k(\tilde{x}, x^k) - \Gamma^k(\bar{x}, x^k) \\
&\quad = (h(\tilde{x}) - h(\bar{x}) - \langle \nabla h(x^k) - \nabla h(\bar{x}), \tilde{x} - \bar{x}\rangle - \langle \nabla h(\bar{x}), \tilde{x} - \bar{x}\rangle) \\
&\qquad + (\mu_k/m)(\langle F(\tilde{x}), \bar{x} - \tilde{x}\rangle + \varphi(\bar{x}) - \varphi(\tilde{x}) + \langle F(\tilde{x}) - F(\bar{x}), x^k - \bar{x}\rangle).
\end{aligned}$$

From the strong convexity of h and since $\tilde{x}$ is a solution of problem (P), we deduce that

$$\begin{aligned}
\Gamma^k(\tilde{x}, x^k) - \Gamma^k(\bar{x}, x^k) \geq\ & (\beta/2)\|\tilde{x} - \bar{x}\|^2 \\
& - \langle \nabla h(x^k) - \nabla h(\bar{x}), \tilde{x} - \bar{x}\rangle \\
& + (\mu_k/m)\langle F(\tilde{x}) - F(\bar{x}), x^k - \bar{x}\rangle.
\end{aligned}$$

Then, if we take the limit on $k \in K$, the weak continuity of ∇h implies that

$$\tilde{\Gamma} - \bar{\Gamma} \geq (\beta/2)\|\tilde{x} - \bar{x}\|^2.$$

Since the roles of $\bar{x}$ and $\tilde{x}$ can be reversed, we also have that

$$\bar{\Gamma} - \tilde{\Gamma} \geq (\beta/2)\|\bar{x} - \tilde{x}\|^2.$$

Combining these two inequalities, we conclude that $\bar{x} = \tilde{x}$, which proves the uniqueness of the weak limit point for $\{x^k\}_{k\in\mathbb{N}}$.

Let x^* denote this weak limit point. To obtain conclusion 3, we will show that when F is strongly monotone on C, we also have that $\|x^k - x^*\| \to 0$. If we put together relations (3.5), (3.6), (3.9) from the proof of Lemma 3.1, we obtain that

$$\begin{aligned}
&\Gamma^{k+1}(x^*, x^{k+1}) - \Gamma^k(x^*, x^k) \\
&\quad \leq -(\beta/2)\,\|x^{k+1} - x^k\|^2 \\
&\qquad + \mu_k\,[\langle F(x^k), x^* - x^k\rangle + \varphi(x^*) - \varphi(x^k) \\
&\qquad\quad + (1/m)\langle F(x^k) - F(x^*), x^k - x^{k+1}\rangle].
\end{aligned}$$

When F is strongly monotone on C (with constant $\bar{\alpha} > 0$), since x^* is a solution of problem (P), we have that

$$\langle F(x^k), x^k - x^*\rangle + \varphi(x^k) - \varphi(x^*) \geq \bar{\alpha}\|x^k - x^*\|^2.$$

We deduce that

$$\begin{aligned}
&\Gamma^{k+1}(x^*, x^{k+1}) - \Gamma^k(x^*, x^k) \\
&\quad \leq -(\beta/2)\,\|x^{k+1} - x^k\|^2 - \mu_k\,\bar{\alpha}\,\|x^k - x^*\|^2 \\
&\qquad + (\mu_k/m)\langle F(x^k) - F(x^*), x^k - x^{k+1}\rangle.
\end{aligned}$$

Let us now pass to the limit on k in this inequality. Since $\{\Gamma^k(x^*, x^k)\}_{k\in\mathbb{N}}$ is convergent, $\{\|x^{k+1} - x^k\|\}_{k\in\mathbb{N}}$ converges to zero, F is weakly continuous, and $0 < \underline{\mu} \leq \mu_k \leq \overline{\mu}$, we conclude that $\{\|x^k - x^*\|\}_{k\in\mathbb{N}}$ converges to zero. This completes the proof. □

Remark 3. If we particularize Theorem 4.3 to the finite dimensional case, we obtain the (strong) convergence of the sequence $\{x^k\}_{k\in\mathbb{N}}$ to some solution of problem (P) provided that all assumptions of Theorem 4.2 are satisfied.

Remark 4. When φ is not approximated, our results generalize those obtained by Zhu and Marcotte in [42]. When $F = 0$ and $C = H$, our scheme amounts to the bundle algorithm to minimize φ on H, and our convergence results reduce to well-known ones (see, e.g., [11]).

5. Numerical tests. The computational experience reported here has been performed with the software MATLAB. Five examples of operator F have been tested. The first example is $F = 0$ so that problem (P) amounts to minimizing φ over C. For the other examples, F is of the form $F(x) = Q\,x$, where Q is an $(n \times n)$ nonsymmetric matrix chosen in such a way that F satisfies the Dunn property, i.e.,

$$\exists\,\sigma > 0 \quad \text{such that for all } x \in \mathbb{R}^n,\ x^T\,Q\,x \geq \sigma\,x^T\,Q^T\,Q\,x.$$

In the examples, the matrices Q_1 and Q_3 are positive definite, while the matrices Q_2 and Q_4 are singular. The function φ is defined on $\mathbb{R}^n$ as the maximum of five

quadratic functions. This classical nonsmooth test function is taken from [23, Test Problem 1: MAXQUAD, p. 151]. Except when $F = 0$, where we take $C = \mathbb{R}^n$, the following constraint set is used in the other cases:

$$\mathcal{C} = \left\{ x \in \mathbb{R}^n : \sum_{i=1}^{n} x_i \geq 1, \quad -5 \leq x_i \leq 5,\ i = 1, \ldots, n \right\}.$$

In the bundle algorithm, we choose $h(x) = (1/2)\|x\|^2$ and

$$\theta^i = \max_{i_k \leq j \leq i-1} \{\varphi(y^j) + \langle s(y^j), \cdot - y^j \rangle\} \quad \text{for all } i : \ i_k < i \leq i_{k+1}.$$

Consequently, each subproblem (P_i^k) can be equivalently written under the following form:

$$(P_i^k) \begin{cases} \min_{x \in C, v \in \mathbb{R}} \{ v + \langle r(x^k), x - x^k \rangle + \frac{1}{2\mu_k} \|x - x^k\|^2 \} \\ \text{subject to } v \geq \varphi(y^j) + \langle s(y^j), x - y^j \rangle, \qquad j = i_k, \ldots, i-1. \end{cases}$$

Observe that, at the solution (y^i, v^i), we have that $v^i = \theta^i(y^i)$. Note also that subproblem (P_i^k) amounts to a quadratic minimization problem. Moreover, for fixed k, each quadratic subproblem is the previous one with an additional linear inequality constraint. The computation of the trial points y^i can thus be made very efficiently.

For the sake of clarity, we adapt the algorithm by taking into account the above considerations.

Bundle algorithm to solve problem (P). Let an initial point x^0 be given, together with tolerances $m \in]0, 1[$, $\epsilon > 0$, and a positive sequence $\{\mu_k\}_{k \in \mathbb{N}}$. Set $y^0 = x^0$ and $k = 0, i = 1, i_0 = 0$.

Step 1. Compute (y^i, v^i) as the unique solution of the quadratic problem

$$(P_i^k) \begin{cases} \min_{x \in C, v \in \mathbb{R}} \{ v + \langle F(x^k), x - x^k \rangle + \frac{1}{2\mu_k} \|x - x^k\|^2 \} \\ \text{subject to } v \geq \varphi(y^j) + \langle s(y^j), x - y^j \rangle, \qquad j = i_k, \ldots, i-1. \end{cases}$$

Step 2. If $\varphi(x^k) - \varphi(y^i) \geq m\,[\varphi(x^k) - v^i] + (1 - m)\,\langle F(x^k), y^i - x^k \rangle$, then set $x^{k+1} = y^i$.

If $\|x^{k+1} - x^k\| \leq \epsilon$, then **STOP**.
Otherwise set $i_{k+1} = i$ and increase k by 1.

Step 3. Increase i by 1 and go to Step 1.

Table 5.1 reports the results obtained with the following choices for the parameters: $(x^0)_i = 1$ for all $i = 1, \ldots, n$; $m = 0.4$; $\epsilon = 10^{-4}$; $\mu_k = \mu = 0.05$ for all k. From this table, we can observe that the number of quadratic programming subproblems per outer iteration is relatively small. The efficiency of the algorithm depends on the values taken for the parameters m and μ. Indeed, the stopping test in Step 2 will be all the more difficult to satisfy as m approaches 1. Our numerical experience shows that it becomes prohibitive relating to the cpu time if m is chosen too close to 1. A good compromise between the number of outer iterations and the number of inner iterations seems to be $m \approx 0.4$. Relating to the choice for μ, we note that if μ is chosen too large, too many null-steps are made between two serious-steps, and the cpu

TABLE 5.1
Number of inner iterations for each outer iteration denoted by k and cpu time in seconds.

	$n = 10$			$n = 20$	
k	$F = 0$ $C = \mathbb{R}^n$	$F(x) = Q_1 x$ $C = \mathcal{C}$	$F(x) = Q_2 x$ $C = \mathcal{C}$	$F(x) = Q_3 x$ $C = \mathcal{C}$	$F(x) = Q_4 x$ $C = \mathcal{C}$
1	3	3	3	5	5
2	3	3	3	6	6
3	4	4	4	9	9
4	4	5	5	8	8
5	6	6	5	9	9
6	7	7	7	12	12
7	8	8	8	12	11
8	8	9	8	43	12
9	9	9	8		16
10	16	11	10		
11	16	22	11		
12	11	13	12		
13	19	15	13		
14	13	17	14		
15	10	17	15		
16	14	18	16		
17	22	20	17		
18	14				
19	14				
20	20				
21	16				
cpu	37	34	29	41	36

time increases drastically. This comes from the fact that the term $(1/2\mu_k)\,\|x - x^k\|^2$ in subproblem (P_i^k) grows slowly when $\|x - x^k\|$ increases. In consequence, the trial points y^i are far from x^k, and the model is not suitable. On the contrary, when μ is chosen too small, too many serious (but small) steps are taken and the process stagnates. This can be justified by the fact that the term $(1/2\mu_k)\,\|x - x^k\|^2$ grows rapidly when $\|x - x^k\|$ increases such that the trial points y^i are close to x^k. The stopping test is then rapidly satisfied, but the corresponding serious-step is small. A good compromise seems to be $\mu \approx 0.05$.

Acknowledgments. The authors are particularly grateful to the associate editor, Claude Lemaréchal, and to two anonymous referees whose comments, questions, and suggestions greatly improved this paper.

REFERENCES

[1] Y. I. ALBER, A. N. IUSEM, AND M. V. SOLODOV, *On the projected subgradient method for nonsmooth convex optimization in a Hilbert space*, Math. Program., 81 (1998), pp. 23–35.
[2] G. AUCHMUTY, *Variational principles for variational inequalities*, Numer. Funct. Anal. Optim., 10 (1989), pp. 863–874.
[3] J. P. AUBIN AND I. EKELAND, *Applied Nonlinear Analysis*, Wiley, New York, 1984.

[4] A. Auslender, *Optimisation. Méthodes Numériques*, Masson, Paris, 1976.
[5] R. S. Burachik, C. A. Sagastizàbal, and B. F. Svaiter, *Epsilon–enlargement of maximal monotone operators with application to variational inequalities*, in Reformulation—Nonsmooth, Piecewise Smooth, Semismooth and Smoothing Methods, M. Fukushima and L. Qi, eds., Kluwer, Dordrecht, 1997, pp. 25–43.
[6] R. S. Burachik, C. A. Sagastizàbal, and B. F. Svaiter, *Bundle methods for maximal monotone operators*, in Ill–posed Variational Problems and Regularization Techniques, R. Tichatschke and M. Théra, eds., Springer, Berlin, 1999, pp. 49–64.
[7] R. D. Bruck, *An iterative solution of a variational inequality for certain monotone operators in Hilbert space*, Bull. American Math. Soc., 81 (1975), pp. 890–892.
[8] G. Cohen and D. L. Zhu, *Decomposition coordination methods in large scale optimization problems: The nondifferentiable case and the use of augmented Lagrangians*, in Advances in Large Scale Systems Theory and Applications, Vol. 1, J. B. Cruz, ed., JAI Press, Greenwich, CT, 1984, pp. 203–266.
[9] G. Cohen, *Nash equilibria: Gradient and decomposition algorithms*, Large Scale Systems, 12 (1987), pp. 173–184.
[10] G. Cohen, *Auxiliary problem principle extended to variational inequalities*, J. Optim. Theory Appl., 59 (1988), pp. 325–333.
[11] R. Correa and C. Lemaréchal, *Convergence of some algorithms for convex minimization*, Math. Program., 62 (1993), pp. 261–275.
[12] R. W. Cottle, F. Giannessi, and J. L. Lions, *Variational Inequalities and Complementarity Problems: Theory and Applications*, Wiley, New York, 1980.
[13] J. P. Crouzeix, *Pseudomonotone variational inequality problems: Existence of solutions*, Math. Program., 78 (1997), pp. 305–314.
[14] J. Eckstein, *Approximate iterations in Bregman–function–based proximal algorithms*, Math. Program., 83 (1998), pp. 113–124.
[15] I. Ekeland and R. Temam, *Convex Analysis and Variational Inequalities*, North–Holland, Amsterdam, 1976.
[16] R. Glowinski, J. L. Lions, and R. Trémolières, *Numerical Analysis of Variational Inequalities*, North–Holland, Amsterdam, 1981.
[17] P. T. Harker and J. S. Pang, *Finite-dimensional variational inequality and nonlinear complementarity problems. A survey of theory, algorithms and applications*, Math. Program., 48 (1990), pp. 161–220.
[18] A. N. Iusem, *On some properties of paramonotone operators*, J. Convex Anal., 5 (1998), pp. 269–278.
[19] D. Kinderlehrer and G. Stampacchia, *An Introduction to Variational Inequalities and Their Applications*, Classics in Appl. Math. 31, SIAM, Philadelphia, 2000.
[20] I. Konnov, *Combined Relaxation Methods for Variational Inequalities*, Lecture Notes in Econom. and Math. Systems 495, Springer-Verlag, Berlin, 2001.
[21] I. Konnov, *A combined relaxation method for a class of nonlinear variational inequalities*, Optimization, 51 (2002), pp. 127–143.
[22] B. Lemaire, *Coupling optimization methods and variational convergence*, in Trends in Mathematical Optimization, K. H. Hoffmann, J. B. Hiriart-Urruty, C. Lemaréchal, and J. Zowe, eds., Int. Ser. Numer. Math., Birkhauser-Verlag, Basel, 1988, pp. 163–179.
[23] C. Lemaréchal and R. Mifflin, *Nonsmooth Optimization*, IIASA Proceedings Series, Vol. 3, Pergamon Press, Oxford, 1978.
[24] P. L. Lions and B. Mercier, *Splitting algorithms for the sum of two nonlinear operators*, SIAM J. Numer. Anal., 16 (1979), pp. 964–979.
[25] S. Makler-Scheimberg, V. H. Nguyen, and J. J. Strodiot, *Family of perturbation methods for variational inequalities*, J. Optim. Theory Appl., 89 (1996), pp. 423–452.
[26] B. Martinet, *Régularisation d'inéquations variationnelles par approximations successives*, Rev. Française Automat. Inform. Rech. Opér., 4 (1970), pp. 154–159.
[27] M.A. Mataoui, *Contributions à la décomposition et à l'agrégation des problèmes variationnels*, Ph.D. thesis, Ecole des Mines de Paris, Fontainebleau, 1990.
[28] A. Moudafi and M. Noor, *New convergence results of iterative methods for set-valued mixed variational inequalities*, Math. Inequal. Appl., 3 (2000), pp 295–303.
[29] P. Panagiotopoulos and G. Stavroulakis, *New types of variational principles based on the notion of quasidifferentiability*, Acta Mech., 94 (1994), pp. 171–194.
[30] J. S. Pang, *Asymmetric variational inequality problems over product sets: Applications and iterative methods*, Math. Program., 31 (1985), pp. 206–219.
[31] D. Pascali and S. Sburlan, *Nonlinear Mappings of Monotone Type*, Sijthoff and Noordhoff, Alphen aan den Rijn, The Netherlands, 1978.

[32] B. T. Polyak, *Introduction to Optimization*, Optimization Software, New York, 1987.
[33] R. T. Rockafellar, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.
[34] R. T. Rockafellar, *On the maximal monotonicity of subdifferential mappings*, Pacific J. Math., 33 (1970), pp. 209–216.
[35] R. T. Rockafellar, *Monotone operators and the proximal point algorithm*, SIAM J. Control Optim., 14 (1976), pp. 877–898.
[36] G. Salmon, *Perturbed Auxiliary Problem Methods to Solve Generalized Variational Inequalities*, Presses Universitaires de Namur, Namur, Belgium, 2001.
[37] G. Salmon, V. H. Nguyen, and J. J. Strodiot, *Coupling the auxiliary problem principle and the epiconvergence theory to solve general variational inequalities*, J. Optim. Theory Appl., 104 (2000), pp. 629–657.
[38] G. Salmon, J. J. Strodiot, and V. H. Nguyen, *A perturbed auxiliary problem method for paramonotone multivalued mappings*, in Advances in Convex Analysis and Global Optimization, N. Hadjisavvas and P. Pardalos, eds., Nonconvex Optim. Appl. 54, Kluwer, Dordrecht, 2001, pp. 515–529.
[39] Y. Sonntag, *Convergence au sens de Mosco: Théorie et applications à l'approximation des solutions d'inéquations*, Ph.D. thesis, Université de Provence, 1982.
[40] P. Tseng, *Applications of a splitting algorithm to decomposition in convex programming and variational inequalities*, SIAM J. Control Optim., 29 (1991), pp. 119–138.
[41] D. Zhu, *The Decomposition Method for the Variational Inequalities with Multivalued Mapping*, private communication, Centre de Recherche sur les Transports, Université de Montréal, 1999.
[42] D. L. Zhu and P. Marcotte, *Co-coercivity and its role in the convergence of iterative schemes for solving variational inequalities*, SIAM J. Optim., 6 (1996), pp. 714–726.

SIAM J. OPTIM.
Vol. 14, No. 3, pp. 894–913

NEW PRECONDITIONERS FOR KKT SYSTEMS OF NETWORK FLOW PROBLEMS*

A. FRANGIONI† AND C. GENTILE‡

Abstract. We propose a new set of preconditioners for the iterative solution, via a preconditioned conjugate gradient (PCG) method, of the KKT systems that must be solved at each iteration of an interior point (IP) algorithm for the solution of linear min cost flow (MCF) problems. These preconditioners are based on the idea of extracting a proper triangulated subgraph of the original graph which strictly contains a spanning tree. We define a new class of triangulated graphs, called *brother-connected trees* (BCTs), and discuss some fast heuristics for finding BCTs of "large" weight. Computational experience shows that the new preconditioners can complement tree preconditioners, outperforming them both in iterations count and in running time on some classes of graphs.

Key words. min cost flow problems, interior point algorithms, preconditioned conjugated gradient method, triangulated graphs

AMS subject classifications. 90C51, 65F10

DOI. 10.1137/S105262340240519X

1. Introduction.

The linear min cost flow (MCF) problem is the following linear program (LP):

$$\min\{cx : Ex = b,\ 0 \le x \le u\}, \tag{1.1}$$

where E is the node-arc incidence matrix of a directed graph $G = (N, A)$, c is the vector of arc costs, u is the vector of arc upper capacities, b is the vector of node deficits, and x is the vector of flows. This problem has a huge set of applications, either in itself or, more often, as a submodel of more complex and demanding problems [1]. This is evidenced by the enormous amount of research that has been devoted to developing efficient solution algorithms for MCF problems [1], either by specializing LP algorithms—such as the simplex method—to the network case, or by developing ad hoc approaches.

Recently, interior point (IP) methods for linear programming have established a reputation as efficient algorithms for large-scale problems; a detailed description of the IP algorithms and their underlying theory can be found in the extensive literature on the subject and in many recent linear programming textbooks, e.g., [19, 24]. At each iteration of these methods, linear systems of the form

$$\left(E\Theta E^T\right)\Delta y = d \tag{1.2}$$

have to be solved, where Θ and d are an $m \times m$ diagonal matrix ($m = |A|$) with positive entries and a vector of $\mathbb{R}^n$ ($n = |N|$), respectively, which depend on the current solution and on the IP algorithm chosen. These systems are often referred to as

*Received by the editors April 8, 2002; accepted for publication (in revised form) August 29, 2003; published electronically March 23, 2004. This work has been partially supported by project CNRG00A4E0 "Interior Point Methods for Structured Linear Programs" of program CNR-Agenzia2000 of the Italian National Research Council (CNR).

http://www.siam.org/journals/siopt/14-3/40519.html

†Dipartimento di Informatica, Università di Pisa, Via Buonarroti 2, 56125 Pisa, Italy (frangio@di.unipi.it).

‡Istituto di Analisi dei Sistemi ed Informatica "Antonio Ruberti" del CNR, Viale Manzoni 30, 00185 Roma, Italy (gentile@iasi.rm.cnr.it).

KKT systems, because they represent the computational core of a "slackened" KKT system for the problem. Although the form (1.2) is not, strictly speaking, the most general, it has the advantage of being the same for many variants of IP algorithms. Furthermore, in the MCF case, the matrix $M = E\Theta E^T$ has close relationships with several extensively studied objects in both linear algebra and graph theory. When $\Theta = I$, the matrix M is closely related to the *Laplacian* of the *undirected* version of G [2, 7], which has been exploited to explore topological properties of graphs through the spectral properties of some associated matrices [6]. Conversely, the graph G can be thought of as a combinatorial representation of certain algebraic properties of M [20], some of which will be recalled below.

The solution of (1.2) typically represents by far the main computational burden of IP algorithms. Thus, developing a specialized approach for the solution of (1.2) for specially structured matrices E can substantially improve the performance of an IP method. Since the form of the KKT system is independent of the specific variant of IP algorithm used, the same specialized solver for (1.2) can be used to implement all the variants of IP algorithms.

As $M = E\Theta E^T$ is symmetric and positive semidefinite, (1.2) is often solved through a Cholesky factorization, which is computationally effective and numerically stable. That is, a lower triangular *Cholesky factor* L with all diagonal entries equal to 1 and a diagonal matrix D with a positive (nonnegative) diagonal are found such that $M = LDL^T$; this can be done in $O(n^3)$, and, once the factorization has been computed, systems involving M can be solved in $O(n^2)$ with two backsolves on L. However, a well-known drawback of the Cholesky factorization is the *fill-in* phenomenon: a sparse matrix M may have a dense Cholesky factor L. The density of the Cholesky factor may vary by reordering the rows of the matrix E; hence, IP codes usually make an effort at finding a permutation of the rows of E which (approximately) minimizes the fill-in effect. This is only done at the beginning of the algorithm, since the structure of the nonzeros in M (and therefore of its Cholesky factor) does not depend on Θ and therefore does not change with the iterations. The problem of finding the reordering which produces the least fill-in is known to be $\mathcal{NP}$-hard [25]; however, several effective heuristics have been developed for computing a "good" such permutation [19]. Yet, in general the fill-in phenomenon cannot be avoided [4] except in some specific cases, so that alternative methods have been proposed for MCF [17, 15, 13, 14, 23] and other network-structured problems [4]. Most of these methods solve the system using a preconditioned conjugate gradient (PCG) method. The critical choice is therefore that of the preconditioner: it must be inexpensive to compute and invert while delivering a consistent reduction of the number of conjugate gradient iterations required to (approximately) solve (1.2).

The first PCG-based IP algorithm specifically tailored for MCF problems was proposed in [17]. Following suggestions from [12] and [22], the *tree preconditioner* was defined, which is a preconditioner of the form

$$M_S = E_S \Theta_S E_S^T, \tag{1.3}$$

where S is a spanning tree of G, E_S is the node-arc incidence matrix of S, and Θ_S is the restriction of Θ to the arcs in S. In particular, S is chosen as an (approximate) maximum-weight spanning tree, the weight of each arc (i, j) being the corresponding θ_{ij}. Such a tree can be constructed in $O(m)$ with a variant of the classical Kruskal algorithm where arcs are only approximately sorted using a "bucket" data structure

with m buckets. The linear systems involving M_S can then be solved in $O(n)$, at each step of the PCG method, by considering the three linear systems with coefficient matrix E_S, Θ_S, and E_S^T, respectively; it is well known [1] that these systems can be solved by visiting the tree S. The preconditioner M_S can be expected to be spectrally effective, especially in the final iterations of an IP algorithm; in fact, the analysis of IP methods shows that, if the optimal solution of the underlying MCF is unique, the weights θ_{ij} tend to zero on all arcs, except on those corresponding to the basic optimal solution [19] that form a spanning tree; hence $M_S \approx E\Theta E^T$ in the last iterations of the IP method. The analysis in [10] and the experimental results show that the tree preconditioner in fact has good spectral properties in the final iterations of an IP algorithm even in the degenerate case. Finally, a different rationale for the choice of S as a maximum-weight spanning tree has been given in [7].

Unfortunately, tree preconditioners are less effective in the first iterations; this has suggested a hybrid preconditioning technique [17], where the diagonal preconditioner is used in the first iterations, and then some heuristic rules are used to switch to the tree preconditioner in a later stage. The implementation of this approach, refined with better stopping criteria [18] and a custom primal-infeasible/dual-feasible IP algorithm [15], has shown to be competitive with well-known combinatorial MCF codes.

In [13], the tree preconditioner is "extended" by using

$$M'_S = M_S + \rho\,\mathrm{diag}(M - M_S) \tag{1.4}$$

as the preconditioner, where $\mathrm{diag}(X)$ is the diagonal matrix having as the diagonal elements those of X. This has the advantage of incorporating information about all arcs, rather than about only those in S. The parameter ρ can be chosen according to the structure of the MCF problem at hand, with different values proposed in [13] for different classes of MCF problems. The relationships between M'_S and M_S, from the spectral viewpoint, have been analyzed in [10]. Finally, a different preconditioner has been proposed in [14] for the special case of transportation problems, based on an incomplete QR factorization of M, that has been reported as being more effective than the tree preconditioner for this particular class of MCF instances in the early iterations of the algorithm. For a more detailed description of these preconditioners and their relationships the interested reader is referred to [16] and [10].

Our aim is to improve the effectiveness of IP methods for MCF problems by designing new classes of preconditioners. The basic idea is that of extracting a proper subgraph $S = (N, A_S)$ of G ($A_S \subseteq A$) which contains—possibly strictly—a spanning tree, but such that the corresponding matrix M_S defined as in (1.3) can still be efficiently factored. We will refer to these preconditioners as *subgraph based*, and to S as the *support* of M_S. One way for ensuring efficient factorization is to select S as a *triangulated* (also known as *chordal*) graph [20], so that there exists an ordering of the nodes for which M_S has no fill-in. Other ideas can then be exploited for further improving the effectiveness of these preconditioners, yielding a large variety of preconditioners, some of which provide a better trade-off between the cost of finding S and factoring M_S and the cost of the PCG iterations.

The structure of the paper is the following: in section 2 we introduce and prove the properties of a large family of new preconditioners. In sections 3 and 4 the algorithmic issues related to this new family of preconditioners are discussed. In section 5 the results of a computational experience aimed at assessing the effectiveness of the new preconditioners are presented. Finally, conclusions are drawn in section 6.

2. Subgraph-based preconditioners. We propose to choose S as a *triangulated* graph [20], i.e., such that every cycle of length at least 4 has an edge joining two nonconsecutive vertices in the cycle. Such an edge is called a *chord*, whence the alternative name of *chordal* graphs. Since M_S has the same nonzero structure of the node-node adjacency matrix of S, there exists a "good" ordering of the nodes, i.e., an $n \times n$ permutation matrix P_n, such that the reordered matrix $P_n M_S P_n^T$ has a Cholesky factorization without fill-in. This is in fact a generalization of the result that is exploited when tree preconditioners are used: a P_n exists such that $P_n E_S$ is lower triangular. For the case of trees, P_n corresponds to any permutation $\mathcal{P}$ of the nodes such that if (i, j) is an arc of S with i father of j, then row j precedes row i in $\mathcal{P}$; these permutations include reverse depth-first visit and reverse breadth-first visit. Note that the definition of a father-son relationship implies that a root has been chosen for the spanning tree.

Thus, a natural way to generalize the tree preconditioner would be to choose S as a maximum-weight triangulated subgraph of G. Unfortunately, this does not appear to be an easy problem; although no conclusive evidence is known, the problem is conjectured in [11] to be $\mathcal{NP}$-hard.

However, choosing a maximum-weight triangulated subgraph of G, even if it were computationally feasible, would not necessarily be a good idea in this application. This is due to the fact that, as shown in section 5, for MCF problems the tree preconditioner is already very effective, and only a limited (although sizable) increase of the spectral efficiency of the method can be expected, especially in the last IP iterations. Thus, it is crucial that the extra cost of finding and factoring a "fatter" preconditioner M_S is kept low for the approach to have some chance of improving on the tree preconditioner. Indeed, the most efficient implementations of IP methods for MCF based on the tree preconditioner use an *approximate* algorithm for finding the maximum-weight spanning tree, even though the optimal tree could be found in (low) polynomial time.

Hence, a generalization of the tree preconditioner is sought for finding a large-weight triangulated graph with only slightly more effort than that required for finding an approximate maximum-weight spanning tree. We remark here that, for our application, finding the graph S is not enough; the "good" permutation $\mathcal{P}$ also has to be provided. This can always be done in linear time [21], but in general it is not free.

Not much along these lines has been done in the literature. In [11], a class of triangulated graphs, the k-windmills, is defined in the context of finding the "best" Markov network model of manageable size which "explains" some observed data, a problem that can be recast as that of finding a maximum-weight triangulated subgraph with "small" cliques of a given graph. An approximation algorithm with guaranteed performance is given for the maximum-weight k-windmill problem, but the algorithm requires the solution of an LP and a rounding operation and is therefore not feasible for our application.

A different way to achieve similar results has been proposed in the more general setting of *M-matrices*; the preconditioners are constructed by adding "a few" extra arcs to a spanning tree T, carefully balancing the extra cost of the incurred fill-in with the gain in the number of iterations [22, 3, 5]. This can be done, e.g., by splitting the node set into a small number k of disjoint components of size about n/k, each one spanned by a subtree of T, and then adding to S the arc with largest weight connecting any two of the components. The approach in [9] is similar although more involved and is mostly motivated by the need for finding a preconditioner that parallelizes well: since the graph is recursively subdivided into smaller graphs of roughly equal

size, the workload can be evenly divided among parallel processors. In both cases, a small number of components ensures a "limited" increase in the cost for factoring the preconditioner, given that fill-in is expected.

2.1. Brother-connected trees. We now define a new family of preconditioners of the form (1.3), based on the characterization of a new class of triangulated graphs, strictly containing spanning trees.

DEFINITION 2.1. *A subgraph $S=(N, A_S)$ of G is a* brother-connected tree *(BCT) if it is either a spanning tree $T = (N, A_T)$ of G or it contains a spanning tree T of G such that the subgraph $S' = (N, A_S \setminus A_T)$ obtained by removing all the arcs of T from S is formed of a certain number $k \geq 1$ of node-disjoint connected components $S'_1 = (N_1, A_1), \ldots, S'_k = (N_k, A_k)$ such that all the nodes in N_i are "brothers" (sons of the same node) in T, and each S'_i is a BCT.*

Definition 2.1 is inherently recursive and operational in nature; a BCT can be constructed by iteratively taking a family of BCTs (which may be ordinary trees) and joining all their nodes in a tree, where all the nodes of any one of the original BCTs are sons of the same node. Note that, conversely, it is *not* required that all the sons of the same node in T belong to the same connected component. In particular, the connected components can be composed by only one node; in this case, we consider the empty set of arcs to be a spanning tree (and, therefore, a BCT) for the component. In other words, the arc set A_S of a BCT S is the union of the arc sets of a family $\mathcal{T} = \{T_1, \ldots, T_q\}$ of arc-disjoint subtrees T_i of G. The family $\mathcal{T}$ itself has a tree structure, where a tree T_i is the son of a tree T_j in $\mathcal{T}$ if all the nodes in T_i are brothers in T_j.

Thus, an important characteristic of a BCT S is its *depth*, which is the depth of the associated tree $\mathcal{T}$, i.e., the number of times that the composition operation has to be applied, starting from an empty graph, in order to construct S. A BCT of depth 1 is an ordinary tree, a BCT of depth 2 contains a spanning tree T such that the removal of all the arcs in T leaves a forest, and so on. For example, consider the graph of Figure 2 in section 3: solid arcs define T, dashed arcs are the forest at the second level, and dotted arcs do not belong to the BCT. The BCT depicted on the left side of Figure 2 has a family $\mathcal{T} = \{T_1, T_2, T_3\}$, where $T_1 = T$ are the solid arcs, T_2 is composed of the dashed arcs linking nodes 1, 2, 3, 4, and 5, and T_3 is only the dashed arc $(6, 7)$. The tree structure of $\mathcal{T}$ is simply that T_1 is father of both T_2 and T_3; therefore, the depth of the BCT is 2.

It is easy to show that BCTs are triangulated graphs by induction on the depth. A BCT of depth 1 is a tree, hence a triangulated graph. When building a BCT of depth $k+1$ out of a number of disjoint BCTs of depth at most k, all newly created cycles have length 3. Thus, there exists a permutation $\mathcal{P}$ of the nodes that allows us to factor E_S without fill-in if S is a BCT. Something more, however, can be said: the "good" ordering is "embedded" in the description of the BCT in terms of the associated tree $\mathcal{T}$. Thus, if the description is—as in the case of the heuristics that we propose—available "for free," then the BCT immediately provides all the necessary information for factoring the associated preconditioner without fill-in. This is what we are going to prove in the following.

A well-known property of the Cholesky factorization is that, given a matrix M' with Cholesky factor L', any symmetric positive definite matrix

$$M = \begin{bmatrix} M' & m \\ m^T & \mu \end{bmatrix} \text{ has a Cholesky factor of the form } L = \begin{bmatrix} L' & 0 \\ l^T & 1 \end{bmatrix}.$$

Furthermore, the values in a row of the Cholesky factor L depend only on the values on the same row and on the previous ones. Therefore, if M' is a matrix with no fill-in, then M can have fill-in only in its final row.

In graph terms, the above operation corresponds to adding to the graph representing the nonzero structure of the matrix M a new node, possibly connected with all other nodes. Therefore, the following result easily follows.

LEMMA 2.2. *Consider any finite number $k \geq 1$ of node-disjoint triangulated graphs $G_i = (N_i, A_i)$ and their corresponding "good" orderings $\mathcal{P}_i$; the graph $G = (N, A)$ obtained as the union of all the graphs G_i plus a new node u linked by an arc to each node in each of the graphs G_i is triangulated, and the corresponding "good" ordering $\mathcal{P}$ is obtained by composing the permutations $\mathcal{P}_i$ in arbitrary order and placing the new node u as the last node in the ordering.*

Proof. Apply the above observations: M' corresponds to all the G_i, and the new row l in the Cholesky factor of M is dense (completely nonzero), but this corresponds to the fact that S has $n-1$ arcs more than S'; i.e., the row $[m^T \mu]$ is completely nonzero too. □

We can now prove the main result.

THEOREM 2.3. *Given a brother-connected tree S in G, its representation as a tree $\mathcal{T}$ allows us to compute a "good" ordering of the nodes of G (such that M_S has a Cholesky factor L with no fill-in).*

Proof. We will proceed by double nested induction: the first on the depth of S, the second on the number of nonterminal nodes in the tree T contained in the BCT.

The basic step of the (outer) induction corresponds to depth 1; i.e., S is a tree. As we already recalled, any ordering of the nodes such that node j precedes node i if (i, j) is an arc of S and i is the father of j has the desired property. This ordering can be found in linear time.

For the inductive step, we assume that the ordering is available for any BCT with depth at most h and show how to construct it for BCTs of depth $h+1$. Once again we proceed by induction, this time on the number of nonterminal nodes of the spanning tree T included in S.

The basic step of the (inner) induction corresponds to the case where there is only one nonterminal node u; i.e., T is a "star tree," where any other node but u is a leaf. Since S is a BCT, the subgraph S' obtained by removing u (and all its incident arcs) from S is formed of $k \geq 1$ node-disjoint BCTs *of depth at most h*. Therefore, for the (outer) inductive hypothesis we know a good ordering for S', and we can find the one for S as shown in Lemma 2.2.

For the (inner) inductive step, consider a nonterminal node u such that all its sons are leaves of T; call V the set of the sons of u. Let S' be the subgraph of S induced by the nodes $V \cup \{u\}$ and let S'' be the subgraph of S induced by the nodes in $N \setminus V$; note that both subgraphs contain node u. Now we can apply the (inner) inductive hypothesis to S'', as we have reduced the number of nonterminal nodes by one unit; hence, we can find a good ordering $\mathcal{P}''$ of $N \setminus V$. Furthermore, since S is a BCT, then $S' \setminus \{u\}$ is a set of node-disjoint BCTs of depth at most h, and, as in the basic step of the (inner) induction, we can find a good ordering $\mathcal{P}'$ of $V \cup \{u\}$, where u must be the last node. We can then construct an ordering $\mathcal{P}$ of N by simply joining $\mathcal{P}'$ and $\mathcal{P}''$ in a sequence that corresponds to $\mathcal{P}'$ on V and to $\mathcal{P}''$ on $N \setminus V$ such that all the nodes in V precede those in $N \setminus V$. Therefore, the corresponding reordered M_S can be written as

$$M_S = \left[\begin{array}{c|c|c|c} M_{S'\setminus\{u\}} & 0 & m & 0 \\ \hline 0 & 0 & 0 & 0 \\ \hline m^T & 0 & \mu & 0 \\ \hline 0 & 0 & 0 & 0 \end{array}\right] + \left[\begin{array}{c|c|c|c} 0 & 0 & 0 & 0 \\ \hline 0 & & & \\ \hline 0 & & M_{S''} & \\ \hline 0 & & & \end{array}\right],$$

where $[m^T, \mu]$ is the (completely nonzero) row corresponding to node u in the matrix $M_{S'}$. The two matrices in the right-hand side share only one nonzero position in the row and column associated with u. Hence, the first part of the Cholesky factor L of M_S is equal to that of $M_{S'}$, and so it is the part of the factorization relative to the nondiagonal elements in row u. Thus, the Cholesky factorization of (the reordered) M_S in the first $|V|$ rows/columns has no fill-in. As S' is a graph, the associated matrix $M_{S'}$ is rank deficient and the value d'_u for its LDL^T factorization is zero. Hence, the value d_u for the factorization of M_S is equal to d''_u computed in the factorization of $M_{S''}$. Therefore, the second part of the factorization of M_S is exactly the factorization of $M_{S''}$. For the inductive hypothesis we know that the Cholesky factor of $M_{S''}$ (reordered with $\mathcal{P}''$) has no fill-in, and this finally allows us to conclude that $\mathcal{P}$ is a good ordering for M_S. □

The above result can be easily generalized with the following proposition.

PROPOSITION 2.4. *Let M be a positive definite matrix with a BCT support; then, the ordering of Theorem* 2.3 *is "good" for M.*

Proof. In the general case, the computation of the element d_u in the proof of Theorem 2.3 may depend on the submatrix associated with S'. Let d'_u and d''_u be the values computed in the factorization of $M_{S'}$ and $M_{S''}$, respectively. The first part of the factorization of M_S is equal to the factorization of $M_{S'}$ for rows in $S' \setminus \{u\}$, and the rest can be obtained as the factorization of $M_{S''}$, but starting from the value $d_u = d'_u + d''_u$. Therefore, M can be factored without fill-in. □

To summarize, for a BCT with depth 2, $\mathcal{P}$ must be such that

- The matrix E_T associated with its spanning tree T is lower triangular; i.e., if (i, j) is an arc of T with i father of j, then row j of E_S precedes row i in $\mathcal{P}$;
- For each nonterminal node u of T, each subset of its sons which belong to the same subtree $T_h = S'_h$ (once the arcs of T have been removed) is ordered in the permutation according to the order defined by T_h; i.e., if (i, j) is an arc of T_h with i father of j, then row j precedes row i in $\mathcal{P}$. The roots of the subtrees and the sequence of the trees can be arbitrarily chosen.

In general, $\mathcal{P}$ can be recursively constructed by ordering the nodes of the BCTs of depth h and then composing these orders into orders for the BCTs of depth $h+1$. This can be done with a bottom-up postvisit of the tree $\mathcal{T}$ associated with the BCT, i.e., by visiting the tree $\mathcal{T}$ from the leaves to the root with the constraint that each node of $\mathcal{T}$ can be visited only after all of its sons.

The induction process in Theorem 2.3 suggests an algorithm that performs the Cholesky factorization of M_S without fill-in in $O(nh^2)$, where h is the depth of the BCT. All the trees at the same depth q can be represented with a unique predecessor function $Pred[q]$ defined on the nodes, such that $Pred[q][u] = v$ if v is the father of u at depth q, and $Pred[q][u] = nil$ (null value) if u is a root, i.e., it has no father. For instance, in a BCT of depth 2 the function $Pred[1]$ represents the spanning tree T whose removal leaves a forest F, which is represented by the function $Pred[2]$. The algorithm for computing the LDL^T factorization of a matrix M_S with a BCT support S is shown with the pseudocode in Figure 1. It requires a description of the BCT (of depth h) in terms of the predecessor functions $Pred[q]$, and a description

```
Procedure CholeskyFactorBCT(h, n, M, Pred, Order, L, D)
  begin
    for i = j, ..., n; i = 1, ..., j − 1 do
      L[i, j] = M[i, j];
    for i = 1 ... n do
      D[i] = M[i, i];
      L[i, i] = 1;
    for i = 1 ... n − 1 do
      u = Order[i];
      for q = h ... 1 do
        w = Pred[q][u];
        if w ≠ nil then
          L[w, u] = L[w, u]/D[u];
          D[w] = D[w] − L[w, u]^2 ∗ D[u];
      for q = h ... 1 do
        w = Pred[q][u];
        if w ≠ nil then
          for r = q − 1 ... 1 do
            y = Pred[r][w];
            if y ≠ nil then
              L[y, w] = L[y, w] − L[w, u] ∗ L[y, u] ∗ D[u];
  end.
```

FIG. 1. *Pseudocode for factorization of a matrix with BCT support.*

of a "good" ordering $\mathcal{P}$ in an array $Order[]$. By performing a bottom-up visit of the tree $\mathcal{T}$, it outputs the Cholesky factor L and the diagonal matrix D. The algorithm is similar to the usual procedure for the Cholesky factorization, but it exploits the fact that the fill-in cannot be produced, so nonzero elements of L correspond to pairs (y, w) such that $y = Pred[s][w]$ for some level s. Indeed, the Cholesky factorization using the ordering $\mathcal{P}$ would be

$$L[y,w] = \frac{1}{D[w]}\left[M[y,w] - \sum_{u <_{\mathcal{P}} w} L[w,u]L[y,u]D[u]\right],$$
$$D[w] = M[w,w] - \sum_{u <_{\mathcal{P}} w} L[w,u]^2 D[u],$$

where "$<_{\mathcal{P}}$" is the ordering contained in $Order[]$.

Using the same data structures, $Pred$ and $Order$, an $O(nh)$ algorithm that solves systems of the form $M_S r = v$—which is what is actually required if M_S is used as a preconditioner in a PCG algorithm—can be constructed; any iteration of a PCG algorithm which uses a BCT-based preconditioner has a complexity of $O(nh + m)$. In our implementation, we have considered only the case of BCTs of depth 2; this simplifies and streamlines the algorithms, while still leaving room for almost doubling the number of arcs to be put in S with respect to a standard tree preconditioner (a BCT of depth 2 can have up to $2n - 3$ arcs).

Thus, BCT preconditioners can be easily integrated with standard tree preconditioners, and they do not need general-purpose Cholesky factorization routines; in fact, the construction and factorization of the preconditioner are easily and efficiently per-

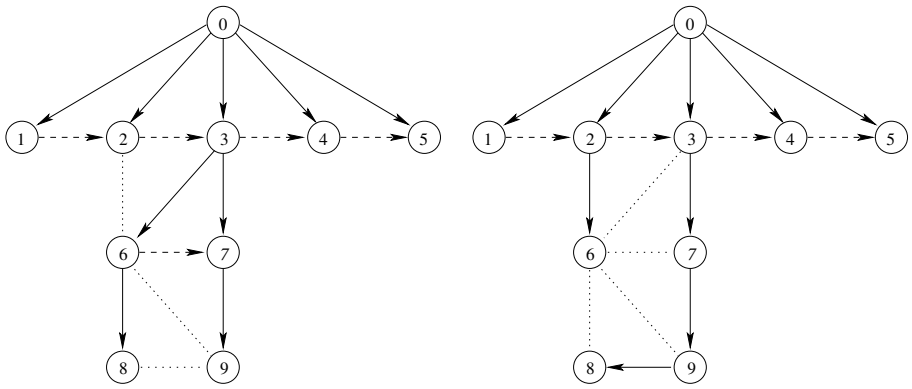


FIG. 2. *Two maximal BCTs on the same graph with different cardinality.*

formed using the data structures naturally produced by the construction of the BCT.

3. Finding brother-connected trees. The complexity of the problem of finding the maximum-weight BCT in a given graph G is not known to us. However, the exact solution of this problem is not crucial in this application; even in the case of tree preconditioners, an approximate solution is usually preferred although the exact solution can be obtained in low polynomial time. It is very unlikely that a maximum-weight BCT can be found with a comparable efficiency, because BCTs are not matroids. This is shown by the two BCTs S_1 and S_2 in Figure 2, where the solid arcs belong to the first level, the dashed ones belong to the second level, and, finally, the dotted ones are in the complements $G \setminus S_1$ and $G \setminus S_2$. It is easy to check that S_1 and S_2 are maximal BCTs (of level two) with different cardinality.

3.1. A class of heuristics for maximum-weight BCT. For all the above reasons, we will resort to heuristics to find the BCT to be applied in the PCG method. A large number of different heuristics can be proposed, by combining different variants of two basic ingredients:

(i) how a spanning tree T is chosen;
(ii) how extra arcs forming trees among brothers in T are chosen.

Some results can be proved about the worst-case performances of this kind of heuristics if T is chosen to be a maximum-weight spanning tree for the graph.

PROPOSITION 3.1. *Let G be a graph; denote by $w(MBCT)$ the weight of the maximum-weight BCT with depth 2 on G and by $w(MST)$ the weight of the maximum weight spanning tree on G. Then $w(MBCT) \leq 2w(MST)$.*

Proof. Consider the following problem: given a graph G, find a connected subgraph $S = (N, A_S)$ of maximum weight with the property that S contains at least a spanning tree $T = (N, A_T)$ of G such that the residual graph $S \setminus T = (N, A_S \setminus A_T)$ is acyclic. Obviously, this problem is a relaxation of the maximum-weight BCT with depth 2. Moreover, its optimal objective function value is less than $2w(MST)$: in fact, $w(MST)$ is an upper bound on both $w(T)$ and $w(S \setminus T)$ as the latter one is acyclic. □

COROLLARY 3.2. *All heuristics for constructing a BCT which augment the maximum-weight spanning tree are 2-approximated.*

Thus, choosing the initial tree as an (approximate) maximum-weight spanning tree appears to be a promising choice. In fact, we have experimented with several

different ways for finding an initial spanning tree, described in [8] and not reported here to save on space, but they were almost invariably outperformed by the "standard" maximum-weight spanning tree.

The above bound is asymptotically tight even if we find the maximum-weight spanning tree T on G and then compute a maximum-weight spanning tree on each connected component induced by the sets of brothers in T, as the following example shows.

Example 3.3. Let us consider the graph with n nodes and the following two types of arcs:

- $(i, i+1)$ for $i = 1, \ldots, n-1$ with weight 1;
- $(1, j)$ for $j = 3, \ldots, n$ with weight $1 - \epsilon$.

Clearly, the maximum spanning tree T is the path from 1 to n composed of the arcs of the first type; hence, there are no brothers in T and the heuristic stops. However, the whole graph is a BCT of depth two, with the arcs connecting 1 with each of the other nodes in the first level and the other arcs in the second level with total weight $2n - 3 - (n-2)\epsilon$.

3.2. Enlarging the tree to a BCT. When a tree T is selected, extra arcs forming trees among brothers in T must be chosen (point (ii)). For this task we propose three different variants:

(ii.a) When the tree is selected, the final ordering of the nodes to be considered in the factorization is also arbitrarily fixed as any "good" ordering for T. Then, the arcs out of T are scanned in (approximated) order of decreasing weight and added to the tree if they are compatible with the fixed ordering and they satisfy the condition that the trees on the second level are paths among brothers.

(ii.b) As in case (ii.a), the arcs are scanned in (approximated) order, and the trees in the second level of the BCT are restricted to being paths; however, the ordering between brothers can be changed. The final ordering is found by considering one of the two possible permutations for each path among brother nodes, and then composing these orders, respecting the tree structure of T.

(ii.c) This variant is analogous to case (ii.b), but the trees in the second level of the BCT are not restricted to being paths.

These three variants require different data structures and amounts of computational time (how many times the list of arcs is scanned), and they find different BCTs. Variant (ii.a) is the cheapest one, but it usually adds fewer new arcs. Variant (ii.c) is the most complex, as it requires a new union-find structure to find trees in the second level and to select the corresponding orders, but it may add more arcs. Variant (ii.b) is something in between.

For variants (ii.b) and (ii.c), it is actually possible to modify the original spanning tree T as the algorithm proceeds, in order to add even more arcs. One way to do that is to apply an operation, which we call *promotion*, whereby a node connected with its grandfather is "promoted" as a brother of its former father. That is, let j be a node, k its father in T, and i the father of k in T. If the arc (i, j) is selected from the (approximated) ordering, it is possible, under some conditions, to modify the tree T in such a way that j becomes a son of i and brother of k. This is done by making (i, j) an arc of T (i.e., in the first level of the BCT), while (k, j) becomes an arc of the second level, as shown in Figure 3. In order to apply the promotion, node j must not have incident arcs (j, l) in the second level of the BCT, as after the promotion j and l are no longer brothers. Moreover, for variant (ii.b) the node k must also have at most one connected brother in the second level of the BCT, for otherwise

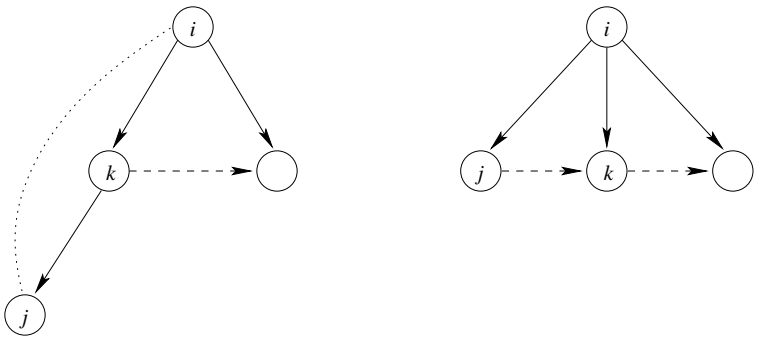


FIG. 3. *The BCT before (left) and after (right) the promotion.*

the tree in the second level would no longer be a path. Note that using the promotion operation in Example 3.3 allows one to discover that the complete graph is indeed a BCT.

In all the above heuristics, an initial ordering of the nodes is assumed that is "good" for the initial tree T; this is done by selecting a root node and performing a visit of the tree. Since this order impacts the heuristics (especially (ii.a), where it is fixed), the selection of the root node is potentially critical. We considered two possible strategies for selecting the root node: choosing the node with the largest adjacency list ("static") or choosing the node with the largest total weight of the set of incident arcs ("dynamic").

Let us remark here that the matrix $M = E\Theta E^T$ has rank equal to n minus the number of connected components in G, i.e., at most $n-1$. It is always possible to assume that G is connected, as otherwise the original MCF problem can be partitioned into a set of smaller subproblems, one for each of the connected components; hence, we can assume that the rank of M is $n-1$. When solving the KKT system, it is therefore possible to work with full-rank matrices by just deleting one row of E; alternatively, it is possible to work with the rank-deficient KKT system, although in this case M_S is rank deficient, too. The choice of the row (node) to be eliminated is arbitrary, yet it may have some consequences; when a node (row of the matrix E) is deleted, we choose it as the one associated with the root node of the tree T, although in principle different choices would be possible.

4. Further improvements. All the preconditioners that we have proposed so far can be further improved by applying two kinds of operations that attempt to incorporate in M_S information regarding arcs which have been left out of the support S.

The first operation amounts to adding to S all arcs (i, j) which are "parallel" to arcs already belonging to S, i.e., every other arc (i, j) or (j, i) belonging to G; we will denote these as "tree/BCT+parallel," or "T/BCT+P" for short, preconditioners. Clearly, this cannot generate fill-in other than that already present in the original M_S, as the support of the two matrices is the same. Note that "parallel" arcs, i.e., multiple copies of the same arc (or of its reverse arc) with different costs and capacities, are often present in MCF problems, e.g., to model piecewise linear convex separable flow cost functions [1]. This kind of operation has not been explicitly described before in the literature of IP approaches for MCF problems, while it is taken for granted when *M-matrices* are approached; our computational experience shows that the option has to be kept open. In fact, when "many" parallel arcs are present, it is useful to set the

weight of each edge $\{i, j\}$ in the MST computation equal to the sum of the weights of all parallel arcs (i, j) or (j, i), in order to correctly estimate the importance of adding any of those parallel arcs (and, therefore, all the others) to the support. However, when "few" parallel arcs are present, the extra computational burden required for computing the weights of the edges is not worth the corresponding improvement in the PCG convergence.

The second operation, proposed in [13] for the tree preconditioner, consists of using as preconditioner the matrix

$$M'_S = M_S + \rho \operatorname{diag}(M - M_S).$$

This cannot have more fill-in than M_S, and it contains at least some information about all arcs. We will denote these as "tree/BCT+diagonal," or "T/BCT+D" for short, preconditioners. Of course, combining the two ideas gives "T/BCT+P+D" preconditioners.

Adding the diagonal can be very useful for some classes of instances, but, as reported in [10] and essentially confirmed by our experience, it is not always convenient, so the option has to be kept open. Note that this operation adds some complexity to the Cholesky factorization of the preconditioner. This is more clearly seen in the case of T+D preconditioners; while the pure tree preconditioner basically does not need any factorization (it can be factored by just permuting the rows), the T+D preconditioner does need a true—although simple—factorization phase. Analogously, the factorization of BCT+D preconditioners requires the modification described by Proposition 2.4. It may also be worth remarking that the factorization routine can be somewhat simplified if $\rho = 1$, which is significant in light of the results reported in the next section.

5. A computational comparison of preconditioners. In this section, we present the results of a large-scale computational test aimed at assessing the effectiveness of our new family of preconditioners.

For our tests, we selected three well-known random generators of MCF problems: `goto` (GridOnTOrus), `gridgen`, and `netgen`.[1] For each generator, we generated a total of 12 classes of instances, with $n = 2^k$ for $k = 8, 10, 12, 14$, and 16 and up to three different densities. In particular, for $k = 8$ we generated instances with density 8, 16, and 32, for $k = 10$ we generated instances with density 8, 32, and 64, for $k = 12$ we generated instances with density 8, 64, and 256, for $k = 14$ we generated instances with density 8 and 64, and for $k = 16$ we generated only instances with density 8. In the following, we will use the form `genX.Y` to refer to the class of instances generated by the generator `gen` (`goto`, `grid`, or `net`), with $k =$ `X` and density equal to `Y`. In each class, five different instances were generated by simply changing the seed of the pseudorandom number generator.

For all the above instances, we ran an implementation of a primal-dual IP method, using a standard tree preconditioner, in order to collect the data for reproducing the matrices M at the IP iterations. Then, the different preconditioners were tested on these matrices, and an estimate of the total time that would be spent by an IP method if using each preconditioner is computed. This way, we ensure that for every preconditioner we solve exactly the same sequence of linear systems; since the systems are

[1] Source code for these generators can be downloaded, e.g., at http://www.di.unipi.it/di/groups/optimize/Data/MMCF.html; parameters for reproducing the instances are available upon request from the authors.

only approximately solved, the sequence of systems solved by different preconditioners within an IP approach would in general be different, so that directly comparing the total time spent in the solution of the linear systems during an IP method using each given preconditioner would have been unfair. The impact of the choice of the preconditioner on the overall optimization process will be analyzed in depth in a forthcoming paper, also taking into account many important details such as the different choices of IP algorithm (primal, dual, primal-dual) with several variants each (affine, barrier, predictor-corrector, etc.), and the required precision in the solution of the systems.

We remark that each one of the preconditioner procedures has been carefully implemented. In particular, during the factorization phase we have exploited as much as possible the structure of the preconditioner in order to speed up operations. For instance, T/T+P preconditioners have a Cholesky factor with entries in $\{1, -1, 0\}$ and where the entries in matrix D depend only on the predecessor arcs of each node; these matrices need not be directly constructed as such, so that both the factorization phase and the solution of the linear systems at each PCG iteration are faster in this case. Analogously, for BCT/BCT+P preconditioners the entries in the Cholesky factor depend only on the brothers and on the predecessor, but they do not depend on the sons in the first level of the BCT, which leads to some simplification in the factorization routine. Since the efficiency of these procedures is crucial, all efforts have been made to obtain the best possible implementation for all the tested preconditioners.

We also remark that we have used an adaptive stopping rule for the PCG: the algorithm for solving (1.2) is stopped when a vector Δy is found such that

$$|d_i - M_i \Delta y| \leq \varepsilon_1 \max(|b_i - E_i \bar{x}|, \varepsilon_2 \max(|b_i|, 1))$$

for all components i, where $\bar{x}$ is the current primal solution of the IP algorithm. It is easy to check that this stopping rule allows early termination in the initial IP iterations, thereby improving the overall efficiency of the IP approach, by ensuring that the PCG is stopped as soon as the system is solved with enough precision to decrease the infeasibility of the primal solution, if it is not feasible yet, or that the violation of the primal constraints is not worsened too much, if the primal solution is already feasible (usually, because of cancellation of errors this is enough to keep the primal solution feasible until termination). The tolerance ε_1 is set to 0.1, while ε_2 is the relative feasibility precision required to the constraints satisfaction, typically set to `1e-6`. We have also tested the alternative "cosine" stopping rule proposed in the literature [18], but we have found it to be less reliable from the IP viewpoint; this is probably due to the fact that we have used a standard primal-dual IP algorithm rather than a primal-infeasible/dual-feasible one [15].

The computational experiments were performed in three phases. In the preliminary phase, a significant subset of the instances were tested with *all* the over 200 possible variants of preconditioners obtained by implementing the ideas presented in sections 3 and 4 and in [8]. This allowed us to discover that certain choices were consistently outperformed, thus reducing the set of promising preconditioners to only eight. In the second phase these preconditioners were tested on the full set of instances, in order to develop automatic rules for choosing the right preconditioner for each instance. Finally, in the third phase we compared the performances of the code having the automatic preconditioner selection rule with that of the corresponding T/T+D (whichever of the two was better) preconditioner. We will report the results of the three phases separately.

5.1. Preliminary experiments. In the preliminary phase, we were able to establish with a high degree of confidence the following facts:

- As already mentioned, using approximated maximum-weight spanning trees as the basis for the heuristics is consistently the best choice.
- The T+D and BCT+D preconditioners were found to be preferable to their "pure" counterparts for the `grid` and `net` classes, while the converse happens for the `goto` problems (except in the very first iteration when $\Theta = I_m$); this basically confirms the results reported in [10].
- When a "+D" preconditioner is used, $\rho = 1$ seems to be the best option in general, at least for the classes of instances at hand.
- Working with the full rank-deficient system M is consistently better than eliminating one row when a "+D" preconditioner is used (this is reasonable, since then the preconditioner is nonsingular even if the whole system is not), while eliminating the row and working with a nonsingular system is preferable if the diagonal is not added.
- When one row (node) has to be removed from the system, the best choice appears to be the one with the largest total weight of the set of incident arcs.
- When working with the rank-deficient system, the choice of the root node—which impacts the heuristics for the maximum-weight BCT computation—has little effect.

We are not reporting the tables relative to the experiments in the preliminary phase in order to save space.

At the end of the preliminary phase, we were therefore able to decide that all preconditioners should find the initial tree with an approximated maximum-weight spanning tree computation. Furthermore, for `goto` problems we did not use the "+D" preconditioners, and therefore we eliminated one row and worked with the full-rank subsystem, while for `grid` and `net` problems we did use the "+D" preconditioners, therefore working with the rank-deficient system M. The remaining choices were about which heuristic was used for finding the BCT ((ii.a), (ii.b), (ii.c), or none, i.e., the tree preconditioner) and whether or not "+P" preconditioners are used, for a grand total of eight different variants. For those we ran the code on all the instances, obtaining the results reported in the next section.

5.2. The second phase. The complete results of the second phase are shown in Table 5.1. There are seven groups of two columns. The first three, labeled *B-a*, *B-b*, and *B-c*, report the results relative to BCT preconditioners where the BCT is found with heuristic (ii.a), (ii.b), and (ii.c), respectively. The fourth group, labeled *TP*, reports the results relative to the T+P preconditioner. Finally, the last three groups, labeled *BP-a*, *BP-b*, and *BP-c*, report the results relative to BCT+P preconditioners. For `grid` and `net` problems only, these preconditioners have to be intended as "+D" also. All the results in the tables are normalized with respect to those obtained by the tree preconditioner (without "+P", and with or without "+D" according to the problem class); that is, the numbers in the columns *Iter* and *Time* are, respectively,

$$Iter = \frac{\text{number of iterations of the corresponding preconditioner}}{\text{number of iterations of the tree preconditioner}}$$

and

$$Time = \frac{\text{running time of the corresponding preconditioner}}{\text{running time of the tree preconditioner}}$$

(averaged among the five instances of each class). This makes it easier to spot where

TABLE 5.1
Comparison of the most promising preconditioners.

	$B-a$		$B-b$		$B-c$		TP		$BP-a$		$BP-b$		$BP-c$	
goto	Iter	Time	Iter	Time	Iter	Time	Iter	Time	Iter	Time	Iter	Time	Iter	Time
8.8	0.97	*	0.87	*	0.87	*	0.90	*	0.88	*	0.80	*	0.79	*
8.16	0.97	*	0.84	*	0.84	*	0.82	*	0.78	*	0.69	*	0.68	*
8.32	0.96	*	0.82	*	0.82	*	0.78	*	0.75	*	0.64	*	0.63	*
10.8	0.99	*	0.86	*	0.85	*	0.77	*	0.76	*	0.70	*	0.70	*
10.32	0.97	0.98	0.81	0.85	0.80	0.86	0.77	0.78	0.74	0.77	0.62	0.68	0.62	0.67
10.64	0.98	1.00	0.86	0.90	0.84	0.88	0.74	0.75	0.69	0.71	0.62	0.66	0.62	0.65
12.8	0.99	1.01	0.86	0.93	0.86	0.92	0.84	0.84	0.82	0.86	0.79	0.85	0.78	0.84
12.64	0.98	0.99	0.84	0.88	0.84	0.87	0.73	0.67	0.71	0.67	0.64	0.62	0.64	0.61
12.256	0.97	0.97	0.80	0.84	0.79	0.83	0.72	0.71	0.68	0.71	0.50	0.56	0.50	0.56
14.8	0.99	1.00	0.76	0.83	0.76	0.83	0.33	0.37	0.33	0.37	0.30	0.39	0.30	0.38
14.64	0.98	1.00	0.78	0.83	0.78	0.83	0.63	0.64	0.62	0.65	0.53	0.62	0.53	0.59
16.8	1.01	1.01	0.72	0.74	0.71	0.74	0.22	0.24	0.22	0.24	0.19	0.22	0.19	0.22
grid	Iter	Time	Iter	Time	Iter	Time	Iter	Time	Iter	Time	Iter	Time	Iter	Time
8.8	1.00	*	0.99	*	0.99	*	0.87	*	0.87	*	0.87	*	0.87	*
8.16	0.99	*	0.99	*	0.98	*	0.97	*	0.97	*	0.95	*	0.95	*
8.32	0.99	*	1.00	*	1.00	*	0.97	*	0.96	*	0.97	*	0.97	*
10.8	1.00	*	1.00	*	1.00	*	0.82	*	0.82	*	0.82	*	0.82	*
10.32	0.99	1.07	0.98	1.12	0.98	1.11	0.94	0.96	0.94	1.02	0.94	1.12	0.94	1.13
10.64	1.00	1.12	0.98	1.28	0.98	1.31	0.99	0.99	0.98	1.09	0.98	1.24	1.00	1.29
12.8	1.00	1.05	1.00	1.08	1.00	1.09	0.63	0.70	0.63	0.73	0.63	0.77	0.63	0.76
12.64	1.00	1.12	1.00	1.29	1.00	1.34	1.00	0.94	1.00	1.05	0.99	1.20	0.99	1.29
12.256	1.00	1.08	0.99	1.26	0.99	1.39	0.97	0.94	0.97	1.06	0.76	1.07	0.96	1.35
14.8	1.00	0.94	1.00	0.98	1.00	1.00	0.38	0.40	0.38	0.41	0.38	0.44	0.38	0.44
14.64	1.00	1.08	1.00	1.20	1.00	1.25	0.91	0.97	0.91	1.03	0.91	1.14	0.92	1.20
16.8	1.00	1.00	1.00	1.03	1.00	1.02	0.31	0.33	0.31	0.34	0.31	0.35	0.31	0.36
net	Iter	Time	Iter	Time	Iter	Time	Iter	Time	Iter	Time	Iter	Time	Iter	Time
8.8	0.99	*	1.00	*	1.00	*	1.00	*	1.00	*	1.00	*	1.00	*
8.16	1.00	*	0.99	*	0.98	*	1.00	*	1.00	*	0.99	*	0.99	*
8.32	1.00	*	1.00	*	1.01	*	1.00	*	1.00	*	1.00	*	1.00	*
10.8	0.99	*	0.99	*	0.99	*	1.00	*	0.99	*	0.99	*	0.99	*
10.32	1.00	1.08	1.00	1.15	1.00	1.17	1.00	1.09	1.00	1.14	1.00	1.19	0.99	1.19
10.64	0.99	1.04	1.00	1.15	1.01	1.21	1.00	0.94	0.99	1.02	1.00	1.12	1.00	1.11
12.8	1.05	1.05	1.00	1.09	1.00	1.09	0.99	1.00	1.03	1.06	1.02	1.14	1.02	1.13
12.64	1.00	1.11	1.00	1.19	1.00	1.25	1.00	1.02	0.99	1.16	0.99	1.21	0.99	1.28
12.256	0.99	1.13	0.99	1.27	0.99	1.35	1.00	1.01	0.99	1.18	0.99	1.33	0.99	1.36
14.8	1.00	0.94	1.00	1.02	1.00	1.11	1.00	1.11	1.00	1.06	1.00	1.13	1.00	1.17
14.64	1.00	1.05	1.00	1.15	1.00	1.19	1.00	0.97	1.00	1.06	1.00	1.16	1.00	1.22
16.8	1.00	1.06	1.00	1.13	1.00	1.15	1.00	1.01	1.00	1.06	1.00	1.13	1.00	1.16

the new preconditioners improve upon the known ones (entries < 1), and it highlights some interesting trends, as we will see later on. However, for the smaller instances we elected not to report running times, as each system was timed separately, and the time required to solve one system was too near to the precision of the timing routines, and therefore too affected by errors, to be significant.

We will now comment on the results for the three classes of problems separately.

`goto` *instances.* For these instances, the new preconditioners are quite competitive with the tree one, obtaining, when parallel arcs are added, improvements of up

TABLE 5.2
More detailed results for 10.32 *instances.*

	goto 10.32				grid 10.32				net 10.32			
	T	B-b	TP	BP-b	T	B-b	TP	BP-b	T	B-b	TP	BP-b
0	729	0.78	0.37	0.26	12	0.95	0.95	0.95	10	1.00	0.98	1.00
1	95	0.78	0.85	0.69	10	0.98	0.96	0.96	11	0.98	1.00	0.98
$k/4$	77	0.81	0.77	0.62	11	1.00	0.98	0.98	11	1.00	0.98	0.98
$k/2$	47	0.82	0.86	0.68	16	0.98	0.94	0.93	15	1.00	1.00	1.00
$3k/4$	27	0.87	0.92	0.80	14	0.99	0.94	0.93	15	1.00	1.00	1.00
k-1	16	0.95	1.00	0.92	7	1.00	0.85	0.85	7	1.00	1.00	1.00
k	16	0.95	1.00	0.92	3	1.00	0.87	0.87	3	1.00	1.00	1.00

to a factor of five in iterations count, and only slightly less so in time. Among BCT preconditioners, the more complex heuristics (ii.b) and (ii.c) clearly outperform the simpler (ii.a), with the most complex one, (ii.c), oftentimes slightly outperforming (ii.b). There does not seem to be a clear trend regarding graph density, with denser graphs sometimes benefiting more and other times benefiting less from BCT preconditioners than sparse ones; however, there is a clear positive trend with graph size, in that larger problems benefit most from BCT preconditioners.

`grid` *instances.* Even for these instances, enriching the support graph by adding more arcs turns out to be in general a good strategy; this time, however, it is the addition of parallel arcs that makes up the largest part of the improvement. In fact, although improvements of up to a factor of three are still obtained, the T+P preconditioner is the most competitive one. BCT preconditioners often obtain smaller iteration counts than the corresponding tree one, but only slightly so, and this does not pay for the extra cost of finding the preconditioner. Among BCT preconditioners, the more complex heuristics (ii.b) and (ii.c) fail, on this class of instances, to obtain more than minor improvements with respect to the simpler (ii.a), so that the most complex one, (ii.c), is usually the slowest one. The same positive trend with graph size as in the `goto` case shows up; this time, however, there appears to be something of a more defined trend with density, too, as improvements tend to be more consistent for problems on sparser graphs.

`net` *instances.* For this class of instances, the new preconditioners are not competitive with the tree one. Although enriching the support graph fairly often decreases the iterations count, the decrease is always minimal, and adding parallel arcs does not help; for these instances, all the mechanisms for enriching the support graph actually increase the total running time required for solving the systems.

In order to better understand the behavior of the preconditioners, it is worthwhile to examine some of the results in greater detail. In Table 5.2 we report some data about the number of iterations required to solve problems of the same size (the class `10.32`) generated by the three different generators. For each generator, we report seven rows corresponding to the systems solved at IP iterations 0, 1, $k/4$, $k/2$, $3k/4$, $k-1$, and k, where k is the index of the last iteration; this is a significant sample of the matrices generated during the IP algorithm. In particular, the systems of iteration 0 are those solved to find an initial interior solution, for which $\Theta = I_m$ (i.e., $M = EE^T$ [7]). For each generator, the column T reports the number of PCG iterations required for solving the system using the tree preconditioner, while the columns *TP*, *B-b*, and *BP-b* have the same meaning as the columns *Iter* in the corresponding sections of Table 5.1.

TABLE 5.3
Number of arcs added to the support graph.

	T	$B-a$		$B-b$		$B-c$	
`goto`	#TP	#B	#BP	#B	#BP	#B	#BP
0	927	2	0	410	8	410	8
1	927	2	0	384	8	384	8
$k/4$	843	54	14	385	33	385	33
$k/2$	724	86	27	370	72	371	73
$3k/4$	722	88	26	368	72	369	73
$k-1$	721	87	26	367	70	367	70
k	721	84	26	354	69	354	69
`grid`	#TP	#B	#BP	#B	#BP	#B	#BP
0	15	5	2	13	6	13	6
1	401	96	35	145	53	188	67
$k/4$	569	5	1	14	3	14	3
$k/2$	546	5	1	12	3	12	3
$3k/4$	544	5	2	13	3	13	3
$k-1$	528	2	1	6	2	6	2
k	498	0	0	1	0	1	0
`net`	#TP	#B	#BP	#B	#BP	#B	#BP
0	5	29	0	41	0	41	0
1	4	8	0	24	0	24	0
$k/4$	4	9	0	23	0	23	0
$k/2$	6	9	0	20	0	20	0
$3k/4$	6	11	0	20	0	20	0
$k-1$	6	7	0	13	0	13	0
k	5	4	0	8	0	8	0

The results show that the systems corresponding to `goto` instances are considerably more difficult to solve than those corresponding to either `grid` or `net` instances. The effect of the BCT preconditioner on `goto` instances is larger in the first iterations, where the tree preconditioner is less effective, and diminishes as the IP algorithm proceeds; for `grid` and `net` instances the effect is very limited across the board, and no clear trend emerges. The effect of the "+P" variant is less easy to characterize, with a decreasing trend showing up for `goto` instances and no clear trend emerging for `grid` instances. It is, however, interesting to note that, for the `goto` instances, in the very final iterations of the IP algorithm the "+P" variant alone does not seem to produce any improvement to the tree preconditioner, while it is capable of helping out, albeit slightly, the BCT one.

The above results can be better understood by looking at Table 5.3, where the number of arcs added to the spanning tree in the different variants of preconditioners is reported. In the table, the three groups of two columns labeled $B-a$, $B-b$, and $B-c$ correspond to the heuristics (ii.a), (ii.b), and (ii.c), respectively, for the maximum-weight BCT computation. In each group, the column #B reports the number of arcs in the second level of the BCT found by the heuristic, and the column #BP reports the number of arcs "parallel" to those in the second level of the BCT. Finally, the column #TP reports the number of arcs "parallel" to those of the original spanning tree. The table shows the (averaged) results for the `10.8` instances for the three different generators; these results can be considered typical. For each generator, we report seven rows corresponding to the systems at the same seven "snapshots" of the optimization process as in Table 5.2.

These results show that the effectiveness of the new preconditioners—at least, relative to that of the tree one—is directly related to the number of arcs that are added to the support graph. In particular, for `goto` instances the heuristic (ii.a) adds considerably fewer arcs than (ii.b) or (ii.c), and in fact it is less effective; furthermore, a large number of "parallel" arcs are added to the support graph, and in fact the corresponding preconditioners improve upon those where this is not done. For `grid` instances, the BCT heuristics are not capable of adding many arcs to the support graph (except in the second iteration), while a large number of "parallel" arcs are added; indeed, adding parallel arcs is what makes the difference for these instances. Finally, for `net` instances very few arcs are added to the support graph by both methods, and this directly translates into the inferior performances of the new preconditioners.

These results lead us to the following conclusions:

- Of all heuristics for finding the BCT, (ii.b) is the one that obtains the best performances, being far more efficient than (ii.a) in adding arcs to the support graph and only slightly less so than (ii.c), but is, however, much more costly; this confirms that balancing the effort for finding/factoring the preconditioner with the improvement in the convergence rate of the PCG is crucial.
- Enriching the support graph turns out to be a good strategy for those problems that are not easily solved by the tree preconditioner, whereas it is less useful for systems that are already very efficiently solved by the tree preconditioner.
- The relative efficiency of the new subgraph-based preconditioners with respect to the tree one is well predicted by the number of arcs added to the spanning tree; this has been confirmed by the analysis of data for all the instances, which we do not report here to save space.

5.3. Final results. Given the results of the previous section, we have tested the effect of an automatic rule for choosing the preconditioner. Sticking to heuristic (ii.b) for finding the BCT, we initially start by using both BCT and "+P" preconditioners. The number of arcs added to the support graph S by both operations are counted; if this number is larger than a fixed threshold, the preconditioner actually includes those arcs; otherwise the operation is disabled in that and all the following IP iterations. This choice is motivated by the fact that the tree preconditioner becomes more and more efficient as the IP algorithm proceeds; hence if adding arcs to the support graph is not likely to help at a given iteration, it is somewhat unlikely that is it going to help later. Permanently disabling the rule is simple and has the advantage of avoiding the cost for finding a BCT and/or parallel arcs that are not going to be used (the cost for factoring M_S is not paid anyway because the decision is taken before the factorization).

The analysis of the obtained results has shown that reasonable thresholds are 45% for the BCT and 10% for "+P"; that is, using the BCT is disabled if it does not add at least as many as $0.45(n-1)$ arcs, and using parallel arcs is disabled if it does not add at least as many as $0.10(n-1)$ arcs. These thresholds appear to work well for all three classes of instances.

The results of using these rules are shown in Table 5.4; as for the previous tables, the results are relative to those obtained by the tree preconditioner ("+ D" or not, according to the class of instances).

The table shows that the rules are, at least in these instances, capable of choosing the right preconditioner at the right time. Most often the chosen preconditioner is always the same for all the IP iterations, but in some cases a switch happens during

TABLE 5.4
Results with the automatic selection rule.

	goto		grid		net	
	Iter	Time	Iter	Time	Iter	Time
8.8	0.80	*	0.87	*	1.00	*
8.16	0.69	*	0.97	*	1.00	*
8.32	0.64	*	0.97	*	1.00	*
10.8	0.70	0.84	0.82	0.90	1.00	1.00
10.32	0.62	0.68	0.94	0.96	1.00	1.00
10.64	0.62	0.66	0.99	0.99	1.00	0.99
12.8	0.79	0.85	0.63	0.70	1.00	1.00
12.64	0.64	0.62	1.00	0.97	1.00	1.00
12.256	0.50	0.56	0.97	0.94	1.00	1.00
14.8	0.30	0.39	0.38	0.40	1.00	1.00
14.64	0.53	0.59	0.91	0.97	1.00	1.00
16.8	0.19	0.22	0.31	0.33	1.00	1.00

the optimization process which may modify the running time w.r.t. the case where the same preconditioner is used throughout the IP algorithm, either decreasing it (as for `goto` 14.64 and `net` 10.64) or increasing it (as for `grid` 12.64) but always by a relatively small amount. More sophisticated selection rules may further improve the results, but the obtained ones already show that BCT preconditioners, if carefully implemented and paired with appropriate automatic selection rules, can effectively complement tree preconditioners as a solution tool for the linear systems arising in IP methods for MCF problems.

6. Conclusion and directions for future work. We have proposed a new family of subgraph-based preconditioners for the solution of the KKT systems arising in the solution of MCF problems through IP methods. For some families of instances, these preconditioners improve on those known in the literature both in iterations count and total time. Also, the new family of preconditioners offers some flexibility in the way to select the subgraph, thereby allowing us to tune the trade-off between the cost of computing and using the preconditioner and the corresponding reduction in the number of PCG iterations. Therefore, we believe that our new preconditioners can be a valuable tool for constructing efficient IP algorithms for MCF problems. Furthermore, they may find broader application for the solution of linear systems with *M-matrices* [3].

Further work along this line of research will involve perfecting our implementation of an IP method for MCF problems and testing it against efficient MCF codes from the literature; the results will be presented in a forthcoming paper, where all the issues relative to the effectiveness of the different variants of preconditioners for different IP algorithms will be discussed. Also, other fast heuristics for the maximum-weight BCT problem will be tested, trying to find an optimal compromise between the quality of the BCT found and the extra cost involved in finding it; that is a critical parameter for the overall efficiency of the approach. Finally, theoretical investigations on the class of BCT graphs may pay off in terms of better heuristics, characterization of some classes of graphs where "large" BCTs can be easily found, and a better understanding of the complexity class of the maximum-weight BCT computation.

REFERENCES

[1] R. Ahuja, T. Magnanti, and J. Orlin, *Network Flows: Theory, Algorithms and Applications*, Prentice Hall, Englewood Cliffs, NJ, 1993.

[2] W. Anderson and T. Morley, *Eigenvalues of the Laplacian of a graph*, Linear and Multilinear Algebra, 18 (1985), pp. 141–145.

[3] M. Bern, J. Gilbert, B. Hendrickson, N. Nguyen, and S. Toledo, *Support-graph preconditioners*, SIAM J. Matrix Anal. Appl., submitted.

[4] J. Castro, *A specialized interior-point algorithm for multicommodity network flows*, SIAM J. Optim., 10 (2000), pp. 852–877.

[5] D. Chen and S. Toledo, *Vaidya's preconditioners: Implementation and experimental study*, Electron. Trans. Numer. Anal., 16 (2003), pp. 30–49.

[6] D. Cvetkovic, M. Doob, and H. Sachs, *Spectra of Graphs*, Academic Press, New York, 1979.

[7] A. Frangioni and S. S. Capizzano, *Spectral analysis of (sequences of) graph matrices*, SIAM J. Matrix Anal. Appl., 23 (2001), pp. 339–348.

[8] A. Frangioni and C. Gentile, *Interior Point Methods for Network Problems*, Tech. report 539, IASI-CNR, Rome, 2000.

[9] K. Gremban, G. Miller, and M. Zagha, *Performance evaluation of a parallel preconditioner*, in Proceedings of the 9th IEEE International Parallel Processing Symposium, Santa Barbara, CA, 1995, pp. 65–69.

[10] J. Júdice, J. Patricio, L. Portugal, M. Resende, and G. Veiga, *A study of preconditioners for network interior point methods*, Comput. Optim. Appl., 24 (2003), pp. 5–35.

[11] D. Karger and N. Srebro, *Learning Markov networks: Maximum bounded tree-width graphs*, in Proceedings of the Twelfth Annual ACM-SIAM Symposium on Discrete Algorithms, ACM Press, New York, SIAM, Philadelphia, 2001, pp. 392–401.

[12] N. K. Karmarkar and K. G. Ramakrishnan, *private comunication*, 1988.

[13] S. Mehrotra and J. Wang, *Conjugate gradient based implementation of interior point methods for network flow problems*, in Linear and Nonlinear Conjugate Gradient-Related Methods, L. Adams and J. Nazareth, eds., SIAM, Philadelphia, 1996, pp. 124–142.

[14] L. Portugal, F. Bastos, J. Júdice, J. Paixao, and T. Terlaky, *An investigation of interior-point algorithms for the linear transportation problem*, SIAM J. Sci. Comput., 17 (1996), pp. 1202–1223.

[15] L. Portugal, M. Resende, G. Veiga, and J. Júdice, *A truncated primal-infeasible dual-feasible network interior point method*, Networks, 35 (2000), pp. 91–108.

[16] M. Resende and P. Pardalos, *Interior point algorithms for network flow problems*, in Advances in Linear and Integer Programming, J. Beasley, ed., Oxford University Press, Oxford, UK, 1996, pp. 147–187

[17] M. Resende and G. Veiga, *An implementation of the dual affine scaling algorithm for minumum-cost flow on bipartite uncapacited networks*, SIAM J. Optim., 3 (1993), pp. 516–537.

[18] M. Resende and G. Veiga, *An efficient implementation of a network interior point method*, in Network Flows and Matching: First DIMACS Implementation Challenge, DIMACS Ser. Discrete Math. Theoret. Comput. Sci. 12, D. Johnson and C. McGeoch, eds., AMS, Providence, RI, 1993, pp. 299–348.

[19] T. Roos, T. Terlaky, and J.-P. Vial, *Theory and Algorithms for Linear Optimization: An Interior Point Approach*, John Wiley, Chichester, UK, 1997.

[20] D. Rose, *Triangulated graphs and the elimination process*, J. Math. Anal. Appl., 32 (1970), pp. 597–609.

[21] R. Tarjan and M. Yannakakis, *Simple linear-time algorithms to test chordality of graphs, test acyclicity of hypergraphs, and selectively reduce acyclic hypergraphs*, SIAM J. Comput., 13 (1984), pp. 566–579.

[22] P. Vaidya, *Solving Linear Equations with Symmetric Diagonally Dominant Matrices by Constructing Good Preconditioners*, Tech. report, Department of Computer Science, University of Illinois at Urbana-Champaign, Urbana, IL, 1990.

[23] C. Wallacher and U. Zimmermann, *A combinatorial interior point method for network flow problems*, Math. Programming, 56 (1992), pp. 321–335.

[24] S. Wright, *Primal-Dual Interior-Point Methods*, SIAM, Philadelphia, 1997.

[25] M. Yannakakis, *Computing the minimum fill-in is NP-complete*, SIAM J. Alg. Disc. Meth., 2 (1981), pp. 77–79.

SIAM J. OPTIM.
Vol. 14, No. 3, pp. 914–935

RIGOROUS LOWER AND UPPER BOUNDS IN LINEAR PROGRAMMING*

CHRISTIAN JANSSON†

Abstract. We consider the computation of rigorous lower and upper error bounds for the optimal value of linear programming problems. The input data of the lp-problem may be exactly given or may vary between given lower and upper bounds. The results are then verified for the family of lp-problems with input data inside these bounds. In many cases only a small computational effort is required. For problems with finite simple bounds, the rigorous lower bound of the optimal value can be computed with $O(n^2)$ operations. The error bounds can be used as well to perform a sensitivity analysis, provided the width of the uncertainties is not too large. Some numerical examples are presented.

Key words. linear programming, interval arithmetic, rigorous error bounds, sensitivity analysis

AMS subject classifications. 90C05, 65G30, 65N15

DOI. 10.1137/S1052623402416839

1. Introduction.

Linear programming problems have been solved very successfully during the last decades, and many efficient algorithms have been developed. However, due to rounding errors and/or uncertainties in the input data, the computed approximations sometimes may be unsatisfactory. Especially for ill-conditioned problems, the results may be quite inaccurate.

Ill-conditioning is not a rare phenomenon. In the recent paper of Ordóñez and Freund [22] it is stated that 72% of the lp-instances in the NETLIB linear programming library [18] are ill-conditioned. They used a condition number which is a scale-invariant reciprocal of the smallest data perturbation that renders the perturbed data instance either primal or dual infeasible. After applying CPLEX 7.1 presolve (a preprocessing heuristic for linear programming), still 19% maintain the property of being ill-conditioned. The NETLIB library contains many industrial problems so that the computation of rigorous error bounds can be valuable in practice.

Beeck [2], Krawczyk [13], and Rump [26] have developed methods for computing rigorous error bounds for lp-problems in which the optimal solution is unique and not degenerate. The common basic idea is to use the simplex method for the computation of an optimal basic index set. Then the optimality of this index set is verified a posteriori with interval methods, and rigorous error bounds for the optimal vertex and the optimal value are calculated. Moreover, these algorithms also can be applied to the calculation of bounds for the family of lp-problems in which the input data vary between given lower and upper bounds. These error bounds require $O(n^3)$ operations, where n denotes the number of variables of the linear programming problem. In Jansson [8] the degenerate case is also considered by applying the well-known graph search method to an appropriately defined graph of degenerate basic index sets. For each basic index set, a linear interval system must be solved, and for all optimal vertices, rigorous error bounds are computed. The computational work is $k \cdot O(n^3)$, where k is greater than or equal to the number of (degenerate) optimal basic index sets.

*Received by the editors October 24, 2002; accepted for publication (in revised form) November 13, 2003; published electronically May 17, 2004.
http://www.siam.org/journals/siopt/14–3/41683.html
†Institut für Informatik III, Technische Universität Hamburg-Harburg, Schwarzenbergstraße 95, Hamburg 21071, Germany (jansson@tu-harburg.de).

The approaches above have in common that for each lp-problem of the family, the existence of optimal solutions is proved. At first, error bounds for the optimal vertices are computed by solving linear interval systems, and then these bounds are used for the computation of error bounds for the optimal value.

The major emphasis in this paper is on the fast computation of rigorous error bounds for the optimal value without using or computing bounds for the optimal vertices. The algorithm presented here avoids the solution of linear interval systems and needs $O(n^2)$ operations for the lower bound, provided that all variables are bounded and an *approximate* optimal solution is known. Nothing is assumed about the quality of this approximate solution. The computational work for the upper bound requires $O(mn + p^3)$ operations, where m is the number of inequalities and p is the number of equations of the lp-problem, and, beyond that, requires the computational costs for applying the lp-solver to a slightly perturbed problem using the known approximate solution as a starting point.

In the following, we do not consider the standard linear programming problem. Our reasons are that transformations to the standard problem may lead to more computational work, and they cause data dependencies in the input data, which may lead to worse error bounds. Therefore, the inequalities are not transformed into equations so that in many cases the number p of equations is small compared to mn, implying only $O(mn)$ operations to compute the upper bound.

It turns out that the lower bound and the upper bound for the optimal value can be computed for both degenerate and large (sparse) problems, provided the lp-routine in use is suited for these problems. The lower bound is based on some duality arguments and is computable even for ill-posed problems. The upper bound uses a technique for underdetermined nonlinear systems of equations which is proposed in Hansen [6] and further investigated by Kearfott [11].

For the large radii of the interval input data and the large absolute value of the simple bounds, the methods presented in [2], [13], and [26] may compute better rigorous bounds, provided that (i) the lp-problems within the interval input data are well-posed and have unique nondegenerate optimal solutions, and (ii) the large $n \times n$ interval linear system corresponding to the optimal basic index set can be solved with narrow bounds.

Rigorous error bounds are not only useful for problems which can be modeled as lp-problems, but also can be used, for example, in global optimization and mixed integer programming whenever linear relaxations must be solved. Independently and at the same time, Neumaier and Shcherbina [21] have developed results which overlap in part with results presented here. However, there are several differences. Their emphasis is more on branch and cut methods for linear mixed integer problems, whereas we discuss problems with uncertain input data, simple bounds which may be infinite, and a different upper bound.

Our paper is organized as follows. In section 2 examples are presented in which a commercial lp-solver computes incorrect approximations, although the problems look simple. Section 3 contains notation, including intervals and their operations. We stress that the main parts of this paper can be understood without knowledge of interval arithmetic (see the hint at the end of section 3). In the subsequent section some comments about solving linear interval systems are given. In section 5 an algorithm for computing a rigorous upper bound of the global minimum value is considered. Then, in section 6 a rigorous lower bound is presented. After some applications in section 7, we present numerical results in section 8. Finally, in section 9 some conclusions are given.

2. Examples. For many practical lp-problems the input data are integer (and therefore exactly representable on a computer), the problem is well-conditioned, and together with an appropriate pivoting rule (for example, Bland's rule) the optimal solution can be computed in a finite number of steps, even in the degenerate case. In addition to implying the linearity of the problem, we may argue that these strong assumptions should imply that the computed approximation is satisfactory. But the following example shows that the commercial lp-solver *linprog* in MATLAB [14] fails, even though all the above assumptions are satisfied.[1] This solver is a variant of the well-known Mehrotra's predictor-corrector algorithm [15].

The problem

$$\begin{array}{rcrcrl} \min -5x_1 & - & 4x_2 & - & 6x_3 & \text{subject to} \\ x_1 & - & 6x_2 & + & x_3 & \leq 20, \\ 3x_1 & + & 2x_2 & + & 4x_3 & \leq 42, \\ 3x_1 & + & 2x_2 & & & \leq 30, \end{array} \tag{1}$$

with $0 \leq x_i \leq \overline{x}_i$ for $i = 1, 2, 3$, is bounded from above by the value 21 for every variable; this can be seen by a short inspection of the second inequality. The optimal solution is $x^* = (0, 15, 3)^T$ with optimal value $f^* = -78$. Hence, the three upper bounds $\overline{x} = (30, 30, 30)^T$, $\overline{x} = (10^{10}, 10^{10}, 10^{10})^T$, and $\overline{x} = (+\infty, +\infty, +\infty)^T$ change neither the set of feasible solutions nor the optimal solution. The MATLAB routine *linprog* computes in all three cases the optimal solution and the optimal value with accuracy of about 11 decimal digits. Now, to the above constraints, we add the linear equation

$$x_1 + x_2 + x_3 = 10 \tag{2}$$

and apply *linprog* for this problem with the previous three upper bounds. Notice that the resulting set of feasible solutions has a nonempty relative interior, and for all three upper bounds the optimal solution is $x^* = (0, 0, 10)$ with optimal value $f^* = -60$.

For the upper bounds $\overline{x} = (30, 30, 30)^T$ and $\overline{x} = (+\infty, +\infty, +\infty)^T$, *linprog* calculates the optimal solution but, surprisingly, for the upper bound $\overline{x} = (10^{10}, 10^{10}, 10^{10})$, the message

both the primal and the dual appear to be infeasible

is displayed. In the case of a warning, the routine *linprog* also provides the last iteration point before the routine stops. For this example, this point is equal to $x = (0.0045, 5.6042, 10.4218)$, which does not satisfy (2).

In the second example, we tried to solve with *linprog* the problem

$$\begin{array}{rcll} \min & \sum_{i=1}^{n} & s_i & \text{subject to} \\ Ax & + & s & = e_1, \\ 0 & \leq x_i, & s_i & \leq 1000 \text{ for } i = 1, \ldots, n, \end{array} \tag{3}$$

where $x, s \in \mathbf{R}^n$, e_1 is the first unit vector, and A is the inverse of the $n \times n$ Pascal matrix. Notice that the Pascal matrix and its inverse have integer entries. A short

[1] By accident, this example was discussed during a course of lectures about optimization for students studying electrical engineering. During this course, branch and bound strategies for linear mixed integer problems and the influence of the simple bounds were investigated. We used the MATLAB routine *linprog* and, surprisingly, by increasing the upper bounds, the set of feasible solutions became empty. Unfortunately, I cannot find the reference for the data of this lp-problem. It could be that it was originally presented in a book borrowed on interlibrary loan.

calculation shows that the optimal solution is $x = A^{-1}e_1$ (which is 1 in every component of x), and $s = 0$. The basis matrix corresponding to the optimal solution is A. For $n = 9$ the 2-norm condition number of A is 2.9×10^8. Because MATLAB uses double precision, one would expect that about 8 decimal places of the optimal solution are correct. In fact, this is true if the solution of the linear system $Ax = e_1$ is computed by the linear system solver in MATLAB. However, the MATLAB routine *linprog* displays the message

the primal appears to be infeasible (and the dual unbounded).

We mention that, for our first example, the results of MATLAB Versions 5.3 and 6.0 are identical. For the second example they differ slightly. The message above comes from Version 5.3. Version 6.0 displays, after 86 iterations, the message

maximum number of iterations exceeded,

and the component $x(9)$ of x in the last iteration is 4.858×10^{-15}, even though the optimal component is 1. Increasing the maximum number of iterations to 1000 does not change anything. For $n = 10$ the 2-norm condition number of A is 4.2×10^9, and also in Version 6.0 the message

the primal appears to be infeasible (and the dual unbounded)

is displayed.

We have investigated the examples above by using the NAG-solver E04MBF [17], which is an implementation of the simplex-method. This solver has produced good approximations in these cases.

A further example can be found in Neumaier and Shcherbina [21]. They give an innocent-looking mixed linear integer problem, where the commercial, high quality solver CPLEX [7] and several others failed. The reason is that the linear programming relaxations are not solved correctly by the lp-algorithm.

In backward error analysis it is described that many algorithms are backward stable; that is, the computed solution is the exact result of a problem with slightly perturbed input data. A main assumption in most proofs for backward stability of algorithms is that the perturbations of the input data vary componentwise independently. But this assumption is not fulfilled for many real-life problems.

A short inspection of our examples shows that *linprog* is not backward stable. But even backward stable results may be unsatisfactory in applications. This is pointed out by Neumaier and Shcherbina in [21] as follows:

> *However, backward error analysis has no relevance for integer linear programs with integer coefficients, since slightly perturbed coefficients no longer produce problems of the same class.*

Algorithms for solving optimization problems search for an optimal solution. Sometimes the optimum cannot be found, although a convergence theory is well-elaborated and a well-known termination criteria is implemented. However, a linkage of search algorithms with rigorous error bounds makes it possible to obtain mathematically correct results for the optimal value of linear programming problems with limited computational work. It turns out that the computation of error bounds can be added a posteriori in the form of a subroutine at the end of each linear programming code. Even in some cases (see the following examples) where unsatisfactory

approximations are computed, useful bounds can be obtained. Hence, non–state-of-the-art solvers can also be used in a safe manner if postprocessing of the computed approximation with rigorous bounds is done.

3. Notation. Throughout this paper we use the following notation. We denote by $\mathbf{R}$, $\mathbf{R}^n$, and $\mathbf{R}^{m\times n}$ the sets of real numbers, real vectors, and real $m \times n$ matrices, respectively. Comparisons $\leq$ and absolute value $|.|$ are used entrywise. The coefficients of a real $m \times n$ matrix A are denoted by A_{ij}, its columns by $A_{:j}$, its rows by $A_{i:}$, and its transpose by A^T.

We require only some basic definitions of interval arithmetic that are described in this paper. There are a number of textbooks on interval arithmetic and self-validating methods that we highly recommend to readers. These include Alefeld and Herzberger [1], Kearfott [11], Moore [16], and Neumaier [19], [20].

If $\mathbf{V}$ is one of the spaces $\mathbf{R}$, $\mathbf{R}^n$, $\mathbf{R}^{m\times n}$, and $\underline{v}, \overline{v} \in \mathbf{V}$, then the box

$$\mathbf{v} := [\underline{v}, \overline{v}] := \{v \in \mathbf{V} : \ \underline{v} \leq v \leq \overline{v}\} \tag{4}$$

is called an *interval quantity* in $\mathbf{IV}$ with *lower bound* $\underline{v}$ and *upper bound* $\overline{v}$. In particular, $\mathbf{IR}$, $\mathbf{IR}^n$, and $\mathbf{IR}^{m\times n}$ denote the set of real intervals $\mathbf{a} = [\underline{a}, \overline{a}]$, the set of real interval vectors $\mathbf{x} = [\underline{x}, \overline{x}]$, and the set of real interval matrices $\mathbf{A} = [\underline{A}, \overline{A}]$, respectively. The real operations $A \circ B$ with $\circ \in \{+, -, \cdot, /\}$ between real numbers, real vectors, and real matrices can be generalized to interval operations. The result $\mathbf{A} \circ \mathbf{B}$ is defined as the interval hull of all possible real results, that is,

$$\mathbf{A} \circ \mathbf{B} := \bigcap\{\mathbf{C} \in \mathbf{IV} : \ A \circ B \in \mathbf{C} \quad \text{for all} \quad A \in \mathbf{A}, B \in \mathbf{B}\}. \tag{5}$$

All interval operations can be easily executed by working appropriately with the lower and upper bounds of the interval quantities. For example, in the simple case of addition, we obtain

$$\mathbf{A} + \mathbf{B} = [\underline{A} + \underline{B}, \overline{A} + \overline{B}]. \tag{6}$$

For interval quantities $\mathbf{A}, \mathbf{B} \in \mathbf{IV}$ we define

$$\check{\mathbf{A}} := (\underline{A} + \overline{A})/2 \quad \text{as the } \textit{midpoint}, \tag{7}$$

$$\operatorname{rad}(\mathbf{A}) := (\overline{A} - \underline{A})/2 \quad \text{as the } \textit{radius}, \tag{8}$$

$$|\mathbf{A}| := \sup\{|A| : \ A \in \mathbf{A}\} \quad \text{as the } \textit{absolute value}, \tag{9}$$

$$\mathbf{A}^+ := \{A \in V : \ A \in \mathbf{A},\ A \geq 0\}, \tag{10}$$

$$\mathbf{A}^- := \{A \in V : \ A \in \mathbf{A},\ A \leq 0\}. \tag{11}$$

Moreover, the comparison in $\mathbf{IV}$ is defined by

$$\mathbf{A} \leq \mathbf{B} \quad \text{iff} \quad \overline{A} \leq \underline{B},$$

and other relations are defined analogously.

Real quantities are embedded in the interval quantities by identifying $v = [v, v]$, and sometimes they are called *point quantities* or *point intervals*. For point quantities the real matrix-vector operations, comparisons, and absolute value coincide with the interval operations, interval comparisons, and interval absolute value. Therefore, *a reader not familiar with interval arithmetic can read the main parts of this paper just by* (i) *interpreting each interval quantity (fat symbols) as a real quantity, and*

(ii) *replacing the statement that a real quantity is contained in an interval quantity with the statement that the real quantity is equal to the interval quantity.*

In general, due to data dependencies, interval arithmetical expressions may lead to overestimation. Even interval matrix-vector expressions (defined as the interval hull of all possible real results) may cause overestimation, since matrices transform interval vectors onto parallelograms. However, for the special cases of interval inner products and for the multiplication of an interval matrix with a point vector, overestimation is absent. Hence, for two vectors $\mathbf{c}, \mathbf{x}$ the equality

$$\sup(\mathbf{c}^T\mathbf{x}) = \sup\{c^T x : c \in \mathbf{c}, x \in \mathbf{x}\}$$

holds.

4. Linear interval systems. Linear systems of equations with inexact input data are treated in interval arithmetic by working with an interval matrix $\mathbf{A} \in \mathbf{IR}^{n\times n}$ and an interval right-hand side $\mathbf{b} \in \mathbf{IR}^n$. The aim of this treatment is to compute an interval vector $\mathbf{x} \in \mathbf{IR}^n$ containing the *solution set*

$$\Sigma(\mathbf{A}, \mathbf{b}) := \{x \in \mathbf{R}^n : \ Ax = b \text{ for some } A \in \mathbf{A},\, b \in \mathbf{b}\}. \tag{12}$$

If all $A \in \mathbf{A}$ are nonsingular, then the solution set is bounded and satisfies, by definition, the property

$$\text{for all } A \in \mathbf{A}, \text{ for all } b \in \mathbf{b}\ \exists\, x \in \mathbf{x} : \ Ax = b. \tag{13}$$

Computing an enclosure $\mathbf{x}$ of the solution set is an NP-hard problem, but there are several methods that compute $\mathbf{x}$ with $O(n^3)$ operations for certain types of interval matrices. A precise description of such methods, required assumptions, and approximation properties can be found, for example, in Neumaier [19]. Roughly speaking, it turns out that for interval matrices with $\|I - R\mathbf{A}\| < 1$ (R is an approximate inverse of the midpoint $\check{\mathbf{A}}$) there are several methods which compute an enclosure $\mathbf{x}$, and the radius $\mathrm{rad}(\mathbf{x})$ decreases linearly with decreasing radii $\mathrm{rad}(\mathbf{A})$ and $\mathrm{rad}(\mathbf{b})$. For the computation of enclosures in the case of large linear systems, the reader is referred to Rump [27].

The computation of rigorous lower and upper bounds for the optimal value requires considering a modified problem. There, a nonsquare interval matrix $\mathbf{A} \in \mathbf{IR}^{m\times n}$ with $m < n$ and a right-hand side $\mathbf{b} \in \mathbf{IR}^m$ are given, and the goal is to compute an interval vector $\mathbf{x} \in \mathbf{IR}^n$ such that property (13) is fulfilled.

Since $m < n$ (in most cases m is much smaller than n), the solution set $\Sigma(\mathbf{A}, \mathbf{b})$ is in general unbounded, whereas property (13) requires finding only an enclosure of a part of the solution set.

Obviously, there are many possibilities for computing such a part of the solution set. We need to compute such an enclosure $\mathbf{x}$ with respect to a given vector $\tilde{x}$ and an index set $I := (\beta_1, \ldots, \beta_m) \subseteq \{1, \ldots, n\}$, and proceed as follows:

1. Set $\mathbf{A}_{:I} := (\mathbf{A}_{:\beta_1}, \ldots, \mathbf{A}_{:\beta_m})$, and denote the remaining matrix $\mathbf{A}$ by $\mathbf{A}_{:N} := (\mathbf{A}_{:\gamma_1} \ldots, \mathbf{A}_{:\gamma_{n-m}})$, where $N := \{\gamma_1, \ldots, \gamma_{n-m}\} = \{1, \ldots, n\}\backslash I$.
2. Set $\mathbf{x}_{\gamma_1} := \tilde{x}_{\gamma_1}, \ldots, \mathbf{x}_{\gamma_{n-m}} := \tilde{x}_{\gamma_{n-m}}$.
3. Compute an enclosure $(\mathbf{x}_{\beta_1}, \ldots, \mathbf{x}_{\beta_m})^T \in \mathbf{IR}^m$ of the solution set with square interval matrix $\mathbf{A}_{:I}$ and right-hand side

$$\mathbf{b} - \sum_{j=1}^{n-m} \mathbf{A}_{:\gamma_j} \tilde{x}_{\gamma_j} \tag{14}$$

by using an algorithm for square linear interval systems.

This algorithm yields for every $A \in \mathbf{A}$ and $b \in \mathbf{b}$ a real vector $x \in \mathbf{x}$ with

$$x_i \in \mathbf{x}_{\beta_j} \text{ if } i = \beta_j \quad \text{and} \quad x_i = \tilde{x}_{\gamma_j} \text{ if } i = \gamma_j \tag{15}$$

such that $Ax = b$, i.e., property (13) is satisfied, and the part of the solution set matching $\tilde{x}_{\gamma_j}$ for all $\gamma_j \in N$ is filtered out. Because the components $\tilde{x}_{\gamma_j}$ are point quantities, the right-hand side of (14) can, in principle, be computed without overestimation.

Hansen [6, Chapter 12] proposed a general technique to prove existence of a feasible point for m nonlinear equations within a bounded box in $\mathbf{R}^n$, where $m \leq n$ and $m-n$ variables are held fixed as above. This technique was modified and investigated numerically by Kearfott [9], [10] and is also described in his book [11, Chapter 5]. Corresponding algorithms are implemented in his software package GlobSol. In the following, for computing a rigorous upper bound in linear programming, we adapt this technique to the linear case but with particular attention to uncertain input data. In particular, this causes a special choice of deflation parameters, other fixed variables, and an iterative process.

5. A rigorous upper bound for the optimal value. In this section we investigate the computation of a rigorous upper bound for the optimal value and the certificate of existence of optimal solutions for linear programming problems. Frequently, the transformations to the standard lp-problem lead to data dependencies of the interval input data. For example, describing an unsigned variable as the difference of two nonnegative variables introduces two dependent columns into the system matrix. For interval input data, such dependencies lead to overestimations.

Therefore, we consider the nonstandard problem

$$\min_{x \in F} c^T x, \quad F := \{x \in \mathbf{R}^n : \ Ax \leq a, \ Bx = b, \ \underline{x} \leq x \leq \overline{x}\}, \tag{16}$$

where $A \in \mathbf{R}^{m \times n}$, $B \in \mathbf{R}^{p \times n}$, $c, x \in \mathbf{R}^n$, $a \in \mathbf{R}^m$, and $b \in \mathbf{R}^p$.

F is called the set of *(primal) feasible solutions*, F^* denotes the set of *optimal solutions*, and f^* is the *optimal value*. If F is empty, $f^* := +\infty$. This problem is described by the input data

$$P = (A, B, a, b, c) \in \mathbf{R}^{(m+p+1)n+m+p} \tag{17}$$

and the simple bounds $\underline{x} < \overline{x}$, which may be infinite; that is, $\underline{x}_j := -\infty$ or $\overline{x}_j := +\infty$ for some $j \in \{1, \ldots, n\}$. Hereafter we use the arrangement $0 \cdot (+\infty) = 0 \cdot (-\infty) = 0$. The set of indices in which the simple bounds are both infinite is denoted by

$$J^\infty := \{j \in \{1, \ldots, n\} : \ \underline{x}_j = -\infty \text{ and } \overline{x}_j = +\infty\}, \tag{18}$$

and its complement is

$$J^r := \{1, \ldots, n\} \setminus J^\infty. \tag{19}$$

In many applications J^∞ is empty or contains only a few indices. A feasible point $x \in F$ is called *degenerate* if more than n constraints in F are active, that is, at least $n+1$ constraints hold as equations in x.

Some or all input data of a linear programming problem may be uncertain. We describe these uncertainties by considering a family of lp-problems P, where $P \in \mathbf{P}$ and

$$\mathbf{P} := (\mathbf{A}, \mathbf{B}, \mathbf{a}, \mathbf{b}, \mathbf{c}) \in \mathbf{IR}^{(m+p+1)n+m+p}. \tag{20}$$

To indicate the dependency of the notation above from $P \in \mathbf{P}$, we sometimes write $F(P), F^*(P), f^*(P), x(P)$, etc.

The basic idea for computing a rigorous upper bound is the determination of an interval vector $\mathbf{x}$ which contains for every lp-problem of the family a feasible solution being in the relative interior of F. This solution must be close to an optimal solution but sufficiently far away from degeneracy and infeasibility. The next theorem gives favorable characteristics of $\mathbf{x}$.

THEOREM 5.1. *Let $\mathbf{P} := (\mathbf{A}, \mathbf{B}, \mathbf{a}, \mathbf{b}, \mathbf{c})$ be a family of lp-problems with input data $P \in \mathbf{P}$ and simple bounds $\underline{x} < \overline{x}$. Suppose that there exists an interval vector $\mathbf{x} \in \mathbf{IR}^n$ such that*

$$\mathbf{A}\mathbf{x} \leq \mathbf{a}, \ \underline{x} \leq \mathbf{x} \leq \overline{x}, \tag{21}$$

and

$$\text{for all } B \in \mathbf{B}, \text{ for all } b \in \mathbf{b} \ \exists\, x \in \mathbf{x} : \ Bx = b. \tag{22}$$

Then for every $P \in \mathbf{P}$ there exists a primal feasible solution $x(P) \in \mathbf{x}$, and the inequality

$$\sup_{P \in \mathbf{P}} f^*(P) \leq \overline{f}^* := \sup(\mathbf{c}^T \mathbf{x}) \tag{23}$$

is satisfied. Moreover, if the objective function is bounded from below for every lp-problem with input data $P \in \mathbf{P}$, then each problem has an optimal solution.

Proof. Let $P = (A, B, a, b, c) \in \mathbf{P}$ be a fixed chosen problem. The condition (22) implies that there exists an $x(P) \in \mathbf{x}$ with $B \cdot x(P) = b$, and from condition (21) it follows that $x(P)$ is a primal feasible solution. Hence, $c(P)^T x(P) \geq f^*(P)$, and (23) follows. If the objective function $c(P)^T x$ is bounded from below by a finite number $\underline{f}^*$, then the theory of linear programming proves the existence of optimal solutions for P. □

In the next section we show how a rigorous lower bound for the optimal value $\underline{f}^*$ can be calculated. Now, it remains to describe the algorithm for computing a rigorous upper bound $\overline{f}^*$ and an appropriate interval vector $\mathbf{x}$ satisfying the conditions (21) and (22). We assume that an approximate solution $\tilde{x}$ of the midpoint problem $\check{\mathbf{P}}$ is known. Nothing is assumed about the quality of $\tilde{x}$. The algorithm consists of the following seven steps:

(1) Let ε be the adjusted accuracy of the lp-solver, let η be a number greater than the smallest positive floating point number, let $1.5 \leq \alpha \leq 5$, and let e be the vector which is equal to 1 in every component. Compute the *deflation parameters*

$$\varepsilon_a := \alpha(\mathrm{rad}(\mathbf{a}) + \mathrm{rad}(\mathbf{A})|\tilde{x}| + \varepsilon(|\check{\mathbf{a}}| + |\check{\mathbf{A}}||\tilde{x}|)) + \eta \cdot e \tag{24}$$

and $\underline{\varepsilon}_s, \overline{\varepsilon}_s$ which are defined componentwise by

$$(\underline{\varepsilon}_s)_j := \begin{cases} 0 & \text{if } \underline{x}_j = -\infty, \\ \varepsilon|\underline{x}_j| + \eta & \text{otherwise,} \end{cases} \tag{25}$$

$$(\overline{\varepsilon}_s)_j := \begin{cases} 0 & \text{if } \overline{x}_j = +\infty, \\ \varepsilon|\overline{x}_j| + \eta & \text{otherwise.} \end{cases} \tag{26}$$

(2) Define the perturbed problem $P(\varepsilon) := (\check{\mathbf{A}}, \check{\mathbf{B}}, \check{\mathbf{a}} - \varepsilon_a, \check{\mathbf{b}}, \check{\mathbf{c}})$ with simple bounds $\underline{x}(\varepsilon) := \underline{x} + \underline{\varepsilon}_s$ and $\overline{x}(\varepsilon) := \overline{x} - \overline{\varepsilon}_s$.

(3) Compute an approximate optimal solution $\tilde{\tilde{x}}$ of the perturbed problem $P(\varepsilon)$ using the starting point $\tilde{x}$.
If the lp-solver cannot find an approximate feasible solution, then STOP: *Upper bound infinity.*

(4) Redefine $\tilde{\tilde{x}}$; that is, every component of $\tilde{\tilde{x}}$ smaller than the corresponding component of $\underline{x}(\varepsilon)$ is set equal to this lower bound, and every component of $\tilde{\tilde{x}}$ larger than the corresponding component of the simple bound $\overline{x}(\varepsilon)$ is set equal to this upper bound. Then $\underline{x}(\varepsilon) \leq \tilde{\tilde{x}} \leq \overline{x}(\varepsilon)$.
Set $\tilde{x} := \tilde{\tilde{x}}$.

(5) For $p = 0$ set $\mathbf{x} := \tilde{x}$.
Otherwise, choose an index set $I = \{\beta_1, \ldots, \beta_p\}$ such that the submatrix $\check{\mathbf{B}}_I$ is (approximately) nonsingular (for example, by performing an LU or QR decomposition of the midpoint of $\mathbf{B}$). For the nonsquare linear interval system via input data $\mathbf{B}, \mathbf{b}$ compute via the algorithm of section 4, with respect to $\tilde{x}$ and I, the interval vector $\mathbf{x}$ such that condition (22) is fulfilled.
If this algorithm does not compute an enclosure $\mathbf{x}$, then STOP: *Upper bound infinity.*

(6) If the conditions $\mathbf{Ax} \leq \mathbf{a}$, $\underline{x} \leq \mathbf{x} \leq \overline{x}$ are satisfied, then STOP: *Upper bound* $\overline{f}^* = \sup(\mathbf{c}^T\mathbf{x})$.

(7) Increase the deflation parameters by setting $\varepsilon_a := \alpha(\varepsilon_a + 2\max(rad(\mathbf{x})))$, $\underline{\varepsilon}_s := \alpha(\underline{\varepsilon}_s + 2\max(rad(\mathbf{x})))$, $\overline{\varepsilon}_s := \alpha(\overline{\varepsilon}_s + 2\max(rad(\mathbf{x})))$.

(8) If the lp-solver gave no warning in step (3), then goto step (2).

It follows that in each iteration (i.e., one execution of steps (2)–(8)) the deflation parameters $\varepsilon_a, \underline{\varepsilon}_s, \overline{\varepsilon}_s$ are increased relatively and absolutely by adding positive values $\mathrm{rad}(\mathbf{x})$ and η. Hence, in each iteration the set of primal feasible solutions of the perturbed problems $P(\varepsilon)$ is contained in the relative interior of the set of feasible solutions in the previous iteration. Therefore, these sets are shrinking in each iteration. Either the primal feasible set of $P(\varepsilon)$ becomes empty (then the algorithm terminates in step (3) because the lp-solver cannot find an approximate optimal solution), or the algorithm terminates in step (5) or (6) with an appropriate $\mathbf{x}$. If the algorithm terminates in step (5), then one may choose another method for solving linear interval systems, or the radius $\mathrm{rad}(\mathbf{P})$ must be decreased. In our present implementation of the algorithm we stop after at most 10 iterations.

In the case of sufficiently small $\mathrm{rad}(\mathbf{P}), \varepsilon, \eta$, and under the assumption that the lp-solver and the solver for linear interval systems work appropriately, the algorithm computes an upper bound $\overline{f}^*$ if $F(\check{\mathbf{P}})$ has a nonempty relative interior; i.e., the linear system of equations and inequalities

$$\check{\mathbf{A}}x < \check{\mathbf{a}}, \check{\mathbf{B}}x = \check{\mathbf{b}},\ \underline{x} < x < \overline{x} \tag{27}$$

has a solution x. The property that the relative interior of system (27) is nonempty holds for small data perturbations; that is, the set of feasible solutions must be well-posed. On the other hand, degeneracy does not affect the practicability of this algorithm because we search for an interval vector contained in the relative interior of F. However, the number of iterations may increase for degenerate problems.

If several iterations are executed, then in step (3) the previous approximate optimal solution always can be used as a starting point, and good warm-start facilities of the lp-solver limit the computational effort. We remark that the simplex method

has very good warm-start strategies, but up to now interior point methods have no comparable strategies for perturbed problem instances.

It is not required that the perturbed problem $P(\varepsilon)$ be contained in $\mathbf{P}$, but the crucial property is that the linear inequalities in step (6) hold for $\mathbf{x}$. This latter property implicitly demands that the deflation parameters be chosen such that the set of feasible solutions $F(P(\varepsilon))$ is contained in the relative interior of $F(P)$ for every $P \in \mathbf{P}$.

The index set $I = \{\beta_1, \ldots, \beta_p\}$ in step (5) is chosen such that the components of $\mathbf{x}$ with diameter greater than zero are far away from the simple bounds. Then the test $\underline{x} \leq \mathbf{x} \leq \overline{x}$ is easier to satisfy.

The deflation parameters affect the computational work and the error bound. For example, large values of α decrease the number of iterations, but the quality of the bound becomes worse because $\mathbf{x}$ will be moved towards the relative interior of $F(\check{\mathbf{P}})$ and away from optimality. In our present implementation the value $\alpha = 2.5$ is used. The other parameters $\varepsilon_a, \underline{\varepsilon}_s$, and $\overline{\varepsilon}_s$ are chosen in a heuristic manner in order to comprise the interval input data and the rounding errors.

The deflation parameters are used for shrinking the set of feasible solutions. We mention that in interval arithmetic a contrary technique was used by Caprani and Madsen [4]. This technique widens intervals such that the existence and uniqueness of fixed point problems can be proved with Brouwer's fixed point theorem. It is called epsilon-inflation by Rump [25], and it is theoretically and practically analyzed in many subsequent papers.

Example. For the purpose of illustration we consider a degenerate lp-problem, where the constraints are taken from Vanderbei [29, page 36]

$$
\begin{array}{lrcrcrcll}
\min & - & x_1 & - & x_2 & - & 4x_3 & \text{subject to} & \\
 & & x_1 & & & + & 2x_3 & \leq & 2, \\
 & & & & x_2 & + & 2x_3 & \leq & 2, \\
0 \leq & & x_1, & & x_2, & & x_3 & \leq & 2,
\end{array}
\tag{28}
$$

and which is illustrated in Figure 5.1.

Obviously, $x^* = (0, 0, 1)$ is optimal with value $f^* = -4$. The point x^* is the intersection of four of the facets, not of three facets as one would normally expect; i.e., x^* is degenerate. Moreover, it is not a simple degeneracy caused by redundant constraints, since deleting only one of the above constraints changes the set of feasible solutions.

Now, we replace each nonzero coefficient c of the objective function and the constraints by an interval $[c - r, c + r]$ with $0 \leq r \leq \frac{1}{2}$, yielding an interval problem $\mathbf{P}$. Then (28) is the midpoint problem of $\mathbf{P}$.

To simplify the following illustration of the previous algorithm, we assume that the lp-solver calculates the exact optimal solution, and we set $\epsilon := 0$, $\eta := 0$, $\alpha = \frac{3}{2}$, and $\tilde{x} := x^*$. Then in step (1) we obtain

$$
\epsilon_a = \frac{3}{2}\left(\begin{pmatrix} r \\ r \end{pmatrix} + \begin{pmatrix} r & 0 & r \\ 0 & r & r \end{pmatrix} \cdot \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}\right) = \begin{pmatrix} 3r \\ 3r \end{pmatrix}, \ \underline{\varepsilon}_s = \overline{\varepsilon}_s = 0,
$$

yielding in step (2) the constraints of the perturbed problem

$$
x_1 + 2x_3 \leq 2 - 3r, \quad x_2 + 2x_3 \leq 2 - 3r.
$$

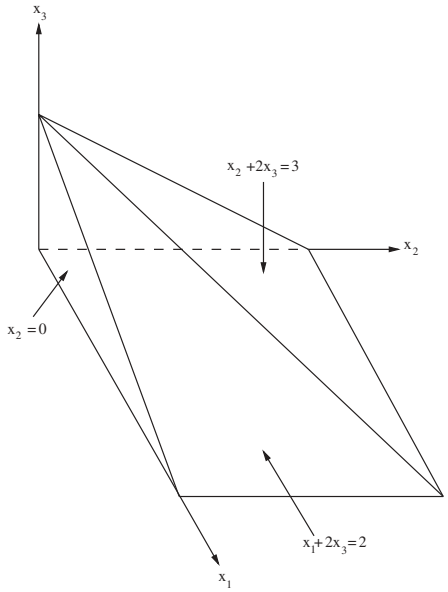


FIG. 5.1. *The set of feasible solutions for the degenerate problem.*

The optimal solution of this perturbed problem is $\tilde{\tilde{x}} = (0, 0, 1 - \frac{3}{2}r)$. In step (5) $p = 0$ and $\mathbf{x} = (0, 0, 1 - \frac{3}{2}r)$. Since $0 \le \mathbf{x} \le 2$ and, furthermore, the inequality

$$[1-r, 1+r] \cdot 0 + [2-r, 2+r]\left(1 - \frac{3}{2}r\right) = \left[2 - 4r + \frac{3}{2}r^2, 2 - 2r - \frac{3}{2}r^2\right] \le [2-r, 2+r]$$

holds for $r \le \frac{1}{2}$, all conditions in step (6) are fulfilled. Hence,

$$\overline{f^*} = \sup\left([-4-r, -4+r] \cdot \left(1 - \frac{3}{2}r\right)\right) = -4 + 7r - \frac{3}{2}r^2. \tag{29}$$

6. A rigorous lower bound for the optimal value. Closely related to problem (16) is the *dual problem*

$$\begin{aligned} \max \quad & a^T y + b^T z + \underline{x}^T u + \overline{x}^T v \quad \text{subject to} \\ & A^T y + B^T z + u + v = c, \\ & y \le 0,\ u \ge 0,\ v \le 0. \end{aligned} \tag{30}$$

A vector $(y, z, u, v)^T \in \mathbf{R}^{2n+m+p}$ satisfying the constraints in (30) is called a *(dual) feasible point*, G denotes the set of dual feasible points in (30), G^* is the set of optimal points, and g^* denotes the optimal value. Observe that with our arrangement the objective value of a dual feasible point $(y, z, u, v)^T$ is finite iff $u_j = 0$ for $\underline{x}_j = -\infty$ and $v_j = 0$ for $\overline{x}_j = +\infty$.

The dual problem is described by the same input data as problem (16), and the theory of linear programming shows that both optimal values f^* and g^* are finite and equal, provided that optimal solutions exist. A feasible solution $(y, z, u, v)^T \in G$ is called *(dual) degenerate* if less than n of its components are nonzero.

The following theorem provides a rigorous lower bound of $f^*(P)$ for all $P \in \mathbf{P}$ by using the dual problem.

THEOREM 6.1. *Let $\mathbf{P} := (\mathbf{A}, \mathbf{B}, \mathbf{a}, \mathbf{b}, \mathbf{c})$ be a family of lp-problems with input data $P \in \mathbf{P}$ and simple bounds $\underline{x} < \overline{x}$. Suppose that there exist interval vectors $\mathbf{y} \in \mathbf{IR}^m$ and $\mathbf{z} \in \mathbf{IR}^p$ such that*

(i) *the sign condition*

$$\mathbf{y} \leq 0 \tag{31}$$

holds;

(ii) *the equations*

$$\begin{aligned} &\text{for all } A \in \mathbf{A}, B \in \mathbf{B}, c \in \mathbf{c}\ \exists\, y \in \mathbf{y}, z \in \mathbf{z}: \\ &(A_{:j})^T y + (B_{:j})^T z = c_j \text{ for } j \in J^\infty \end{aligned} \tag{32}$$

are fulfilled;

(iii) *and further, for the intervals*

$$\mathbf{d}_j := \mathbf{c}_j - (\mathbf{A}_{:j})^T \mathbf{y} - (\mathbf{B}_{:j})^T \mathbf{z} \text{ for } j \in J^r \tag{33}$$

the inequalities

$$\mathbf{d}_j \leq 0, \text{ if } \underline{x}_j = -\infty \text{ and } j \in J^r, \tag{34}$$

$$\mathbf{d}_j \geq 0, \text{ if } \overline{x}_j = +\infty \text{ and } j \in J^r \tag{35}$$

are satisfied.

Then the inequality

$$\inf_{P \in \mathbf{P}} f^*(P) \geq \underline{f}^* := \min \left(\mathbf{a}^T \mathbf{y} + \mathbf{b}^T \mathbf{z} + \sum_{\substack{j \in J^r \\ \overline{\mathbf{d}}_j > 0}} \underline{x}_j \mathbf{d}_j^+ + \sum_{\substack{j \in J^r \\ \underline{\mathbf{d}}_j < 0}} \overline{x}_j \mathbf{d}_j^- \right) \tag{36}$$

is fulfilled, and $\underline{f}^*$ is a finite lower bound of the global minimum value. Moreover, if

(a) *all input data are point data (i.e., $P = \mathbf{P}$);*

(b) *P has an optimal solution (y^*, z^*, u^*, v^*);*

(c) *$\mathbf{y} := y^*$, $\mathbf{z} := z^*$;*

(d) *the quantities in* (33) *and* (36) *are calculated exactly,*

then conditions (i), (ii), *and* (iii) *are satisfied, and the optimal value $f^*(P) = \underline{f}^*$; that is, this lower error bound is sharp for point input data and exact computations.*

Proof. Let $\mathbf{y}, \mathbf{z}$ be interval vectors satisfying conditions (i), (ii), and (iii), and let $y \in \mathbf{y}, z \in \mathbf{z}$ be vectors which satisfy (32) for a fixed chosen problem $P = (A, B, a, b, c) \in \mathbf{P}$. By definition (33),

$$d_j := c_j - (A_{:j})^T y - (B_{:j})^T z \in \mathbf{d}_j \tag{37}$$

for every $j \in J^r$. We define componentwise for $j = 1, \ldots, n$ two vectors $u, v \in \mathbf{R}^n$ by

$$u_j := \begin{cases} d_j & \text{for } j \in J^r \text{ with } d_j > 0, \\ 0 & \text{otherwise,} \end{cases} \tag{38}$$

$$v_j := \begin{cases} d_j & \text{for } j \in J^r \text{ with } d_j < 0, \\ 0 & \text{otherwise.} \end{cases} \tag{39}$$

Using (32), (37), (38), and (39), it follows immediately that the vector (y, z, u, v) satisfies

$$(A_{:j})^T y + (B_{:j})^T z + u_j + v_j = c_j \quad \text{for} \quad j = 1, \ldots, n.$$

The sign condition (31) and the definitions above yield $y \leq 0$, $u \geq 0$, and $v \leq 0$, implying that this vector is dual feasible. Hence, using the inequalities and equations in (16), we obtain for every primal feasible $x \in F$

$$\begin{aligned} &a^T y + b^T z + \underline{x}^T u + \overline{x}^T v \\ &\leq x^T A^T y + x^T B^T z + x^T u + x^T v \\ &= x^T c. \end{aligned}$$

Noting that $u_j = v_j = 0$ for $j \in J^\infty$, it follows that the corresponding terms $\underline{x}_j u_j$ and $\overline{x}_j v_j$ vanish in the dual objective function, and

$$f^*(P) = \min_{x \in F} c^T x \geq a^T y + b^T z + \sum_{j \in J^r} \underline{x}_j u_j + \sum_{j \in J^r} \overline{x}_j v_j. \tag{40}$$

The bounds (33) yield $d_j \in \mathbf{d}_j$ for every $j \in J^r$, and definitions (38) and (39) imply

$$\underline{x}_j u_j = 0 \quad \text{or} \quad \underline{x}_j u_j \in \underline{x}_j \mathbf{d}_j^+ \quad \text{for} \quad j \in J^r, \tag{41}$$

$$\overline{x}_j v_j = 0 \quad \text{or} \quad \overline{x}_j v_j \in \overline{x}_j \mathbf{d}_j^- \quad \text{for} \quad j \in J^r. \tag{42}$$

If we put (41) and (42) into (40), and if we consider all $P \in \mathbf{P}$, we obtain the rigorous lower bound (36). The inequalities (34) and (35) imply that no infinite terms appear in (36). Hence, the lower bound is finite.

For the point input data let (y^*, z^*, u^*, v^*) be a dual optimal solution. Then define $\mathbf{y} := y^*$, $\mathbf{z} := z^*$. If the d_j are calculated exactly, then (38), (39) yield $u = u^*$, $v = v^*$ and the constructed vector $(\mathbf{y}, \mathbf{z}, u, v)$ is the optimal vector of the dual problem. A short inspection shows the validity of (i), (ii), and (iii), and the duality theorem of linear programming proves the last assertion. □

A consequence is the following error bound for standard linear programming problems.

COROLLARY 6.1. *Let $B \in \mathbf{R}^{p \times n}$, $c \in \mathbf{R}^n$, $b, z \in \mathbf{R}^p$, and $d := c - B^T z$. Then the optimal value f^* of the standard lp-problem*

$$\min_{x \in F} c^T x, \quad F := \{x \in \mathbf{R}^n : \ Bx = b, \ 0 \leq x \leq \overline{x}\} \tag{43}$$

with simple upper bound $\overline{x} \in \mathbf{R}^n$, $0 < \overline{x}$, satisfies the inequality

$$f^* \geq b^T z + \sum_{\substack{j=1 \\ d_j < 0}}^{n} \overline{x}_j \cdot d_j. \tag{44}$$

Proof. This follows from Theorem 6.1 by noticing that A is empty, all simple bounds are finite, and $\underline{x} = 0$. Definition (33) yields the vector d for given z. □

Obviously, the bound (44) depends only on the quality of z. If z is close to a dual optimal solution of (43), then f^* is close to the right-hand side of (44). This follows immediately from the dual problem of (43), which is

$$\max b^T z + \overline{x}^T v \quad \text{subject to} \quad c - B^T z = v, v \leq 0. \tag{45}$$

More generally, the lower bound $\underline{f}^*$ in Theorem 6.1 depends mainly on the radius r of the interval input data and on the chosen interval vectors $\mathbf{y}$ and $\mathbf{z}$. From the duality theory, it follows that the quality of this bound is improved if $\mathbf{y}$ and $\mathbf{z}$ are close to or contain the corresponding parts y^*, z^* of the dual optimal solution.

At first we discuss the algorithm for computing $\underline{f}^*$ in the case where all simple bounds are finite. Later the general case is treated.

We assume that approximations $\tilde{y}, \tilde{z}$ of a dual optimal solution (y^*, z^*, u^*, v^*) of the midpoint problem of $\mathbf{P}$ are already computed.

(1) Redefine $\tilde{y}$ by setting all positive components equal to zero.
(2) Set $\mathbf{y} := \tilde{y}$ and $\mathbf{z} := \tilde{z}$.
(3) Compute $\mathbf{d} := \mathbf{c} - \mathbf{A}^T\mathbf{y} - \mathbf{B}^T\mathbf{z}$.
(4) Compute $\underline{f}^*$ with formula (36).

Since all simple bounds are finite, conditions (ii) and (iii) are fulfilled by definition, and because $\mathbf{y} \leq 0$ (see steps (1) and (2)), condition (i) is also satisfied, yielding the rigorous bound $\underline{f}^*$. This bound requires only linear operations between interval quantities. Since $\mathrm{rad}(\mathbf{y}) = \mathrm{rad}(\tilde{y}) = 0$ and $\mathrm{rad}(\mathbf{z}) = \mathrm{rad}(\tilde{z}) = 0$, the bound is linearly decreasing for increasing radius r as long as no additional terms $\underline{x}_j\mathbf{d}_j^+$ or $\overline{x}_j\mathbf{d}_j^-$ occur in the two sums. For large values $|\underline{x}_j|$ and $|\overline{x}_j|$, the bound may be poor if one of the products $\underline{x}_j\mathbf{d}_j^+$ or $\overline{x}_j\mathbf{d}_j^-$ appears in the sums. In this case we recommend putting these indices into J^∞. Then the few equations corresponding to these indices must be added, but the bound will be better.

The additional computational costs for computing $\underline{f}^*$ are of order $0((m+p)\cdot n)$. The computed solution $(\tilde{y}, \tilde{z}, \tilde{u}, \tilde{v})$ approximates an optimal vertex of the dual problem yielding at most n nonzero components. Hence, at most n components of $\mathbf{y}$ and $\mathbf{z}$ are nonzero, and the lower bound needs only $O(n^2)$ operations.

The approximate solution can be computed by any lp-method for large or sparse systems. Since the lower bound requires only a few matrix-vector operations, the sparse structure can be easily utilized. Moreover, degeneracy of the lp-problem does not cause difficulties at all.

Example. We consider the dual problem of (28):

$$\max\ 2y_1 + 2y_2 + 2v_1 + 2v_2 + 2v_3 \text{ subject to}$$

$$\begin{pmatrix} 0 & 1 \\ 1 & 0 \\ 2 & 2 \end{pmatrix} \cdot \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} + \begin{pmatrix} u_1 \\ u_2 \\ u_3 \end{pmatrix} + \begin{pmatrix} v_1 \\ v_2 \\ v_3 \end{pmatrix} = \begin{pmatrix} -1 \\ -1 \\ -4 \end{pmatrix}, \tag{46}$$

$$y_1, y_2, v_1, v_2, v_3 \leq 0, \quad u_1, u_2, u_3 \geq 0.$$

A short inspection shows that a dual optimal solution is given by $y_1^* = y_2^* = -1$ and all other variables are set equal to zero. The optimal value is $g^* = -4 = f^*$. Also the dual problem is degenerate, since less than three components of the optimal solution are nonzero, and the dual optimal solution is not unique. For example, the point $y_1^* = -\frac{1}{2}, y_2^* = -\frac{1}{2}, {v_1}^* = -\frac{1}{2}, {v_2}^* = -\frac{1}{2}$ is also optimal.

To obtain a lower bound $\underline{f}^*$, we set

$$\mathbf{y} := \begin{pmatrix} y_1^* \\ y_2^* \end{pmatrix} = \begin{pmatrix} -1 \\ -1 \end{pmatrix} \leq 0,$$

and a short computation for the interval problem as defined in the previous section yields

$$\mathbf{d}_1 = [-2r, 2r], \quad \mathbf{d}_2 = [-2r, 2r], \quad \mathbf{d}_3 = [-3r, 3r]$$

and the lower bound

$$\begin{aligned} \underline{f}^* &= [2-r,2+r]\mathbf{y}_1 + [2-r,2+r]\mathbf{y}_2 \\ &\quad +2[-2r,0] + 2[-2r,0] + 2[-3r,0] \\ &= -4 - 16r. \end{aligned} \tag{47}$$

Example. In order to illustrate some other effects, we consider the lp-problem

$$\min c^T x \quad \text{subject to} \quad a_{11}x \le a_1, a_{21}x \le a_2, \underline{x} \le x \le \overline{x}, \tag{48}$$

with integer coefficients and simple bounds

$$c = 100,\ a_{11} = 100,\ a_1 = 102,\ a_{21} = -90,\ a_2 = -90,\ \underline{x} = -2,\ \overline{x} = 2. \tag{49}$$

Because of the two inequalities $100x \le 102$, $-90x \le -90$ the set of feasible solutions satisfies $1 \le x \le 1.02$. Hence, this is a bounded and well-conditioned problem with optimal solution $x^* = 1$ and optimal value $f^* = 100$. Now, we allow that all coefficients a of this problem (with the exception of the simple bounds) can vary within a radius r; that is, we consider intervals of the form $[a-r, a+r]$.

For $r \ge 1$ this is a slightly complicated one-dimensional interval problem. The difficulties arise because some point problems within the interval input data contain no feasible solutions, some contain feasible solutions, and some are degenerate and ill-posed. Choosing

$$c = 100,\ a_{11} = 101,\ a_1 = 101,\ a_{21} = -89,\ a_2 = -91 \tag{50}$$

yields coefficients within the radius $r = 1$, and the inequalities $101x \le 101$, $-89x \le -91$ are contradictory. Hence, this problem is not feasible. If we change two coefficients within the interval input data to

$$a_{21} = -90,\ a_2 = -90, \tag{51}$$

then we obtain an ill-posed degenerate problem which has exactly one feasible solution $x = 1$.

The dual problem of (49) is

$$\begin{aligned} &\max 102y_1 - 90y_2 - 2u + 2v \quad \text{subject to} \\ &100y_1 - 90y_2 + u + v = 100,\ y_1, y_2, v \le 0,\ u \ge 0, \end{aligned} \tag{52}$$

yielding the dual optimal solution

$$y_1^* = u^* = v^* = 0,\ y_2^* = -\frac{10}{9}. \tag{53}$$

Assume that the approximate solution coincides with the dual optimal solution. Then step (2) of our algorithm yields

$$\begin{aligned} &\mathbf{y}_1 = y_1^* = 0, \mathbf{y}_2 = y_2^* = -\frac{10}{9} \quad \text{and} \\ &\mathbf{d} = [100-r, 100+r] - [-90-r, -90+r] \cdot \left(-\frac{10}{9}\right) = \left[-\frac{19r}{9}, \frac{19r}{9}\right]. \end{aligned} \tag{54}$$

Now with formula (36) we obtain the lower bound

(55)
$$\inf_{P\in[P]} f^*(P) \geq \min\left([-90-r,-90+r]\cdot\left(-\frac{10}{9}\right) - 2\frac{19r}{9} + 2\left(-\frac{19r}{9}\right)\right) = 100-\frac{86r}{9}.$$

The lower bound decreases linearly for increasing radius r and coincides with the optimal value in the case $r = 0$. This is consistent with Theorem 6.1. Moreover, the influence of the simple bounds can be seen. In the example case of $\underline{x} := 0$, $\overline{x} := 2$ the lower bound is $100-48r/9$, demonstrating that some constraint propagation heuristic for improving the simple bounds can be very useful.

We remark that none of the $O(n^3)$ methods mentioned in section 1 can compute any bounds for this example, since these methods additionally verify that all real problems within the interval problem have optimal solutions and are well-posed, which is not true for this example.

The previous example starts with a midpoint problem which is well-posed and has a finite optimal value. The assumptions of Theorem 6.1 do not require that even one real problem $P \in \mathbf{P}$ have primal feasible solutions. It is interesting to see what happens with the lower bound (36) in such situations.

Example. We consider again example (48) with input data (50). This problem has an empty set of feasible solutions, and $f^* = +\infty$. If we apply *linprog* to this problem, then we get the correct message

the primal appears to be infeasible (and the dual unbounded)

and the approximate values[2]

$$\tilde{f} = 1.015200942446940 \times 10^2, \quad \tilde{x} = 1.015200942446940$$

for the primal problem, and

$$\begin{aligned} &\tilde{y}_1 = -7.846817927076620 \times 10^7, \quad \tilde{y}_2 = -8.904815962188074 \times 10^7, \\ &\tilde{u} = 2.039165444641289 \times 10^{-10}, \quad \tilde{\tilde{v}} = -7.434652599734950 \times 10^{-9} \end{aligned}$$

for the dual problem. Obviously, the calculated approximations $\tilde{f}, \tilde{x}$ are completely wrong. But surprisingly, the approximations of the dual problem satisfy the dual constraint very sharply and, consequently, we obtain the lower bound $1.780964192437591\times 10^8$ which is clearly nearer to $f^* = +\infty$ than $\tilde{f}$. In a branch and bound framework such a bound can serve to eliminate a subproblem. For a varying radius of the interval input data, the lower bound (36) is of the same order of magnitude for all radii $0 \leq r \leq 0.21$. But for $r = 0.22$ we get $-6.171554418163598 \times 10^6$, the sign has changed, and this bound is poor. The reason is that for values $r \leq 0.21$ the first two terms in (36) remain much larger than the last two sums. This behavior changes for values close to or greater than $r = 0.22$.

In the following we consider the algorithm for computing the rigorous lower bound in the general case. This algorithm is very similar to the algorithm for computing an upper bound. Condition (ii) of Theorem 6.1 requires the solution of a linear interval system, which can be carried out as before.

Additionally, variables x_j with $j \in J^r$ may occur which are bounded only on one side. Then the inequalities (34) and (35) must be satisfied. In order to obtain

[2]In this example we display all decimal digits because cancellation occurs in the dual constraint.

appropriate interval vectors $\mathbf{y}$ and $\mathbf{z}$, we solve a slightly perturbed linear programming problem. If we take the midpoint problem $\check{\mathbf{P}} := (\check{\mathbf{A}}, \check{\mathbf{B}}, \check{\mathbf{a}}, \check{\mathbf{b}}, \check{\mathbf{c}})$ and change $\check{\mathbf{c}}$ to $c(\varepsilon)$ componentwise by

$$c_j(\varepsilon) := \begin{cases} \check{\mathbf{c}}_j + \varepsilon_j & \text{if } \underline{x}_j = -\infty, \\ \check{\mathbf{c}}_j - \varepsilon_j & \text{if } \overline{x}_j = +\infty, \\ \check{\mathbf{c}}_j & \text{otherwise,} \end{cases} \tag{56}$$

where $\varepsilon_j > 0$, then the exact dual optimal solution $(y(\varepsilon)^*, z(\varepsilon)^*, u(\varepsilon)^*, v(\varepsilon)^*)$ of this perturbed problem $P(\varepsilon)$ (provided the solution exists) satisfies for $\underline{x}_j = -\infty$ the equation

$$\check{\mathbf{c}}_j + \varepsilon_j = (\check{\mathbf{A}}_{:j})^T y(\varepsilon)^* + (\check{\mathbf{B}}_{:j})^T z(\varepsilon)^* + v_j(\varepsilon)^*. \tag{57}$$

Notice that ${u_j}^*(\varepsilon)$ must be zero because of the term $\underline{x}_j u_j$ in the objective function of the dual problem. Hence,

$$d_j(\varepsilon) := \check{\mathbf{c}}_j - (\check{\mathbf{A}}_{:j})^T y(\varepsilon)^* + (\check{\mathbf{B}}_{:j})^T z(\varepsilon)^* = v_j(\varepsilon)^* - \varepsilon_j < 0, \tag{58}$$

and similarly for $\overline{x}_j = +\infty$ we obtain

$$d_j(\varepsilon) := \check{\mathbf{c}}_j - (\check{\mathbf{A}}_{:j})^T y(\varepsilon)^* + (\check{\mathbf{B}}_{:j})^T z(\varepsilon)^* = u_j(\varepsilon)^* + \varepsilon_j > 0. \tag{59}$$

Summarizing, for the perturbed problem $P(\varepsilon)$ the conditions (34) and (35) are fulfilled, provided $\varepsilon_j > 0$ and a sufficiently good approximation $\tilde{y}, \tilde{z}, \tilde{u}, \tilde{v}$ of the optimal solution of $P(\varepsilon)$ is known. In order for these conditions to also hold for the interval problem $\mathbf{P}$, we have to comprise the interval input data and the rounding errors into ε_j as for the upper bound in the previous section.

We assume that approximate dual solutions $\tilde{y}, \tilde{z}$ of the midpoint problem $\check{\mathbf{P}}$ are known. Nothing is assumed about the quality of these approximations. The algorithm consists of the following steps:

(1) Compute the deflation parameters

$$\begin{aligned} \varepsilon_j \;=\; & \alpha(\operatorname{rad}(\mathbf{c}_j) + \operatorname{rad}(\mathbf{A}_{:j})^T|\tilde{y}| + \operatorname{rad}(\mathbf{B}_{:j})|\tilde{z}|) \\ & + \alpha \cdot \varepsilon(|\check{\mathbf{c}}_j| + |(\check{\mathbf{A}}_{:j})^T|\,|\tilde{y}| + |(\check{\mathbf{B}}_{:j})^T|\,|\tilde{z}|) + \eta \end{aligned} \tag{60}$$

for j with $\overline{x}_j = +\infty$ or $\underline{x}_j = -\infty$.

(2) Define the perturbed problem $P(\varepsilon) := (\check{\mathbf{A}}, \check{\mathbf{B}}, \check{\mathbf{a}}, \check{\mathbf{b}}, c(\varepsilon))$, where $c(\varepsilon)$ is given by (56).

(3) Compute an approximate dual solution $\tilde{\tilde{y}}, \tilde{\tilde{z}}$ of the perturbed problem $P(\varepsilon)$ using the starting point $\tilde{y}$, $\tilde{z}$.
If the lp-solver cannot find an approximate solution, then STOP: *Lower bound minus infinity.*

(4) Redefine $\tilde{\tilde{y}}$ by setting all positive components equal to zero, and set $\tilde{y} := \tilde{\tilde{y}}$, $\tilde{z} := \tilde{\tilde{z}}$.

(5) If $|J^\infty| = 0$, set $\mathbf{y} := \tilde{y}$ and $\mathbf{z} := \tilde{z}$. Otherwise, choose an index set $I = \{\beta_1, \ldots, \beta_{|J^\infty|}\}$ consisting initially of the indices corresponding to the components of $\tilde{z}$, and if more indices for I are required (i.e., $|J^\infty| > p$), then choose those following indices which correspond to the components of y with largest absolute values. For $\tilde{y}$, $\tilde{z}$, and I compute via the algorithm for non-square linear interval systems (cf. section 4) interval vectors $\mathbf{y}$ and $\mathbf{z}$ such that condition (ii) of Theorem 6.1 is satisfied. If this algorithm does not compute enclosures $\mathbf{y}$ and $\mathbf{z}$, then STOP: *Lower bound minus infinity.*

(6) If the conditions (i) and (iii) of Theorem 6.1 are satisfied, then STOP: *Lower bound is* $\underline{f}^*$.
(7) Set $\varepsilon_j = \overline{\alpha}(\varepsilon_j + \eta)$ for j with $\overline{x}_j = +\infty$ or $\underline{x}_j = -\infty$.
(8) If the lp-solver gave no warning in step (3), then goto step (2).

The linear interval system (32) must be solved with $m + p$ variables and $|J^\infty|$ equations. Notice that in many applications, $m + p$ is much larger than the number of unbounded variables $|J^\infty|$, yielding only a small quadratic linear interval system.

Several modifications of this algorithm are possible and may yield improvements, at least for special classes of problems. For example, in cases where some of the simple bounds are zero and some are very large, it is advantageous to choose the perturbed problem by defining

$$c_j(\varepsilon) := \begin{cases} \check{\mathbf{c}}_j + \varepsilon_j & \text{if } |\underline{x}_j| \geq |\overline{x}_j|, \\ \check{\mathbf{c}}_j - \varepsilon_j & \text{if } |\underline{x}_j| < |\overline{x}_j|. \end{cases} \tag{61}$$

Also other choices of the deflation parameters may improve the computing time and the bounds.

Last, we want to mention that a seemingly good modification of the above algorithm would be to convert free variables as the difference of two nonnegative variables. Then the computation of an enclosure for the linear interval system corresponding to the free variables would be avoided. However, such a transformation leads to an ill-posed linear programming problem which contains in each neighborhood of the input data problems with an empty dual feasible domain. This would imply the trivial lower bound minus infinity.[3]

7. Applications. A frequently used approach for solving global optimization problems is to use a branch and bound framework together with relaxations; see, for example, Floudas [5], Quesada and Grossmann [23], and the literature cited therein. A linear relaxation is an lp-problem such that each feasible solution of the global optimization problem is also feasible for the linear relaxation, and the linear objective function is an underestimator of the original objective function. Hence, a linear relaxation provides a lower bound for the global minimum value.

In order to discard subproblems in a branch and bound algorithm, it is important to have fast methods available for computing rigorous lower bounds for the optimal value of a linear relaxation. If this lower bound is greater than a known upper bound for the global minimum value, then this subproblem can be eliminated because it cannot contain a global minimizer.

Therefore, computing rigorous upper bounds of the optimal value and certifying that optimal solutions exist are also important. But in contrast to the rigorous lower bound, which has to be computed in each branching step, the computation of the rigorous upper bound and the certification of existence of optimal solutions is only necessary once for the best approximate solution at the end of the branch and bound algorithm. However, it is also important that the upper bound should be computable in the degenerate case and for large systems.

A similar situation occurs for mixed integer problems. There the linear relaxation is defined by relaxing the integer variables, i.e., the integer variables are treated as continuous variables. Rigorous bounds of linear programming problems are also required for verification methods of nonlinear minimax problems. For details the reader is referred to the recent manuscript of Kearfott [12].

[3] I wish to thank Arnold Neumaier who told me this observation at a conference in Dagstuhl (2003).

8. Numerical experiments. The following results were obtained by using MATLAB [14] and the interval toolbox INTLAB (see Rump [28]). In all experiments interval input data $\mathbf{P}$ are considered with $\mathrm{rad}(\mathbf{P}) = r \cdot \check{\mathbf{P}}$; that is, relative perturbations with respect to the midpoint problem are treated.

At first, we look at the two examples of section 2. For the first example, defined by (1) and (2), we obtain for the simple upper bounds $\overline{x} = (30, 30, 30)^T$ and $r = 1.0 \times 10^{-5}$ the rigorous upper bound $\overline{f}^* = -5.99982 \times 10^1$ and the rigorous lower bound $\underline{f}^* = -6.00042 \times 10^1$. Since both bounds are finite, Theorem 5.1 shows that all lp-problems within the interval data have optimal solutions and, moreover, the computed enclosure

$$\mathbf{x} = \begin{pmatrix} [2.421003440006751 \times 10^{-14}, 2.421003440006753 \times 10^{-14}] \\ [7.477872102096182 \times 10^{-16}, 7.477872102096187 \times 10^{-16}] \\ [9.999799997799970, 1.000020000219998] \end{pmatrix}$$

contains for every lp-problem a primal feasible solution that is close to optimality. For $r = 1.0 \times 10^{-2}$, the lower bound $\underline{f}^* = -6.4200 \times 10^1$ and the upper bound $\overline{f}^* = -5.8174 \times 10^1$ are computed. In the case of simple upper bounds $\overline{x} = (10^{10}, 10^{10}, 10^{10})^T$ and r as before, although *linprog* gives the warning that *both the primal and the dual appear to be infeasible*, we obtain, somewhat surprisingly, $\underline{f}^* = -2.42318 \times 10^3$ and $\overline{f}^* = -4.88103 \times 10^1$, and thus the certification that for all lp-problems within the interval data there exist optimal solutions. The computed enclosure is

$$\mathbf{x} = \begin{pmatrix} [4.631674545609419 \times 10^{-3}, 4.631674545609423 \times 10^{-3}] \\ [5.591684607388289, 5.591684607388292] \\ [4.403483715866093, 4.403883720266105] \end{pmatrix}.$$

Typically, in branch and bound algorithms a subproblem is discarded if the lp-solver detects infeasibility (see, for example, Borchers and Mitchell [3]). Observe that the previous lower bound, although very poor when compared with the optimal value, would prevent a branch and bound algorithm from discarding subproblems which contain optimal points.

For the second example (cf. (3)) we display in Table 8.1 below, for the dimensions $n = 8, \ldots, 12$ and fixed radius $r = 1.0 \times 10^{-12}$, the 2-norm condition number of A; the symbol $*$ if *linprog* gives the warning that *the primal appears to be infeasible (and the dual unbounded)*; and the error bounds $\underline{f}^*$ and $\overline{f}^*$.

TABLE 8.1
Results for the second example (3).

n	Condition	Warning	$\underline{f}^*$	$\overline{f}^*$
8	2.0×10^7		-3.3600×10^{-18}	2.1816×10^{-28}
9	2.9×10^8	*	-9.0261×10^{-4}	9.2276×10^{-20}
10	4.5×10^9	*	-3.6902×10^7	4.1743×10^{-13}
11	6.0×10^{10}	*	-1.1294×10^8	3.3209×10^{-15}
12	8.7×10^{11}	*	-3.4427×10^8	$+\infty$

Rigorous upper and lower bounds are computed until dimension $n = 11$, implying the existence of optimal solutions for all lp-problems within the interval input data and, furthermore, the upper bound is close to the optimal value. The lower bound is poor for $n \geq 10$, which is caused by the unsatisfactory dual approximation computed by *linprog*. However, the primal approximation and the enclosure $\mathbf{x}$ are rather fair regarding the large condition number. Below we display the components $\mathbf{x}(1), \mathbf{x}(11), \mathbf{s}(1)$, and $\mathbf{s}(11)$ in the case of dimension $n = 11$:

$$\begin{pmatrix} [0.9999986005449631, 1.000001399410638] \\ \\ [0.9902735279885911, 1.009726272224163] \\ \\ \left[1.779069799688461 \times 10^{-15}, 1.779069799688465 \times 10^{-15}\right] \\ \\ \left[1.623200692326961 \times 10^{-21}, 1.623200692326963 \times 10^{-21}\right] \end{pmatrix}.$$

Note that the optimal solution for the midpoint problem ${x_i}^* = 1, {s_i}^* = 0$ for $i = 1, \ldots, n$ is close to this enclosure. It is only for dimension $n \geq 12$ that an upper bound cannot be computed, and the existence of optimal solutions cannot be proved. This is not surprising since the radius is almost equal to the reciprocal of the condition number.

Last, we consider some random test problems, whose construction is given in Rosen and Suzuki [24]. All components of the primal solution x^* are uniformly distributed in the interval $(0, 1)$. The lower and upper bounds $\underline{x}_i$ and $\overline{x}_i$ are set equal to zero and one for every component x_i, respectively. The dual vector y^* is generated by uniformly distributing $n - p$ components in $(-1, 0)$, while the remaining components are set equal to zero. The coefficients of z^*, A, and B are uniformly distributed in $(-1, 1)$. The right-hand side a is chosen such that the first $n - p$ inequalities are active. The right-hand side is $b := Bx^*$. The coefficients of the objective function are generated by the equation $c := A^T y^* + B^T z^*$. This construction ensures that the optimum is known; that is, x^* is primal optimal and $(y^*, z^*, 0, 0)$ is dual optimal. For the following numerical results, we consider interval input data with $\mathrm{rad}(\mathbf{P}) = r \cdot \check{\mathbf{P}}$, $r := 10^{-8}$, and $\check{\mathbf{P}}$ is the previously defined random problem.

In Table 8.2 we display the number of inequalities m, the number of variables n, the number of equations p, the relative error $|\overline{f}^* - \underline{f}^*|/|\underline{f}^*|$ of the rigorous bounds, and the ratios $t_{\underline{f}^*}/t_s$, $t_{\overline{f}^*}/t_s$, where t_s denotes the time required by *linprog* applied to the midpoint problem, $t_{\underline{f}^*}$ is the time for the lower bound, and $t_{\overline{f}^*}$ denotes the time required by the upper bound.

The symbol $*^1$ means that in this case *linprog* gave the warning

the dual appears to be infeasible (and the primal unbounded)

and has computed poor approximate optimal solutions with an accuracy of only two decimal digits. Note that, because of the above construction, there exist optimal solutions. The symbol $*^2$ means that *linprog* gives results without any warning, but the accuracy of the dual optimal solution was poor, yielding a poor rigorous lower bound. However, the upper bound $\overline{f}^*$ was very accurate.

From Table 8.2 it follows that the bounds are close together with respect to the interval data, provided sufficiently good approximations are calculated by *linprog*. The time for the lower bound $t_{\underline{f}^*}$ is only a fraction of the time t_s needed by the lp-solver, and only one iteration is required for these examples. The time for the

TABLE 8.2
Results for randomly generated problems.

m, n, p	$\lvert\overline{f}^* - \underline{f}^*\rvert / \lvert\underline{f}^*\rvert$	$t_{\underline{f}^*}/t_s$	$t_{\overline{f}^*}/t_s$
20, 10, 0	1.5777×10^{-6}	0.092	1.283
100, 20, 0	7.3356×10^{-7}	0.025	1.240
300, 30, 0	2.8784×10^{-6}	0.0039	0.985
500, 40, 0	2.7805×10^{-6}	0.001	0.947
1000, 50, 0	$*^1 : \underline{f}^* = 1.5015 \times 10^1$ $\overline{f}^* = \infty$	0.002	1.031
0, 20, 10	1.01822×10^{-5}	0.198	2.207
0, 100, 30	$*^2 : 5.1683 \times 10^{-2}$	0.202	2.128
0, 1000, 300	$*^2 : 2.1452 \times 10^{-2}$	0.021	1.160
20, 10, 5	8.0556×10^{-5}	0.186	2.583
50, 20, 10	2.2092×10^{-6}	0.065	2.383
200, 40, 20	$*^2 : 7.2848 \times 10^{-2}$	0.007	3.316

upper bound $t_{\overline{f}^*}$ is larger or of the same order of magnitude as t_s. The reason is that for the upper bound an approximate solution of the perturbed problem must be computed in the first step, and *linprog* allows no reoptimization. Except for the last three cases in Table 8.2, which have required two to three iterations, only one iteration was necessary.

Because of the curious results, we discuss separately a randomly generated problem, where $m = 500$, $n = 100$, and $p = 50$. By our construction we know the optimal value $f^* = 2.6136 \times 10^1$. The routine *linprog* gave the warning that *the dual appears to be infeasible (and the primal unbounded)* and has computed the poor approximate optimal value -1.5733 and an approximation of a primal feasible point, which is not within the simple bounds. One consequence was that the rigorous upper bound was computed equal to $+\infty$. However, the rigorous lower bound was equal to 8.8983, closer to the optimal value than the approximate value. The required time for *linprog* was 2538 seconds, and the time for the lower bound was 0.7 seconds.

9. Conclusions and future work. In this paper algorithms for computing rigorous lower and upper bounds of the optimal value of a linear programming problem are considered. It turns out that such bounds can also be useful for well-posed problems because even commercial solvers like the MATLAB routine *linprog* may compute curious and unsatisfactory results.

If linear relaxations especially are used in branch and bound methods, then these rigorous bounds are advantageous because the user cannot see and judge the intermediate results, and solutions of the original problem cannot drop away. A benefit is that the lower bound, which must be computed in each branching step, needs only a small part of the computing time required for the lp-solver.

For interval input data the bounds allow a rough sensitivity analysis. This is usually not possible in classical sensitivity analysis for linear programming problems, where only special input data are permitted to vary.

In our future work we want to generalize these results to convex programming problems and to use these bounds in algorithms for mixed nonlinear integer problems. Moreover, by using software other than the MATLAB optimization toolbox, we want to investigate experimentally the range of applicability of these rigorous error bounds for large and sparse problems.

REFERENCES

[1] G. Alefeld and J. Herzberger, *Introduction to Interval Computations*, Academic Press, New York, 1983.

[2] H. Beeck, *Linear Programming with Inexact Data,* Technical report 7830, Abteilung Mathematik, TU München, Munich, Germany, 1978.

[3] B. Borchers and J. E. Mitchell, *An improved branch and bound algorithm for mixed integer nonlinear programs*, Comput. Oper. Res., 21 (1994), pp. 359–367.

[4] O. Caprani and K. Madsen, *Iterative methods for interval inclusion of fixed points*, BIT, 18 (1978), pp. 42–51.

[5] C. A. Floudas, *Deterministic Global Optimization—Theory, Methods and Applications*, Nonconvex Optimization and Its Applications 37, Kluwer Academic Publishers, Dordrecht, Boston, London, 2000.

[6] E. R. Hansen, *Global Optimization Using Interval Analysis*, Marcel Dekker, New York, 1992.

[7] *ILOG CPLEX* 7.1 *User's Manual*, ILOG, France, 2001.

[8] C. Jansson, *A self-validating method for solving linear programming problems with interval input data*, in Scientific Computation with Automatic Result Verification, Comput. Suppl. 6, Springer, Vienna, 1988, pp. 33–45.

[9] R. B. Kearfott, *On proving existence of feasible points in equality constrained optimization problems*, Math. Programming, 83 (1998), pp. 89–100.

[10] R. B. Kearfott, *On Verifying Feasibility in Equality Constrained Optimization Problems*, Technical report, Department of Mathematics, University of Southwestern Louisiana, Lafayette, LA, 1996. Available online at http://interval.louisiana.edu/preprints/big_constrai.pdf

[11] R. B. Kearfott, *Rigorous Global Search: Continuous Problems*, Kluwer Academic Publishers, Dordrecht, 1996.

[12] R. B. Kearfott, *On Verification of Solutions to Nonlinear Minimax Problems*, manuscript, 2002.

[13] R. Krawczyk, *Fehlerabschätzung bei linearer Optimierung,* in Interval Mathematics, K. Nickel, ed., Lecture Notes in Comput. Sci. 29, Springer-Verlag, Berlin, 1975, pp. 215–222.

[14] *MATLAB User's Guide, Version* 6, The MathWorks Inc., Natick, MA, 2000.

[15] S. Mehrotra, *On the implementation of a primal-dual interior point method*, SIAM J. Optim., 2 (1992), pp. 575–601.

[16] R. E. Moore, *Methods and Applications of Interval Analysis*, SIAM Studies in Appl. Math. 2, SIAM, Philadelphia, 1979.

[17] *NAG Foundation Toolbox User's Guide*, Numerical Algorithms Group, Ltd., Oxford, UK, 1995.

[18] *NETLIB Linear Programming Library*, http://www.netlib.org/lp.

[19] A. Neumaier, *Interval Methods for Systems of Equations*, Encyclopedia of Mathematics and Its Applications 37, Cambridge University Press, Cambridge, UK, 1990.

[20] A. Neumaier, *Introduction to Numerical Analysis*, Cambridge University Press, Cambridge, UK, 2001.

[21] A. Neumaier and O. Shcherbina, *Safe bounds in linear and mixed-integer programming*, Math. Program., 99 (2004), pp. 283–296.

[22] F. Ordóñez and R. M. Freund, *Computational experience and the explanatory value of condition measures for linear optimization*, SIAM J. Optim., 14 (2003), pp. 307–333.

[23] I. Quesada and I. E. Grossmann, *A global optimization algorithm for linear fractional and bilinear programs*, J. Global Optim., 6 (1995), pp. 39–76.

[24] J. B. Rosen and S. Suzuki, *Construction of nonlinear programming test problems*, Comm. ACM, 8 (1965), p. 113.

[25] S. M. Rump, *Kleine Fehlerschranken bei Matrixproblemen*, Ph.D. thesis, Universität Karlsruhe, Karlsruhe, Germany, 1980.

[26] S. M. Rump, *Solving algebraic problems with high accuracy,* in A New Approach to Scientific Computation, W. Kulisch and W. L. Miranker, eds., Academic Press, New York, 1983, pp. 51–120.

[27] S. M. Rump, *Validated solution of large linear systems*, in Validation Numerics: Theory and Applications, Comput. Suppl. 9, R. Albrecht, G. Alefeld, and H. J. Stetter, eds., Springer-Verlag, Berlin, 1993, pp. 191–212.

[28] S. M. Rump, *INTLAB—Interval Laboratory, Version* 4.1.2, 1998. http://www.ti3.tu-harburg.de/rump/intlab/index.html.

[29] R. J. Vanderbei, *Linear Programming: Foundations and Extensions*, Kluwer Academic Publishers, Dordrecht, 1996.

SIAM J. OPTIM.
Vol. 14, No. 3, pp. 936–937

CORRIGENDUM: SEMIDEFINITE PROGRAMS: NEW SEARCH DIRECTIONS, SMOOTHING-TYPE METHODS, AND NUMERICAL RESULTS*

CHRISTIAN KANZOW† AND CHRISTIAN NAGEL†

Abstract. We correct an error in Lemma 4.2 of [C. Kanzow and C. Nagel, *Semidefinite programs: New search directions, smoothing-type methods, and numerical results*, SIAM J. Optim., 13 (2002), pp. 1–23]. With this correction, all results in that paper remain true.

Key words. semidefinite programs, smoothing-type methods, Newton's method, global convergence, superlinear convergence

AMS subject classifications. 90C22, 90C46

DOI. 10.1137/S1052623403433109

Let $L_A(X) := AX + XA$ be the Lyapunov operator associated with a given matrix $A \in \mathbb{R}^{n\times n}$. In Lemma 4.2(c) of [1] we asserted that the composition $L_A \circ L_B$ is strongly monotone for two symmetric positive definite matrices $A, B \succ 0$. While the given proof holds if A and B commute, inequality (4.8) in [1], namely

$$\operatorname{tr}(X(BA + AB)X) \geq 0,$$

does not hold in general for all $X \in \mathcal{S}^{n\times n}$: Setting

$$A := \begin{pmatrix} 10 & 5 \\ 5 & 6 \end{pmatrix}, \quad B := \begin{pmatrix} 1 & 2 \\ 2 & 6 \end{pmatrix}, \quad X := \begin{pmatrix} 15 & -7 \\ -7 & 3 \end{pmatrix},$$

an easy calculation shows that $A, B \succ 0$ but $\operatorname{tr}(X(BA + AB)X) = -588 < 0$. Therefore, to prove part (c), (d) of Lemma 4.2 in [1], we have to add the assumption

$$BA + AB \succeq 0. \tag{1}$$

In Proposition 4.4 we used Lemma 4.2(d) with $A := E - S$ and $B := E - X$ (where $E := \left(X^2 + S^2 + 2\tau^2 I\right)^{1/2}$ for some $\tau > 0$) in order to show that a certain linear mapping is bijective. We now have to prove that these matrices satisfy inequality (1). This will be done in the following Lemma.

LEMMA 1. *Let* $E := \left(X^2 + S^2 + 2\tau^2 I\right)^{1/2}$, $A := E - S$, *and* $B := E - X$. *Then* $BA + AB \succeq 0$.

Proof. We have

*Received by the editors August 11, 2003; accepted for publication (in revised form) September 25, 2003; published electronically May 17, 2004.
http://www.siam.org/journals/siopt/14-3/43310.html
†Institute of Applied Mathematics and Statistics, University of Würzburg, Am Hubland, 97074 Würzburg, Germany (kanzow@mathematik.uni-wuerzburg.de, nagel@mathematik.uni-wuerzburg.de). This research was partially supported by the DFG (Deutsche Forschungsgemeinschaft).

$$\begin{aligned} BA + AB &= (E - X)(E - S) + (E - S)(E - X) \\ &= E^2 - XE - ES + XS + E^2 - SE - EX + SX \\ &= X^2 + S^2 + 2\tau^2 I - E(X + S) - (X + S)E + E^2 + XS + SX \\ &= (X + S - E)^2 + 2\tau^2 I \succeq 0. \end{aligned}$$

This completes the proof. □

REFERENCE

[1] C. Kanzow and C. Nagel, *Semidefinite programs: New search directions, smoothing-type methods, and numerical results*, SIAM J. Optim., 13 (2002), pp. 1–23.

SIAM J. OPTIM.
Vol. 14, No. 4, pp. 939–964

ON THE CONVERGENCE OF ASYNCHRONOUS PARALLEL PATTERN SEARCH*

TAMARA G. KOLDA† AND VIRGINIA J. TORCZON‡

Abstract. In this paper we prove global convergence for asynchronous parallel pattern search. In standard pattern search, decisions regarding the update of the iterate and the step-length control parameter are synchronized implicitly across all search directions. We lose this feature in asynchronous parallel pattern search since the search along each direction proceeds semiautonomously. By bounding the value of the step-length control parameter after any step that produces decrease along a single search direction, we can prove that all the processes share a common accumulation point and, if the function is continuously differentiable, that such a point is a stationary point of the standard nonlinear unconstrained optimization problem.

Key words. asynchronous parallel optimization, pattern search, unconstrained optimization, global convergence analysis

AMS subject classifications. 90C56, 90C30, 65K05, 65Y05, 68W15

DOI. 10.1137/S1052623401398107

1. Introduction.

Asynchronous parallel pattern search (APPS) was introduced in [5] as a way to solve in a parallel or distributed computing environment nonlinear optimization problems of the form

$$\min_{x \in \mathbb{R}^n} f(x), \quad \text{where } f : \mathbb{R}^n \to \mathbb{R}. \tag{1.1}$$

In this paper, we prove that a subsequence of the sequence of iterates produced by APPS converges to a stationary point of (1.1), when f is continuously differentiable.

To do so, we build on the global convergence results for pattern search established in [7, 10, 11]. What distinguishes this analysis from the earlier work is the need to address the new concerns introduced by the asynchronism. The analyses in [7, 10, 11] rely on the fact that the more usual implementations of pattern search have complete knowledge of information acquired during the course of the search when making decisions about how to proceed. In contrast, APPS partitions out each search direction to a single process and, to eliminate idle time, does away with the close synchronization of the searches along each direction. This means that the search along the single direction governed by an individual process is allowed to proceed semiautonomously. By this we mean that each process is allowed to make its own

*Received by the editors November 13, 2001; accepted for publication (in revised form) September 8, 2003; published electronically May 25, 2004.

http://www.siam.org/journals/siopt/14-4/39810.html

†Computational Sciences and Mathematics Research Department, Sandia National Laboratories, Livermore, CA 94551–9217 (tgkolda@sandia.gov). The research of this author was sponsored by the Mathematical, Information, and Computational Sciences Division at the U.S. Department of Energy and by Sandia National Laboratories, a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the U.S. Department of Energy under contract DE-AC04-94AL85000.

‡Department of Computer Science, College of William & Mary, P.O. Box 8795, Williamsburg, VA 23187–8795 (va@cs.wm.edu). The research of this author was funded by the Computer Science Research Institute at Sandia National Laboratories and by the National Science Foundation under grant CCR-9734044.

decisions regarding the update of the iterate and the length of the next step, based only on the information currently available to it, even though that information may not be up-to-date with respect to the other processes. Further, there is no single controlling process. Instead, information between processes is exchanged intermittently so that eventually all processes learn of every reasonable candidate for the minimizer. The only assumption we make is that information about success (i.e., a decrease in the value of f) on one process reaches all other processes in a finite amount of time. We make no assumption about the order in which such information is received. Thus the processes act as a loose confederation of agents working toward a single goal: the identification of a stationary point of (1.1). The advantage of allowing processes to proceed semiautonomously is that we can eliminate synchronization barriers so that we achieve good computational performance when working in a parallel or distributed computing environment, as our tests in [5] demonstrate.

The critical issue for our analysis is that APPS makes decisions about updating the length of the next step and the best point in the absence of complete information about the progress of the searches along the other directions. Therefore, at any given time in the search, neither the value of the parameter each process uses to determine the length of the step nor the value of the best point may be the same across participating processes. Another minor aspect in which we differ from previous analysis is that we do not fix the contraction and expansion parameters used to update the step lengths. These differences require significant extensions to the analyses found in [7, 10, 11]. The key to safeguarding the overall outcome of the search lies in bounding the values which the parameter that controls the lengths of the steps is allowed to assume after any step that produces decrease on f (i.e., after a *successful* step).

In section 2 we describe a synchronous variant of parallel pattern search and use it to motivate APPS. In section 3 we outline APPS and introduce the extensive notation required for our analysis. We hasten to add that most of this bookkeeping, which is essential to our analysis, is not required in practice. A full treatment of the practical design and implementation of APPS is deferred to [5]. Since the notational overhead required for the analysis is significant, we refer interested readers to [6] for an example of APPS applied to a simple function, an illustration of the associated notation, and a discussion of those features of the asynchronous algorithms that most complicate the analysis. In this paper, we concentrate on the analysis, which is broken into four parts, covered in sections 5–8. In section 9 we close with some remarks regarding further extensions that could be made to the analysis.

Standard notation. We denote by $\mathbb{R}$, $\mathbb{Q}$, $\mathbb{Z}$, and $\mathbb{N}$ the sets of real, rational, integer, and natural numbers, respectively.

We use $\text{pow}(\Lambda, \ell)$ to indicate that Λ is raised to the power ℓ, so that $\text{pow}(\Lambda, \ell) \equiv \Lambda^\ell$. We adopt this notational convention to eliminate any ambiguities that could arise when we introduce superscripts for use as indices.

2. Parallel pattern search. We start by considering a *synchronous* version of parallel pattern search (PPS) to clarify the notation and motivate APPS.

We assume that we have p independent processes, each of which is generating a sequence of trial points. We denote the set of processes as

$$\mathcal{P} = \{1, \ldots, p\}.$$

We work with a finite set of search directions

$$\mathcal{D} = \{d_1, \ldots, d_p\} = \{Bc_1, \ldots, Bc_p\}, \tag{2.1}$$

where

- $B \in \mathbb{R}^{n\times n}$ is a real nonsingular matrix,
- $c_i \in \mathbb{Q}^n$ for each $i \in \mathcal{P}$, and
- the vectors in the set $\mathcal{D}$ form a positive spanning set [7] for $\mathbb{R}^n$.

To each process $i \in \mathcal{P}$ we assign the constant search direction $d_i \in \mathcal{D}$. We constrain the vectors c_i, $i = \{1, \dots, p\}$, to the rationals to ensure that all iterates lie on a *rational lattice*, which, as we see in section 5, is required for the proof of Theorem 5.2. However, we allow a mapping of the rational vectors c_i to the real vectors d_i through the use of a fixed real nonsingular matrix B.

We denote by x_i^k the *best point* (i.e., one with the least function value) known by process i at iteration k. We denote by Δ_i^k the scalar that controls the length of the step taken along the direction d_i to construct a new trial point at iteration k. We refer to Δ_i^k as the *step-length control parameter*. For the synchronous version of pattern search, the subscript i on x and Δ is redundant since the synchronization ensures that the values of x_i^k and Δ_i^k are equivalent for all $i \in \mathcal{P}$; however, this subscript becomes meaningful in the asynchronous case, so we introduce the notation here for comparison.

Each process $i \in \mathcal{P}$ constructs a trial point by computing

$$x_i^k + \Delta_i^k d_i \tag{2.2}$$

and then evaluates f at this point. After the evaluation has finished on process i, process i broadcasts the result to all the other processes in $\mathcal{P}$ and then waits until it has received results from all the other processes in $\mathcal{P}$. This is the point of synchronization; no further action can be taken on process i until all the results from all the other processes in $\mathcal{P}$ are known. Once all p results are known to all p processes, a decision is made simultaneously as to which point is now best, and then x_i^k and Δ_i^k are updated to produce x_i^{k+1} and Δ_i^{k+1}. We assume that any ties are broken in a way that ensures all processes arrive at an identical choice for the new best point.

Because it is convenient for what follows, we replace the notion of iterations with the notion of occurrences at certain time steps as measured by a *global clock* like that used in other asynchronous convergence proofs; cf. [2]. Let the infinite set

$$\mathcal{T} = \{0, 1, 2, \dots\}$$

be the index of time steps. We assume that the time steps are of fine enough resolution that at most one *event* (i.e., a change in the best known point and/or the value of the step-length control parameter) occurs per time step, per process. In the synchronous case, iterations can be thought of as coarse time steps.

Using our global clock, we can represent the consequence of a single iteration, say k, for a single process, say $i \in \mathcal{P}$, on a timeline as illustrated in Figure 2.1. At time step t_0, process i starts a function evaluation at its trial point given by

$$x_i^{t_0} + \Delta_i^{t_0} d_i.$$

Observe that the notation introduced in (2.2) has changed. Now the time step replaces the iteration number in the superscript and, from now on, we use time steps as our indices. At time step t_1, process i finishes its evaluation of $f(x_i^{t_0} + \Delta_i^{t_0} d_i)$ and broadcasts its result to the remaining processes. We assume that at some time step t_2, all processes in $\mathcal{P}$ have received the results from all other processes, so each independently decides on the point that is now best. Since each process knows the

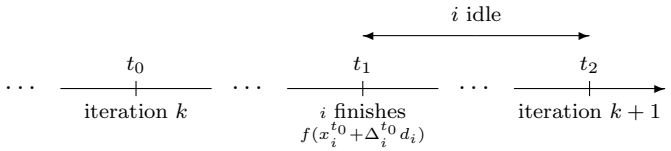


FIG. 2.1. *Timeline for synchronous pattern search for process i.*

results from all p processes in $\mathcal{P}$, and since ties are broken in a consistent fashion, all p processes will arrive at the same conclusion as to which point is now best. Each process then updates its copies of the best point and the step-length control parameter to obtain $x_i^{t_2}$ and $\Delta_i^{t_2}$. Iteration $k+1$ then begins. Note that from time step t_1 until time step t_2, process i is idle.

For process $j \in \mathcal{P}$, $j \neq i$, the procedure differs in only two respects. First, the trial point is calculated using a different search direction $d_j \in \mathcal{D}$ to yield

$$x_j^{t_0} + \Delta_j^{t_0} d_j.$$

Recall that $x_j^{t_0} = x_i^{t_0}$ and $\Delta_j^{t_0} = \Delta_i^{t_0}$ due to the synchronization. Second, we have no guarantee that the evaluation of f at the trial point will take the same number of time steps on process j as it did on process i. At one extreme is the possibility that the evaluation of f takes only a single time step, which would give us the scenario illustrated in Figure 2.2, where $\hat{t}_1$ denotes the time step at which the function evaluation on process j finishes. In this case, $\hat{t}_1 = t_0 + 1$ and process j is idle from time step $t_0 + 1$ to time step t_2. At the other extreme, we have the scenario in Figure 2.3, so that there is effectively no idle time on process j. Note that in this case we have assumed that the communication is instantaneous—our theory allows for this possibility as well as the possibility that communication may take up to a finite number of time steps.

We stress that even though the time required to finish a function evaluation may vary from process to process and from iteration to iteration, the synchronization ensures that, across all processes, iteration k begins at time step t_0 while iteration $k+1$ begins at time step t_2.

The goal of asynchronous parallel pattern search is to eliminate the synchronization since it potentially can waste CPU cycles, as our examples in Figures 2.1 and 2.2 demonstrate and our experimental evidence in [5] confirms. As we see in the next section, APPS allows each process to update its x_i^t and Δ_i^t independently whenever a function evaluation finishes and/or a new message arrives.

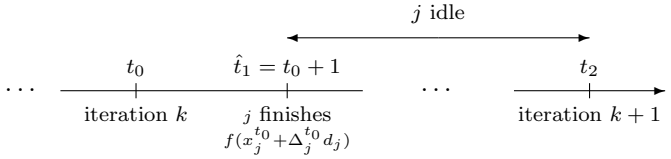


FIG. 2.2. *Timeline for synchronous pattern search for process j.*

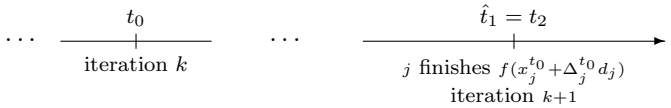


FIG. 2.3. *Alternate timeline for synchronous pattern search for process j.*

3. Asynchronous parallel pattern search. Like PPS, APPS [5] uses p processes collectively to solve (1.1). Each process is in charge of searching along a single search direction from its best known point, and the best known point and the value of the step-length control parameter are varied according to internal and external events. The difference is that individual processes in APPS no longer wait for information from the other processes before making a local decision as to the next best point. Once the decision is made, the process then updates its record of the best point and the step-length control parameter, constructs a new trial point, and immediately begins a new evaluation of the objective function.

Because we no longer have synchronization after every function evaluation, decisions now depend on the time step at which they are made. Therefore, we index according to the global clock described previously. We then define the following for each process $i \in \mathcal{P}$ and time step $t \in \mathcal{T}$:

$$x_i^t = \text{the best known point at time step } t \text{ for process } i, \text{ and}$$
$$\Delta_i^t = \text{the step-length control parameter at time step } t \text{ for process } i.$$

In APPS, the current values of the best point and the step-length control parameter can be different across processes at the same time step $t \in \mathcal{T}$. Therefore, the subscript i is no longer redundant, and it is possible that $x_i^t \neq x_j^t$ and/or $\Delta_i^t \neq \Delta_j^t$. On a single process $i \in \mathcal{P}$, we are guaranteed that at any time step $t \in \mathcal{T}$, $f(x_i^{t+1}) \leq f(x_i^t)$.

The values of x_i^t and Δ_i^t are not necessarily changed at every time step. Let

$$\mathcal{T}_i = \text{the set of time steps at which } x_i^t \text{ and/or } \Delta_i^t \text{ is changed,} \tag{3.1}$$

so that $\mathcal{T}_i \subseteq \mathcal{T}$. For each process $i \in \mathcal{P}$ we categorize each time step $t \in \mathcal{T}$ as either *successful* or *unsuccessful*. We also need to observe further distinctions within each of these two categories, which we detail in sections 3.2 and 3.3.

3.1. Assumptions. As a practical matter, we assume that at the start of the search the best point and the value of the step-length control parameter are equal for all $i \in \mathcal{P}$; that is, there exist $x^0 \in \mathbb{R}^n$ and $\Delta^0 \in \mathbb{R}$, $\Delta^0 > 0$ such that

$$x^0 = x_1^0 = x_2^0 = \cdots = x_p^0 \qquad \text{and} \qquad \Delta^0 = \Delta_1^0 = \Delta_2^0 = \cdots = \Delta_p^0. \tag{3.2}$$

We further assume that the value $f(x^0)$ is known by all processes.

As is standard for pattern search analysis, we assume

$$\mathcal{L}(x^0) = \{\, x \in \mathbb{R}^n \,:\, f(x) \leq f(x^0) \,\} \text{ is bounded.} \tag{3.3}$$

Also, to ensure that APPS always converges to a stationary point of (1.1), we assume that f is continuously differentiable on the closure of $\mathcal{L}(x^0)$, though we need this assumption only for our final result, Theorem 8.5.

We assume that $\mathcal{D}$, the set of search directions, is fixed and finite and of the form given in (2.1), with the conditions that B is a real nonsingular matrix, that $c_i \in \mathbb{Q}^n$ for each $i \in \mathcal{P}$, and that the vectors in the set $\mathcal{D}$ form a positive spanning set for $\mathbb{R}^n$.

We assume that the initial step-length control parameter is constrained by

$$0 < \Delta^{\min} \leq \Delta^0 \leq \Delta^{\max} < +\infty, \tag{3.4}$$

where $\Delta^{\min}$ and $\Delta^{\max}$ are constants. These same constants are used to bound Δ_i^t after any step that produces decrease on f (i.e., after any successful time step). This condition is given in (3.10) and is described fully in section 3.2.1.

We assume that both the maximum time for a function evaluation and the maximum time for a single communication are finite; we quantify those as

(3.5) $\eta =$ maximum number of time steps for evaluating f at a given x, and

(3.6) $\gamma =$ maximum number of time steps for communicating a message.

We assume that the minimum time for evaluation and communication are one and zero time steps, respectively.

3.2. Successful time steps. On process i, we characterize any time step $t \in \mathcal{T}$ at which we identify a point with a strictly lower value of f as *successful*. We further distinguish between *internal* and *external* successes depending on whether the information that identified improvement in the value of f was computed locally or received in the form of a message from another process; we detail these distinctions in sections 3.2.1 and 3.2.2.

We pay special attention to points that produce equal values of f since we must break ties in a consistent fashion. This becomes particularly critical in the asynchronous case since equivalent function values are likely to become known to each process at different time steps and perhaps in reverse order. To ensure the convergence of the overall search, we must ensure that when faced with equivalent function values, every one of the participating processes arrives at the same decision as to which of the points known to produce the same function value should be considered "best." Thus, we may have reason to classify some time steps as successful, even when they do not strictly improve the value of f. We describe such situations in more detail in section 3.2.2.

3.2.1. Internal successes. The first type of successful time step is an *internal success*, which can occur when a process finishes a function evaluation. Suppose that on process $i \in \mathcal{P}$ a function evaluation starts at some time step, say t_0, (using $x_i^{t_0}$ and $\Delta_i^{t_0}$ to generate the trial point) and finishes at some later time step, say t_1. We can represent this on a timeline as in Figure 3.1.

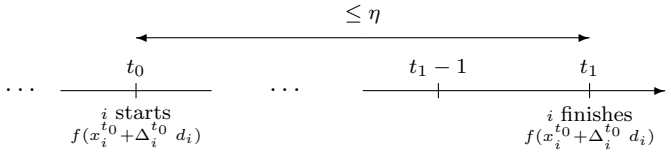


FIG. 3.1. *Timeline for asynchronous pattern search on process i.*

The time step t_1 is considered an internal success when the following condition is satisfied:

$$f\left(x_i^{t_0} + \Delta_i^{t_0}\, d_i\right) < f\left(x_i^{t_1-1}\right). \tag{3.7}$$

We compare $f(x_i^{t_0} + \Delta_i^{t_0}\, d_i)$ to $f(x_i^{t_1-1})$, rather than to $f(x_i^{t_0})$, since it is possible that $x_i^{t_1-1} \neq x_i^{t_0}$ due to an external success, which is described in the next section. When (3.7) is not satisfied, the time step is *unsuccessful*, as described in section 3.3. Otherwise, when (3.7) is satisfied, we say that time step $t_1 \in \mathcal{I}_i$, where

$$\mathcal{I}_i = \text{the set of internal successful time steps for process } i.$$

We then update x_i as follows:

$$x_i^{t_1} = x_i^{t_0} + \Delta_i^{t_0}\, d_i;$$

in other words, $x_i^{t_1}$ is set to the point that produced the best known function value. Further, we update the step-length control parameter Δ_i as follows:

$$\Delta_i^{t_1} = \lambda_i^{t_1} \Delta_i^{t_0},$$

where $\lambda_i^{t_1}$ is the *expansion parameter* for the update at time step t_1. Before we define the expansion parameter for the update, we first define the rational constant

$$\Lambda \in \mathbb{Q}, \quad \Lambda > 1, \tag{3.8}$$

which controls the scaling of all steps. Returning to the choice of λ_i^t, we require it to satisfy two conditions. The first condition is that λ_i^t be a nonnegative integer power of Λ; i.e.,

$$\lambda_i^t = \mathrm{pow}(\Lambda, k_i^t) \tag{3.9}$$

for some

$$k_i^t \in \{0, 1, 2, \dots\}.$$

Since $\Lambda > 1$ and k_i^t is nonnegative, $\lambda_i^t \geq 1$. The second condition on the choice of λ_i^t is that the new step-length control parameter must satisfy

$$0 < \Delta^{\min} \leq \Delta_i^t \leq \Delta^{\max} < +\infty, \tag{3.10}$$

where $\Delta^{\min}$ and $\Delta^{\max}$ are the same constants used in (3.4). Note that (3.10) applies *only* to updates associated with successful time steps. The bounds on Δ_i^t implicitly restrict the value of k_i^t that may be chosen in (3.9).

The lower bound on Δ is new to the asynchronous analysis; in section 4 we give an example that shows why this lower bound is necessary to ensure an accumulation point that is common to all processes. As for the upper bound on Δ, we could use the assumption that $\mathcal{L}(x^0)$ is bounded, given in (3.3), to yield an implicit upper bound on Δ, as is done in the analyses in [7, 10]. For convenience, here we assume the existence of an explicit upper bound and thus eliminate the dependence on f.

Once x_i and Δ_i are updated, process i broadcasts the new best point, its function value, and the new step-length control parameter to all the other processes in $\mathcal{P}$ for them to consider as a candidate for new best. Process i then proceeds with the construction and evaluation of $x_i^{t_1} + \Delta_i^{t_1} d_i$.

3.2.2. External successes. The other type of successful time step is an *external success*. Suppose that an internal success occurs on process i at time step t_1, as just described in section 3.2.1. Then at some time step $t_2 \geq t_1$, process j, $j \neq i$, receives the broadcast from process i with the new best point found by process i, along with its associated function value and step-length control parameter. We assume that process j can immediately assimilate the newly received information *even if it is currently in the midst of a function evaluation*. In the implementation described in [5], we achieve this by executing the function evaluation as a separate thread or process. We represent this example of an external success on the timeline in Figure 3.2.

There are three possibilities when process j receives a message from process i: the function value associated with the incoming point is either better than, equal to, or worse than the function value of the best point at the previous time step. Certainly, if $f(x_i^{t_1}) < f(x_j^{t_2-1})$ holds, then process j now has a new best point,

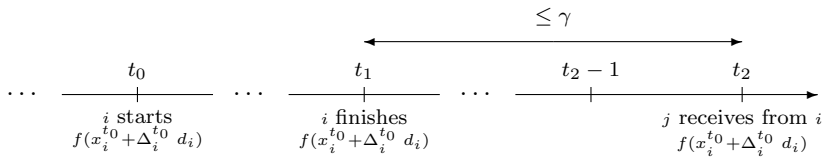


FIG. 3.2. *Timeline for asynchronous pattern search message from process i to process j.*

received from the external process i, and it should update its local values for the best point and the step-length control parameter in light of this new information. However, if $f(x_i^{t_1}) > f(x_j^{t_2-1})$, process j should simply discard the new information since $x_j^{t_2-1}$ is clearly better than $x_i^{t_1}$.

The interesting question is what to do when $f(x_i^{t_1}) = f(x_j^{t_2-1})$. To ensure the robustness of the search procedure, we define a comparison operator $\prec$. Given any $x, y, z \in \mathbb{R}^n$, $\prec$ denotes a comparison that satisfies the following two conditions:

1. $x \prec y$ and $y \prec z$ implies $x \prec z$, and
2. $x = y$ (i.e., neither $x \prec y$ nor $y \prec x$) only if $x[i] = y[i]$ for $i = 1, \ldots, n$, where the notation $x[\cdot]$ denotes the ith entry of the vector x.

We can use any definition for the comparison operator $\prec$ so long as it satisfies these two conditions. For example, we may use the following ordered elementwise comparison. We say $x \prec y$ if there exists $j \in \{1, \ldots, n\}$ such that $x[j] < y[j]$ and $x[i] = y[i]$ for $i = 1, \ldots, j-1$. Given a way to resolve ties, we are now ready to define an external success.

The time step t_2 is considered an external success if either

$$f\left(x_i^{t_1}\right) < f\left(x_j^{t_2-1}\right) \quad \text{or} \quad f\left(x_i^{t_1}\right) = f\left(x_j^{t_2-1}\right) \quad \text{and} \quad x_i^{t_1} \prec x_j^{t_2-1}. \tag{3.11}$$

If (3.11) is satisfied, we then say that $t_2 \in \mathcal{E}_j$, where

$$\mathcal{E}_j = \text{the set of external successful time steps for process } j.$$

The updates are

$$x_j^{t_2} = x_i^{t_1}$$

and

$$\Delta_j^{t_2} = \Delta_i^{t_1}.$$

We assume that the receipt of an external message does not affect the status of a function evaluation that may be executing on the receiving process.

3.2.3. Additional comments on what constitutes a success. Now that we have defined what constitutes both an internal and an external success, we define

$$\mathcal{S}_i = \mathcal{I}_i \cup \mathcal{E}_i = \text{the set of successful time steps for process } i.$$

We emphasize again that although internal successes require strict decrease in the function value as seen in (3.7), external successes relax the requirement of strict decrease and instead use the comparison operator $\prec$ to break ties, as shown in (3.11). This ensures that all processes agree on the best point even when different points generated by different processes have the same function value.

3.3. Unsuccessful time steps. Any time step that is not successful is classified as *unsuccessful*. We let the set

$$\mathcal{U}_i = \mathcal{T} \setminus \mathcal{S}_i$$

denote the unsuccessful time steps on process $i \in \mathcal{P}$. There are two types of unsuccessful time steps.

3.3.1. Contractions. Consider again the function evaluation on process i that starts at time step t_0 and finishes at time step t_1, as shown in Figure 3.1. We say that time step t_1 is a *contraction* if (3.7) is not satisfied *and* $x_i^{t_1-1} = x_i^{t_0}$; i.e., there is no reduction in the function value and x_i has not been updated since time step t_0 (which also means that $\Delta_i^{t_1-1} = \Delta_i^{t_0}$). In terms of time steps, $t_1 \notin \mathcal{I}_i$ and $t \notin \mathcal{E}_i$ for any $t \in \{t_0 + 1, \ldots, t_1 - 1\}$.

In this case, process i is required to reduce the value of its step-length control parameter $\Delta_i^{t_1-1}$ before continuing the search along its direction d_i. This means that $t \in \mathcal{T}_i$ since $\Delta_i^{t_1-1}$, though *not* $x_i^{t_1-1}$, is changed. More specifically, we say that $t_1 \in \mathcal{C}_i$, where

$$\mathcal{C}_i = \text{the set of contraction time steps for process } i.$$

Note that $\mathcal{S}_i \cap \mathcal{C}_i = \emptyset$ since $\mathcal{C}_i \subseteq \mathcal{U}_i$.

We update the step-length control parameter Δ_i as follows:

$$\Delta_i^{t_1} = \theta_i^{t_1} \Delta_i^{t_1-1},$$

where $\theta_i^{t_1}$ is the *contraction parameter* at time step t_1. The choice of the contraction parameter θ_i^t is subject to the following condition, using the same Λ as in (3.9):

$$\theta_i^t = \text{pow}(\Lambda, \ell_i^t) \tag{3.12}$$

for some

$$\ell_i^t \in \{-1, -2, -3, \ldots, \ell^{\min}\}, \tag{3.13}$$

where $\ell^{\min}$ is a finite integer constant. Together, (3.8), (3.12), and (3.13) imply that

$$\theta_i^t \in [\theta^{\min}, \theta^{\max}] \subset (0, 1), \quad \text{where } \theta^{\min} = \text{pow}(\Lambda, \ell^{\min}),\ \theta^{\max} = \text{pow}(\Lambda, -1). \tag{3.14}$$

3.3.2. The trivial case. The final possibility is that no changes to either x_i^t or Δ_i^t occur on process i for a given time step t; in other words, $t \notin \mathcal{T}_i$. This situation could occur for several reasons.

One possibility would be that process i is still evaluating f at a trial point constructed at some time step $t_0 < t$ and that evaluation does not finish during time step t. Thus, $t \notin \mathcal{I}_i$ and $t \notin \mathcal{C}_i$.

A further possibility is that no external candidate arrives from process j, $j \neq i$, or an external candidate does arrive, but it is immediately discarded since its function value does not improve upon $f(x_i^{t-1})$. Thus, $t \notin \mathcal{E}_i$.

A last possibility is that at time step t, process i does finish evaluating f at a trial point constructed at some time step $t_0 < t$ but the function value does not improve upon $f(x_i^{t-1})$, so $t \notin \mathcal{I}_i$. However, before assigning t to $\mathcal{C}_i$, we must verify that $x_i^{t-1} = x_i^t$. If $x_i^{t-1} \neq x_i^t$, that means that at least one external success occurred on process i at some time step $\hat{t} \in \{t_0+1, \ldots, t-1\}$. Let $\hat{t} = \max\{\{t_0+1, \ldots, t-1\} \cap \mathcal{E}_i\}$. In this case, since we have already recorded the external success at time step $\hat{t}$, we construct a new trial point without further changes to $x_i^{\hat{t}}$ or $\Delta_i^{\hat{t}}$ and initiate a new function evaluation. Thus, while $\hat{t} \in \mathcal{E}_i$, $t \in \mathcal{U}_i \setminus \mathcal{C}_i$.

3.4. Multiple decisions in one time step. We allow for the possibility that multiple candidates for the best point may be considered simultaneously at time step $t \in \mathcal{T}$ if, for instance, multiple messages have arrived from external processes or there is both an internal candidate as well as one or more external candidates to consider.

3.5. Identifying the source of a change. If a function evaluation finishes at time step t_1, a new one is started at time step t_1 using the values $x_i^{t_1}$ and $\Delta_i^{t_1}$—at least one of these values is guaranteed to have changed since time step t_0 from either an internal success, an external success, or a contraction.

To identify where a change to x_i^t, and possibly Δ_i^t, was generated (i.e., on which process) and at what time step the corresponding function evaluation started and finished, for each $i \in \mathcal{P}$ and for all $t \in \mathcal{S}_i$ we define the following *generating functions*:

$\omega_i(t) =$ the index of the process generating the update at time step t on process i,

$\tau_i(t) =$ the time index for the initiation of the function evaluation that produced the update at time step t on process i, and

$\nu_i(t) =$ the time index for the completion of the function evaluation that produced the update at time step t on process i.

Here

$$\omega_i(\cdot) : \mathcal{S}_i \to \mathcal{P}, \quad \tau_i(\cdot) : \mathcal{S}_i \to \mathcal{T}, \quad \nu_i(\cdot) : \mathcal{S}_i \to \mathcal{T}, \quad \text{and} \quad 0 \le \tau_i(t) < \nu_i(t) \le t.$$

For our example of an internal success on process i, so that $t_1 \in \mathcal{I}_i$, as illustrated in Figure 3.1, we have $\omega_i(t_1) = i$, $\tau_i(t_1) = t_0$, and $\nu_i(t_1) = t_1$. In fact, $\omega_i(t) = i$ and $\nu_i(t) = t$ for all $t \in \mathcal{I}_i$.

For our example of an external success on process j, so that $t_2 \in \mathcal{E}_j$, as illustrated in Figure 3.2, we have $\omega_j(t_2) = i$, $\tau_j(t_2) = \tau_i(t_1) = t_0$, and $\nu_j(t_2) = \nu_i(t_1) = t_1$.

The generating functions play an important role in the proofs of Lemma 5.1, Theorem 5.2, Lemma 7.4, and Corollary 7.5.

3.6. The definitions for x_i^t and Δ_i^t. For every $t \in \mathcal{T}$, $t > 0$, the best point x_i^t for process $i \in \mathcal{P}$ is defined to be

$$x_i^t = \begin{cases} x_{\omega_i(t)}^{\tau_i(t)} + \Delta_{\omega_i(t)}^{\tau_i(t)}\, d_{\omega_i(t)} & \text{if } t \in \mathcal{S}_i, \text{ and} \\ x_i^{t-1} & \text{otherwise.} \end{cases} \tag{3.15}$$

Recall that we initialize the procedure with x^0 as shown in (3.2). Thus, x_i^t is changed on process $i \in \mathcal{P}$ only at successful time steps $t \in \mathcal{S}_i$.

Changes in Δ_i^t must occur at contraction time steps and may occur at successful (internal or external) time steps. Correspondingly, for every $t \in \mathcal{T}$, $t > 0$, the step-length control parameter Δ_i^t for process $i \in \mathcal{P}$ is defined to be

$$\Delta_i^t = \begin{cases} \lambda_{\omega_i(t)}^{\nu_i(t)}\, \Delta_{\omega_i(t)}^{\tau_i(t)} & \text{if } t \in \mathcal{S}_i, \\ \theta_i^t \Delta_i^{t-1} & \text{if } t \in \mathcal{C}_i, \text{ and} \\ \Delta_i^{t-1} & \text{otherwise.} \end{cases} \tag{3.16}$$

Again, the initialization is as in (3.2), and we assume Δ^0 satisfies (3.4). Recall $\lambda_i^t \ge 1$ is the expansion parameter defined in (3.9) and $\theta_i^t \in (0, 1)$ is the contraction parameter defined in (3.12).

These precise definitions for x_i^t and Δ_i^t play a role in all the results that follow.

4. An overview of the analysis. Now that we have reviewed APPS and introduced most of the notation required for our analysis, we provide an outline of that analysis. Before proceeding, the reader may wish to review the example given in [6]. This example helps establish the definitions given in section 3, including those for the many sets we have introduced to track the progress of the search. Also, [6] illustrates and discusses those features of the asynchronous algorithm that make the analysis more intricate than for the synchronous case.

Our first task is to show that every iterate x_i^t lies on a rational lattice; this is equivalent to Theorem 3.2 in [10] for the synchronous case. The main difference here is that the asynchronism we have introduced in APPS complicates the arguments. Now, for some subset of the t's in $\mathcal{T}$, the x_i^t residing on process i may be the result of an external success—i.e., a point produced by a search along direction d_j on process j, where $j \neq i$. Thus the changes to x_i^t and Δ_i^t may be made without regard to the history of past successes on process i. Nevertheless, in section 5 we show that the algebraic structure found in the synchronous case is still preserved in the asynchronous case.

The lattice structure is the key to ensuring convergence for pattern search. In the synchronous case, the underlying lattice structure makes it possible to prove that a subsequence of the step-length control parameters goes to zero—even though pattern search does not enforce a sufficient decrease condition. In section 7, we show an equivalent result, but we now have p semi-independent sequences of Δ to consider. This makes the arguments more complex than in the synchronous case. Still, we arrive at Theorem 7.8, which says that

$$\liminf_{t \to +\infty} \Delta_j^t = 0 \quad \text{for all } j \in \mathcal{P}.$$

Our definition of Δ_i^t, given in (3.16), ensures that Δ_i^t is decreased only at contractions (i.e., when $t \in \mathcal{C}_i$). In fact, it is these contractions that are of interest for the remainder of the proof. In the synchronous case, there is an accumulation point $\hat{x}$ of the subsequence of iterates associated with contractions which has the property that $0 \leq f(\hat{x})^T d_i$ for all $i \in \mathcal{P}$. The challenge in the asynchronous case lies in showing that all the processes share a *common* accumulation point with this property. In order to do this, in section 8 we show that all the processes share a common subsequence of contraction iterates. This, in turn, relies on the fact that we require $\Delta^{\min} \leq \Delta$ every time we encounter a successful point, which guarantees that in the limit there will be long sequences of unsuccessful iterates. These long sequences of unsuccessful iterates allow us to construct a common sequence of contraction iterates across all processors. This argument culminates with Theorem 8.5, which states that there exists an $\hat{x}$ and, for each $i \in P$, a subset of the contraction iterates $\hat{\mathcal{C}}_i$ such that

$$\lim_{\substack{t \to +\infty \\ t \in \hat{\mathcal{C}}_i}} \Delta_i^t = 0, \quad \lim_{\substack{t \to +\infty \\ t \in \hat{\mathcal{C}}_i}} x_i^t = \hat{x}, \quad \text{and} \quad \lim_{\substack{t \to +\infty \\ t \in \hat{\mathcal{C}}_i}} \nabla f(x_i^t) = \nabla f(\hat{x}) = 0 \quad \text{for all } i \in \mathcal{P}.$$

A fundamental difference in the assumptions for the asynchronous case is that the step-length control parameter is bounded below for all successful iterates (see (3.10)). This assumption is critical to the asynchronous case because it enables us to avoid a so-called race condition. If we did not enforce this lower bound, we might have different processes producing sequences of iterates that converge to different limit points, as the following situation in $\mathbb{R}^2$ illustrates. Let $f(x) = x^T x$. Observe that f is symmetric about both the x-axis and the y-axis and has a unique global minimizer

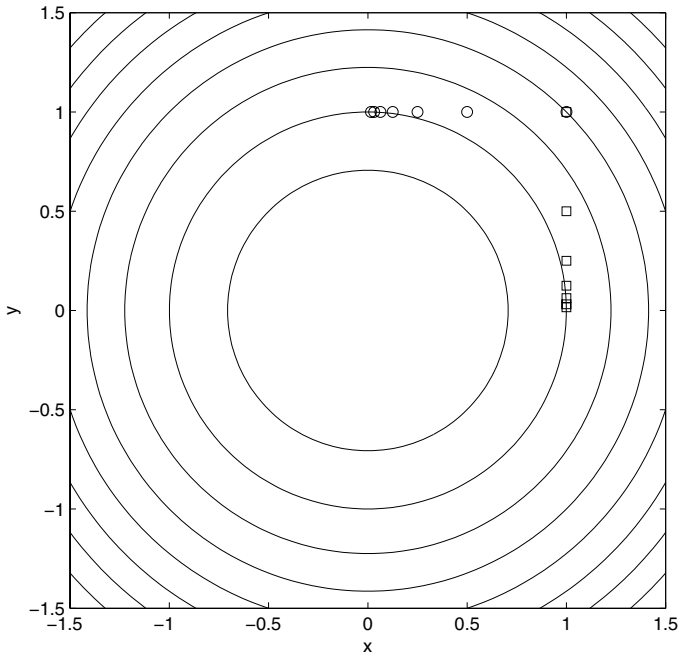


FIG. 4.1. *A potential race condition, which we exclude. In this illustration, if we do not enforce the bounds on* Δ_i^t *after an internal success, then each sequence converges to a different limit; the circles denote the sequence of best iterates on processes* 1 *and* 3 *while the squares denote the sequence of best iterates on process* 2. *The situation is remedied by requiring* $\Delta_i^t \geq \Delta^{\min}$ *for all* $t \in \mathcal{I}_i$.

at $x = (0,0)^T$. Choose

$$\mathcal{D} = \left\{ \begin{pmatrix} -1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ -1 \end{pmatrix}, \begin{pmatrix} 1 \\ 1 \end{pmatrix} \right\},$$

which forms a positive basis for $\mathbb{R}^2$. Let $x^0 = (1,1)^T$, $\Delta^0 = 2$, and $\Lambda = 2$. For every contraction ($t \in \mathcal{C}_i$), choose $\theta = \Lambda^{-2} = 1/4$. For every internal success ($t \in \mathcal{I}_i$), choose $\lambda = \Lambda^1 = 2$, *ignoring* the lower bound restriction in (3.10). Assume further that each function evaluation requires exactly one time step and each communication takes exactly two time steps.

The situation that develops in this case is illustrated in Figure 4.1. Here, both $\{x_1^t\}$ and $\{x_3^t\} \to (0,1)^T$ (denoted by circles) while $\{x_2^t\} \to (1,0)^T$ (denoted by squares). The reason that the three sequences converge to two different limit points, neither of which is the unique stationary point for f, is that the first and the second processes are reducing their steps too quickly, continuing to find internal successes, and rejecting each candidate for an external success because they have already found another, better, internal success. The third process cannot remedy the situation since its direction is always a direction of ascent. The broadcasts of internal successes from the first and the second processes arrive on the third process within the same time step and have the same function value, which is better than any produced on the third process. Ties are broken in favor of the first process, leading to an external success on the third process, so the iterates on the third process also converge to $(0,1)^T$.

Enforcing the lower bound of $\Delta^{\min}$ on Δ_i^t for successful points eliminates this race condition. Choose, for instance, $\Delta^{\min} = 1/4$ and notice that this eventually disrupts the symmetric exchanges between the first and the second processes.

We require a *common* accumulation point $\hat{x}$ so that we can use the fact that our search directions in $\mathcal{D}$ form a positive spanning set, thus ensuring the final conclusion of Theorem 8.5: that $\hat{x}$ is also a stationary point of f.

We close by noting that in practice we stop searching along a given direction d_i once Δ_i^t falls below a certain threshold. Process i then waits until either another process reports a better point, in which case the search along d_i resumes with this new best point, or a sufficient number of other processes have converged to the same point identified by process i, in which case the entire search terminates. (For further details, see [5].) Thus, as a practical matter, $\Delta^{\min}$ need not impede the overall progress of the search, as it can be chosen to be on the order of the threshold used to terminate the search.

5. The algebraic structure of the iterates. We return to the analysis of APPS. We use the formulation for x_i^t given in (3.15) to show that we can, in fact, write any x_i^t as a linear combination of the search directions (translated by x^0). We prove this in Lemma 5.1. Then, in Theorem 5.2, we show that the algebraic structure underlying the sequences $\{x_i^t\}$, for all $i \in \mathcal{P}$, guarantees that all the iterates lie on a rational lattice defined by the search directions. The latter result is equivalent to Theorem 3.2 in [10].

LEMMA 5.1. *For any $i \in \mathcal{P}$ and any $t \in \mathcal{T}$, there exist sets $\hat{\mathcal{I}}_j(i,t) \subseteq \mathcal{I}_j$ for each $j \in \mathcal{P}$ such that*

$$x_i^t = x^0 + \sum_{j \in \mathcal{P}} \delta_j(i,t)\, d_j \quad \textit{with} \quad \delta_j(i,t) = \sum_{\hat{t} \in \hat{\mathcal{I}}_j(i,t)} \Delta_j^{\tau_j(\hat{t})}, \tag{5.1}$$

where $\delta_j(i,t) = 0$ if $\hat{\mathcal{I}}_j(i,t) = \emptyset$.

Proof. We prove this lemma by induction on t. For any $i \in \mathcal{P}$, the case for $t = 0$ is trivial since $x_i^0 = x^0$ by (3.2). Simply choose $\hat{\mathcal{I}}_j(i,0) = \emptyset$ for each $j \in \mathcal{P}$.

Now consider the case for general t for any $i \in \mathcal{P}$. First consider $t \in \mathcal{U}_i$, in which case (3.15) gives us $x_i^t = x_i^{t-1}$. From the induction hypothesis, we have

$$x_i^{t-1} = x^0 + \sum_{j \in \mathcal{P}} \delta_j(i,t-1)\, d_j \quad \text{with} \quad \delta_j(i,t-1) = \sum_{\hat{t} \in \hat{\mathcal{I}}_j(i,t-1)} \Delta_j^{\tau_j(\hat{t})}.$$

In this case, we simply choose $\hat{\mathcal{I}}_j(i,t) = \hat{\mathcal{I}}_j(i,t-1)$ for all $j \in \mathcal{P}$ to yield (5.1).

On the other hand, consider $t \in \mathcal{S}_i$. From (3.15), we have

$$x_i^t = x_{\omega_i(t)}^{\tau_i(t)} + \Delta_{\omega_i(t)}^{\tau_i(t)} d_{\omega_i(t)}.$$

The assumption that the minimum time for a function evaluation is one time step ensures that $\tau_i(t) < t$ for all $i \in \mathcal{P}$. Thus, from the induction hypothesis, we can rewrite the first term as

$$x_{\omega_i(t)}^{\tau_i(t)} = x^0 + \sum_{j \in \mathcal{P}} \delta_j(\omega_i(t), \tau_i(t))\, d_j \quad \text{with } \delta_j(\omega_i(t), \tau_i(t)) = \sum_{\hat{t} \in \hat{\mathcal{I}}_j(\omega_i(t), \tau_i(t))} \Delta_j^{\tau_j(\hat{t})}.$$

By definition, we also have $\tau_i(t) = \tau_{\omega_i(t)}(\nu_i(t))$ and $\nu_i(t) \in \mathcal{I}_{\omega_i(t)}$. Therefore, choosing

$$\hat{\mathcal{I}}_j(i,t) = \begin{cases} \hat{\mathcal{I}}_j(\omega_i(t), \tau_i(t)) \cup \{\nu_i(t)\} & \text{for } j = \omega_i(t), \text{ and} \\ \hat{\mathcal{I}}_j(\omega_i(t), \tau_i(t)) & \text{for } j \neq \omega_i(t) \end{cases}$$

yields (5.1). ∎

The purpose of the sets $\hat{\mathcal{I}}_j(i,t)$ is to track, for each $j \in \mathcal{P}$, which subset of the set of time steps that produced internal successes on process j led to the x_i^t residing on process i at time step t.

Now that we have taken a closer look at x_i^t, let us do the same for Δ_i^t. From (3.16), (3.12), and (3.9), we see that for any $i \in \mathcal{P}$ and for any $t \in \mathcal{T}$ we can express any Δ_i^t as a multiple of an integer power of the Λ from (3.8) times the Δ^0 from (3.4). Let Γ_i^t denote that integer power so that

$$\Delta_i^t = \text{pow}(\Lambda, \Gamma_i^t)\, \Delta^0, \quad \Gamma_i^t \in \mathbb{Z}. \tag{5.2}$$

Since $\Lambda \in \mathbb{Q}$, we can find Λ_{N} and Λ_{D} (here the subscripts denote *numerator* and *denominator*, respectively) such that

$$\Lambda = \frac{\Lambda_{\text{N}}}{\Lambda_{\text{D}}}, \quad \text{where } \Lambda_{\text{D}}, \Lambda_{\text{N}} \in \mathbb{N} \text{ with } \Lambda_{\text{D}}, \Lambda_{\text{N}} \text{ relatively prime.} \tag{5.3}$$

Using (5.3), we can rewrite (5.2) as

$$\Delta_i^t = \text{pow}(\Lambda_{\text{N}}, \Gamma_i^t)\ \text{pow}(\Lambda_{\text{D}}, -\Gamma_i^t)\, \Delta^0, \quad \Gamma_i^t \in \mathbb{Z}. \tag{5.4}$$

Define $\Gamma^{\max}$ to be the least integer such that

$$\text{pow}(\Lambda, \Gamma^{\max})\Delta^0 \geq \Delta^{\max}, \quad \Gamma^{\max} \in \mathbb{Z}. \tag{5.5}$$

From (3.10) we are then guaranteed that

$$\Gamma_i^t \leq \Gamma^{\max} \quad \text{for all } i \in \mathcal{P} \text{ and } t \in \mathcal{T}. \tag{5.6}$$

Finally, we recall the definition of the set of search directions $\mathcal{D}$ given in (2.1). Observe that each search direction $d_i \in \mathcal{D}$ is the product of a real nonsingular matrix B and a rational vector c_i.

Combining our observations on x_i^t and Δ_i^t, with our recollection of the definition of $\mathcal{D}$, we now state and prove the following theorem, which is our analogue of Theorem 3.2 from [10].

THEOREM 5.2. *Let $i \in \mathcal{P}$ and $\Gamma \in \mathbb{Z}$. For any $t \in \mathcal{T}$ such that*

$$\Gamma \leq \min\, \{\, \Gamma_i^{\tau_i(\hat{t})} \,:\, \hat{t} \leq t,\ \hat{t} \in \mathcal{I}_i,\ i \in \mathcal{P} \,\}, \tag{5.7}$$

where Γ_i^t is defined as in (5.2)*, there exists $\zeta_j(i,t,\Gamma) \in \mathbb{Z}$ for each $j \in \mathcal{P}$ such that*

$$x_i^t = x^0 + \frac{\text{pow}(\Lambda_{\text{N}}, \Gamma)}{\text{pow}(\Lambda_{\text{D}}, \Gamma^{\max})}\, \Delta^0\, B \sum_{j \in \mathcal{P}} \zeta_j(i,t,\Gamma)\, c_j, \tag{5.8}$$

where Λ_{N} and Λ_{D} are as defined in (5.3) *and $\Gamma^{\max}$ is as defined in* (5.5)*.*

Further, x_i^t lies on the rational lattice defined by integer multiples of the elements of the set $\{c_1, \ldots, c_p\}$ that is scaled by $\text{pow}(\Lambda_{\text{N}}, \Gamma)\ \text{pow}(\Lambda_{\text{D}}, -\Gamma^{\max})\, \Delta^0$*, translated by x^0, and subject to a (possible) change of basis B. This lattice is denoted by $\mathcal{G}(\mathcal{D}, \Lambda, \Gamma, \Delta^{\max}, \Delta^0, x^0)$.*

Proof. First we make an observation about any $\Delta_i^{\tau_i(\hat{t})}$ such that $i \in \mathcal{P}$, $\hat{t} \leq t$, and $\hat{t} \in \mathcal{I}_i$. From (5.4)–(5.6) we have

$$\Delta_i^{\tau_i(\hat{t})} = \text{pow}(\Lambda_\text{N}, \Gamma_i^{\tau_i(\hat{t})} - \Gamma)\ \text{pow}(\Lambda_\text{D}, \Gamma^\text{max} - \Gamma_i^{\tau_i(\hat{t})})\ \frac{\text{pow}(\Lambda_\text{N}, \Gamma)}{\text{pow}(\Lambda_\text{D}, \Gamma^\text{max})}\ \Delta^0.$$

Recall from (5.3) that Λ_D, $\Lambda_\text{N} \in \mathbb{N} \subset \mathbb{Z}$ and from (5.2) that $\Gamma_i^{\tau_i(\hat{t})} \in \mathbb{Z}$. Further, we have assumed that $\Gamma \in \mathbb{Z}$ and $\Gamma \leq \Gamma_i^{\tau_i(\hat{t})}$, and the assumptions placed on Γ^max ensure that $\Gamma^\text{max} \in \mathbb{Z}$ and $\Gamma^\text{max} \geq \Gamma_i^{\tau_i(\hat{t})}$. Combining these observations, we have that

$$\text{pow}(\Lambda_\text{N}, \Gamma_i^{\tau_i(\hat{t})} - \Gamma)\,\text{pow}(\Lambda_\text{D}, \Gamma^\text{max} - \Gamma_i^{\tau_i(\hat{t})}) \in \mathbb{Z}.$$

In Lemma 5.1 we saw that we could write any x_i^t as the sum of x^0 plus a linear combination of the search directions. Using the definitions of $\hat{\mathcal{I}}_j(i,t)$ and $\delta_j(i,t)$ from Lemma 5.1, we choose

$$\begin{aligned}\zeta_j(i,t,\Gamma) &= \sum_{\hat{t} \in \hat{\mathcal{I}}_j(i,t)} \text{pow}(\Lambda_\text{N}, \Gamma_j^{\tau_j(\hat{t})} - \Gamma)\,\text{pow}(\Lambda_\text{D}, \Gamma^\text{max} - \Gamma_j^{\tau_j(\hat{t})}) \\ &= \frac{\text{pow}(\Lambda_\text{D}, \Gamma^\text{max})\,\delta_j(i,t)}{\text{pow}(\Lambda_\text{N}, \Gamma)\,\Delta^0},\end{aligned}$$

with $\zeta_j(i,t,\Gamma) = 0$ if $\hat{\mathcal{I}}_j(i,t) = \emptyset$. Clearly, $\zeta_j(i,t,\Gamma) \in \mathbb{Z}$. Given that for every $j \in \mathcal{P}$, $d_j = Bc_j$, (5.8) then follows immediately from (5.1). The final statement follows from two facts. First, the set $\{c_1, \ldots, c_p\}$ is finite. Second, each of the c_j's is strictly rational and any finite set of rational numbers can be scaled to the integers. ∎

The importance of Theorem 5.2 will become apparent in Lemma 7.4, where we show that some subsequence of the step-length control parameters must go to zero.

6. The subset of time steps at which changes occur is infinite. Before we proceed to the proof of global convergence, we revisit the set $\mathcal{T}_i$, which we first defined in (3.1), and show that it must be infinite. A review of (3.15) and (3.16) leads to an alternate definition in terms of the subsets $\mathcal{S}_i$ and $\mathcal{C}_i$:

$$\mathcal{T}_i = \mathcal{S}_i \cup \mathcal{C}_i. \tag{6.1}$$

LEMMA 6.1. *$\mathcal{T}_i$ is infinite.*

Proof. Each function evaluation takes at most η time steps, and a new function evaluation is started at the conclusion of each function evaluation. Since $\mathcal{T}$ is infinite, there are infinitely many function evaluations. Recalling the discussion in section 3.5, for each function evaluation we are guaranteed that either an external successful update took place during the function evaluation or either an internal successful update or a contraction took place at the conclusion of the function evaluation. So, there must be at least one update to x_i and/or Δ_i for every function evaluation and, hence, there are infinitely many updates. ∎

This fact about $\mathcal{T}_i$ plays a role in the analysis ahead.

7. A subsequence of the step-length control parameters goes to zero. Once the lattice structure has been established, the next part of the proof of convergence for standard pattern search convergence analysis [10] involves showing that the step-length control parameter Δ goes to zero; i.e.,

$$\liminf_{t \to +\infty} \Delta^t = 0.$$

In this section, we aim to show an equivalent result, but we now have p semi-independent sequences of Δ to consider. Given this complication, the basic outline for our arguments is as follows:

1. If the number of successful time steps for some process is finite, showing that the sequence of step-length control parameters goes to zero is trivial. So, we eliminate this case first in Lemma 7.1.

2. Using Lemma 7.1, we then show, in Lemma 7.2 and Corollary 7.3, that either every process has a set of successful time steps that is finite or none do. From this point forward, we then need only concern ourselves with the case where the number of successful time steps is infinite.

3. Lemma 7.4 is a key result. We show that some subsequence of the set of all step-length control parameters (indexed over all processes and all time steps) must go to zero. This result relies on the fact that every x_i^t lies on a rational lattice.

4. We narrow the scope in Corollary 7.5 to show that a subsequence of step-length control parameters converges to zero on *one* process $i \in \mathcal{P}$.

5. Before we can extend this result to the remaining processes, we introduce some new definitions that help us discover what is happening between successful time steps on any process $j \in \mathcal{P}$, $j \neq i$. In Lemma 7.6, we conclude that the lim sup of the number of time steps between successes on a single process goes to $+\infty$ in these cases.

6. We now can tie together the actions across processes to say, in Lemma 7.7, that *every* process must have a subsequence of step-length control parameters that goes to zero.

7. Combining all these results into Theorem 7.8, we see that whether or not the number of successful time steps is infinite, every process has a subsequence of step-length control parameters that goes to zero.

Now that we have an overall picture of the argument, we begin by showing that for any process i which has only finitely many successful time steps, the sequence of step-length control parameters goes to zero.

LEMMA 7.1. *If $\mathcal{S}_i$ is finite for some $i \in \mathcal{P}$, then*

$$\lim_{t \to +\infty} \Delta_i^t = 0.$$

Proof. Let $t_0 = \max \{ t : t \in \mathcal{S}_i \}$. Then, by (3.16), for any time step $t \in \mathcal{T}$ such that $t > t_0$, either the time step is a contraction or nothing happens. From (6.1), we have $\mathcal{T}_i = \mathcal{S}_i \cup \mathcal{C}_i$, and Lemma 6.1 assures us that $\mathcal{T}_i$ is infinite. Since, by assumption, the set $\mathcal{S}_i$ is finite, we conclude that the set $\mathcal{C}_i$ must be infinite. Hence there are infinitely many contractions after time step t_0. Therefore, the sequence $\{\Delta_i^t\}_{t=t_0}^{+\infty}$ is decreasing and bounded below by zero. Finally, (3.14) guarantees that the contraction parameter $\theta_i^t \leq \theta^{\max} < 1$, which enforces a fraction of decrease at each contraction. We can therefore conclude that the sequence $\{\Delta_i^t\}_{t=t_0}^{+\infty}$ converges to zero. Hence, the claim. ∎

In the next lemma, we show that if one process has infinitely many successful time steps, then *every* process must have infinitely many successful time steps.

LEMMA 7.2. *If $\mathcal{S}_i$ is infinite for some $i \in \mathcal{P}$, then $\mathcal{S}_j$ is infinite for all $j \in \mathcal{P}$.*

Proof. Suppose not; that is, suppose there exists $k \in \mathcal{P}$, $k \neq i$, such that $\mathcal{S}_k$ is finite. Let $t_0 = \max \{ t : t \in \mathcal{S}_k \}$, which implies that $x_k^{t_0}$ is the best point known by process k over all $t \in \mathcal{T}$. The point $x_k^{t_0}$ is considered by process i at some later time step $t_1 \leq t_0 + \gamma$, where γ is defined in (3.6). Since S_i is infinite, $x_k^{t_0}$, whether initially accepted or rejected at time step t_1, is improved upon at some later time step

t_2 with $t_2 > t_1$; together, (3.5) and (3.6) guarantee that t_2 is finite. The point $x_i^{t_2}$ must, in turn, be considered by process k at a later time step $t_3 \leq t_2 + \gamma$. Since $x_i^{t_2}$ is an improvement over $x_k^{t_0}$, we must have $t_3 \in \mathcal{S}_k$; but this contradicts t_0 being the maximum $t \in \mathcal{S}_k$. ∎

The immediate corollary is that if any process has only finitely many successful time steps, then *every* process has only finitely many successful time steps.

COROLLARY 7.3. *If $\mathcal{S}_i$ is finite for some $i \in \mathcal{P}$, then $\mathcal{S}_j$ is finite for all $j \in \mathcal{P}$.*

From Lemma 7.1 and Corollary 7.3, the case for the convergence of the step-length control parameters to zero is trivial when there are finitely many successful time steps. The remainder of this section concentrates on the case where there are infinitely many successful time steps on each process.

The next lemma shows there is a subsequence of step-length control parameters (indexed over all processes) that converges to zero.

LEMMA 7.4. *Suppose $\mathcal{S}_j$ is infinite for all $j \in \mathcal{P}$. Then there exists $i \in \mathcal{P}$ such that*

$$\liminf_{\substack{t \to +\infty \\ t \in \mathcal{S}_i}} \Delta_{\omega_i(t)}^{\tau_i(t)} = 0.$$

Proof. Suppose not. Then there exists $\Delta^* > 0$ such that

$$\Delta_{\omega_j(t)}^{\tau_j(t)} \geq \Delta^* \quad \text{for all } j \in \mathcal{P} \text{ and } t \in \mathcal{S}_j.$$

Choose a $\Gamma^* \in \mathbb{Z}$ that satisfies (5.7) for all $t \in \mathcal{T}$. We are guaranteed that such a Γ^* exists since Δ^* is strictly positive. With this choice of Γ^*, Theorem 5.2 guarantees that (5.8) holds for *every* choice of $t \in \mathcal{T}$, and thus every x_j^t lies on the translated rational lattice $\mathcal{G}(\mathcal{D}, \Lambda, \Gamma^*, \Delta^{\max}, \Delta^0, x^0)$.

Observe that each lattice point in $\mathcal{G}(\mathcal{D}, \Lambda, \Gamma^*, \Delta^{\max}, \Delta^0, x^0)$ can be considered successful at most once by each process. Consider process $k \in \mathcal{P}$. Recall that $\mathcal{S}_k = \mathcal{I}_k \cup \mathcal{E}_k$ and a successful point must satisfy either (3.7) or (3.11). In either case, if $f(x_k^{t_2}) < f(x_k^{t_1})$, then clearly $x_k^{t_1} \neq x_k^{t_2}$. The only other possibility is that $t_2 \in \mathcal{E}_k$ with $f(x_k^{t_2}) = f(x_k^{t_1})$, in which case we must have $x_k^{t_2} \prec x_k^{t_1}$ so that, once again, $x_k^{t_1} \neq x_k^{t_2}$. We conclude, therefore, that for any process $k \in \mathcal{P}$, we cannot have $t_1, t_2 \in \mathcal{S}_k$ with $t_1 < t_2$ such that $x_k^{t_1} = x_k^{t_2}$.

On the other hand, every successful point must lie in $\mathcal{L}(x^0)$. The intersection of the bounded set $\mathcal{L}(x^0)$ with the translated integer lattice $\mathcal{G}(\mathcal{D}, \Lambda, \Gamma^*, \Delta^{\max}, \Delta^0, x^0)$ is finite.

Since any successful point must be in the finite set $\mathcal{G}(\mathcal{D}, \Lambda, \Gamma^*, \Delta^{\max}, \Delta^0, x^0) \cap \mathcal{L}(x^0)$ and no point is successful more than once for each process $j \in \mathcal{P}$, it follows that $\mathcal{S}_j$ must be finite. But this contradicts the assumption that $\mathcal{S}_j$ is infinite for all $j \in \mathcal{P}$. Hence, the claim. ∎

An immediate corollary to the preceding lemma is that there is some process which has a subsequence of step-length control parameters that converges to zero.

COROLLARY 7.5. *Suppose $\mathcal{S}_j$ is infinite for all $j \in \mathcal{P}$. Then there exists $i \in \mathcal{P}$ such that*

$$\liminf_{t \to +\infty} \Delta_i^t = 0. \tag{7.1}$$

Proof. By Lemma 7.4, there exists $i \in \mathcal{P}$ and $\bar{\mathcal{S}}_i \subseteq \mathcal{S}_i$ such that

$$\lim_{\substack{t \to +\infty \\ t \in \bar{\mathcal{S}}_i}} \Delta_{\omega_i(t)}^{\tau_i(t)} = 0.$$

For each $j \in \mathcal{P}$, define $\bar{\mathcal{S}}_{ij} = \{t \in \bar{\mathcal{S}}_i : \omega_i(t) = j\}$, so $\bigcup_{j=1}^{p} \bar{\mathcal{S}}_{ij} = \bar{\mathcal{S}}_i$. Since $\bar{\mathcal{S}}_i$ is infinite, there exists at least one k such that $\bar{\mathcal{S}}_{ik}$ is infinite. So,

$$\lim_{\substack{t \to +\infty \\ t \in \bar{\mathcal{S}}_{ik}}} \Delta_k^{\tau_i(t)} = 0.$$

Hence, the claim. ∎

We need to show that a subsequence of step-length control parameters is going to zero for *every* process. In order to do so, we must first introduce some definitions and an additional lemma.

For each process $i \in \mathcal{P}$, we can decompose the set of unsuccessful time steps (i.e., $t \notin \mathcal{S}_i$) into contiguous blocks as follows:

$$\mathcal{U}_i = \mathcal{T} \setminus \mathcal{S}_i = \mathcal{U}_{i1} \cup \mathcal{U}_{i2} \cup \cdots \cup \mathcal{U}_{iN}, \tag{7.2}$$

where N may be $+\infty$, each $\mathcal{U}_{i\ell}$ is a contiguous index block (e.g., $\mathcal{U}_{i\ell} = \{3, 4, 5, 6\}$), and any pair $\mathcal{U}_{i\ell}$ and $\mathcal{U}_{i,\ell+1}$ is separated by at least one $t \in \mathcal{S}_i$.

It is also useful to define the minimum number of contractions required to reduce $\Delta^{\min}$ to a given $\Delta \in \mathbb{R}$, $\Delta > 0$, as

$$\underline{\kappa}(\Delta) = \min \{ p \in \{0, 1, 2, \dots\} : \operatorname{pow}(\theta^{\min}, p)\, \Delta^{\min} \leq \Delta \}, \tag{7.3}$$

where $\theta^{\min}$ is defined in (3.14) and $\Delta^{\min}$ is defined in (3.10). It is straightforward to see that

$$\lim_{\Delta \to 0} \underline{\kappa}(\Delta) = +\infty. \tag{7.4}$$

Finally, for a given $t \in \mathcal{T}$, we define the last successful time step up to, and possibly including, t and the first successful time step after t as

$$\psi_i(t) = \max \{ \bar{t} \in \mathcal{S}_i \cup \{0\} : \bar{t} \leq t \}, \quad \text{and} \tag{7.5}$$

$$\phi_i(t) = \min \{ \bar{t} \in \mathcal{S}_i : t < \bar{t} \}, \tag{7.6}$$

respectively. We ensure that $\psi_i(t)$ is always defined by setting it to zero in the case that $\{ \bar{t} \in \mathcal{S}_i : \bar{t} \leq t \}$ is empty. In the case that there is no $\bar{t} \in \mathcal{S}_i$ satisfying $t < \bar{t}$, then $\phi_i(t) = +\infty$. Thus, $\psi_i(\cdot) : \mathcal{T} \to \mathcal{S}_i \cup \{0\}$, $\phi_i(\cdot) : \mathcal{T} \to \mathcal{S}_i \cup \{+\infty\}$, and $\psi_i(t) < \phi_i(t)$ for all $t \in \mathcal{T}$.

Using the above definitions, we can show that the lim sup of the number of time steps between successes is going to infinity if a subsequence of the step-length control parameters is going to zero.

LEMMA 7.6. *Suppose $\mathcal{S}_j$ is infinite for all $j \in \mathcal{P}$. Then for all $i \in \mathcal{P}$ satisfying* (7.1)*, we have*

$$\limsup_{\ell \to +\infty} |\mathcal{U}_{i\ell}| = +\infty. \tag{7.7}$$

Proof. Let $i \in \mathcal{P}$ be such that (7.1) holds. By the definition of the limit, for any $\Delta^* > 0$, there exists $t^* \in \mathcal{T}$ such that $\Delta_i^{t^*} < \Delta^*$. Without loss of generality, we assume $t^* \in \mathcal{U}_i$.

Then, using definitions (7.3) and (7.5) from above, there must be at least $\underline{\kappa}(\Delta^*)$ time steps between t^* and $\psi_i(t^*)$ since (3.10) must hold for all $t \in \mathcal{S}_i$. Let ℓ^* be such that $t^* \in \mathcal{U}_{i\ell^*}$. Then

$$|\mathcal{U}_{i\ell^*}| > \underline{\kappa}(\Delta^*).$$

From (7.4), the proof is complete. ∎

We can now show that, in the case of an infinite number of successful time steps, a subsequence of the step-length control parameters converges to zero for every process.

LEMMA 7.7. *Suppose $\mathcal{S}_j$ is infinite for all $j \in \mathcal{P}$. Then for all $j \in \mathcal{P}$,*

$$\liminf_{t \to +\infty} \Delta_j^t = 0.$$

Proof. Suppose not. Then there exists an $i \in \mathcal{P}$ and $\Delta^* > 0$ such that

$$\Delta_i^t \geq \Delta^* \quad \text{for all } t \in \mathcal{T}.$$

Define

$$\bar{\kappa}(\Delta^*) = \min\{p \in \{0, 1, 2, \dots\} : \text{pow}(\theta^{\max}, p)\Delta^{\max} \leq \Delta^*\},$$

where $\theta^{\max}$ is defined in (3.14) and $\Delta^{\max}$ is defined in (3.10). Then $\bar{\kappa}(\Delta^*)$ is the maximum possible number of contractions needed to reduce $\Delta^{\max}$ to Δ^*. So the maximum number of time steps between two successful time steps on process i is

$$\max_{\ell} |\mathcal{U}_{i\ell}| \leq \eta\, \bar{\kappa}(\Delta^*),$$

where η is defined in (3.5) and $\mathcal{U}_{i\ell}$ is defined in (7.2).

Now consider $k \in \mathcal{P}$, $k \neq i$. Since any successful point produced on process k is considered on process i within γ time steps, i has a new minimum within $\eta\, \bar{\kappa}(\Delta^*)$ time steps, and that new minimum is considered by process k within γ more time steps; so the maximum number of time steps between successes on any process k, $k \neq i$, can be at most

$$\max_{\ell} |\mathcal{U}_{k\ell}| \leq \eta\, \bar{\kappa}(\Delta^*) + 2\gamma. \tag{7.8}$$

However, Corollary 7.5 guarantees us that there exists i^* such that (7.1) holds, and our null hypothesis tells us $i^* \neq i$. Further, Lemma 7.6 says (7.7) must hold for i^*, but this contradicts (7.8), which also holds for $k = i^*$. Hence, the claim. ∎

Finally, we show that each process has a subsequence of step-length control parameters that converges to zero—whether there are finitely or infinitely many successful time steps.

THEOREM 7.8. *For every process $j \in \mathcal{P}$, there exists a subsequence of the step-length control parameters that goes to zero; that is,*

$$\liminf_{t \to +\infty} \Delta_j^t = 0 \quad \textit{for all } j \in \mathcal{P}.$$

Proof. If $\mathcal{S}_i$ is infinite for some $i \in \mathcal{P}$, then $\mathcal{S}_j$ is infinite for all $j \in \mathcal{P}$ by Lemma 7.2, in which case the claim follows immediately from Lemma 7.7. Otherwise, all $\mathcal{S}_j$ must be finite for all $j \in \mathcal{P}$ by Corollary 7.3, in which case the claim follows from Lemma 7.1. ∎

The following corollary says that, specifically, the subsequence of time steps at which the step-length control parameters decrease forms a subset of the set of unsuccessful time steps. This corollary is useful in the next section.

COROLLARY 7.9. *The set $\mathcal{C}_j$ is infinite for all $j \in \mathcal{P}$, and*

$$\liminf_{\substack{t \to +\infty \\ t \in \mathcal{C}_j}} \Delta_j^t = 0 \quad \textit{for all } j \in \mathcal{P}. \tag{7.9}$$

Proof. This follows immediately from Theorem 7.8 since for each $j \in \mathcal{P}$, $\Delta_j^t \geq \Delta^{\min}$ for all $t \in \mathcal{S}_j$ and (3.16) confirms that Δ_j^t is unchanged for all $t \in \mathcal{T} \setminus \mathcal{T}_i$. ∎

8. A common accumulation point that is also a stationary point. Our final goal is to show that there exists a common accumulation point for all processes and that this accumulation point has a zero gradient. Our argument is outlined as follows.

1. In Lemma 8.1 we show that on process 1 we can extract a subsequence of contractions for which the step-length control parameter goes to zero and that the subsequence x_1^t associated with these particular contractions has an accumulation point. (We specify the first process for convenience, but we could pick any process.)

2. Still focused on process 1, in Corollary 8.2 we show that the number of unsuccessful time steps before each of these contractions is going to $+\infty$. This means that on process 1 we have a sequence of ever-lengthening contiguous index blocks of unsuccessful time steps.

3. In Lemma 8.3 we show that using the subsequence of contractions on process 1 for which the step-length control parameter goes to zero, each process i, $i \neq 1$, has its own corresponding sequence of contiguous index blocks of unsuccessful time steps.

4. In Lemma 8.4, we show that the sequence of contiguous index blocks of unsuccessful time steps on each process i, $i \neq 1$, is also ever-lengthening. We then extract a sequence of step-length control parameters corresponding to these ever-lengthening blocks of unsuccessful time steps and show that this particular sequence of step-length control parameters must go to zero.

5. Finally, in Theorem 8.5 we show that these blocks of unsuccessful iterates can be used to build a sequence of contraction iterates corresponding to those on process 1—and thus share the same accumulation point. Furthermore, if we use the fact that the set of search directions positively spans $\mathbb{R}^n$, and assume that f is continuously differentiable, then we can show that this common accumulation point is also a stationary point of f.

In essence, our argument for the existence of a common accumulation point is based on the timing of the global clock. Since we have assumed both that the number of time steps required for a function evaluation is finite (3.5) and that the number of time steps required for the communication of a message is finite (3.6), we know that eventually every process must see any candidate for the new best point in a finite amount of time. What we do not know, in general, is in what order each candidate will be considered on any given process. What we show is that there is an infinite sequence of increasingly long blocks of unsuccessful time steps on every process, where the block length is unbounded above as the algorithm proceeds. We also show that every sufficiently long block is a member of a set of such blocks, where all the blocks in a set have start times within γ time steps of one another. Similarly, the same can be said for all finish times. We then show that each set contains one block for each process. For a set of sufficiently long blocks, each process must start and finish a function evaluation within that process's block. The bounds (3.5) and (3.6) mean that all processes start these new function evaluations using the *same* best point. For sufficiently long blocks, all of these function evaluations must be unsuccessful. Thus, in the language of [10], the processes collectively perform a poll about the best point, and this poll is unsuccessful. The sequence of such sets of blocks is infinite, and so an infinite sequence of these best points occurs. The final conclusion, that this accumulation point is also a stationary point of f, follows automatically from our assumptions on $\mathcal{D}$ and f.

Having made these observations, we start the analysis by showing that the first process has a convergent subsequence of x's corresponding to a subsequence of step-length control parameters that goes to zero.

LEMMA 8.1. *There exists* $\hat{x} \in \mathbb{R}^n$ *and* $\hat{\mathcal{C}}_1 \subseteq \mathcal{C}_1$ *such that*

$$\lim_{\substack{t \to +\infty \\ t \in \hat{\mathcal{C}}_1}} \Delta_1^t = 0 \qquad \text{and} \qquad \lim_{\substack{t \to +\infty \\ t \in \hat{\mathcal{C}}_1}} x_1^t = \hat{x}. \tag{8.1}$$

Proof. From Corollary 7.9, we know that $\mathcal{C}_1$ is infinite and that (7.9) holds, so there exists $\mathcal{C}_1' \subseteq \mathcal{C}_1$ such that

$$\lim_{\substack{t \to +\infty \\ t \in \mathcal{C}_1'}} \Delta_1^t = 0.$$

Since the set $\{ x_1^t : t \in \mathcal{C}_1' \}$ is contained in the bounded set $\mathcal{L}(x^0)$, we can extract an infinite subset $\hat{\mathcal{C}}_1 \subset \mathcal{C}_1'$ such that the subsequence converges; i.e., there exists $\hat{x}$ in the closure of $\mathcal{L}(x^0)$ such that the limit in (8.1) holds. ∎

Next, we show that the number of time steps between each $t \in \hat{\mathcal{C}}_1$ and the most recent success on process 1 goes to $+\infty$.

COROLLARY 8.2. *Let* $\hat{\mathcal{C}}_1$ *be as defined in Lemma* 8.1. *Then there exists* $t^* \in \mathcal{T}$ *such that*

$$\underline{\kappa}(\Delta_1^t) > \eta + 2\gamma \quad \text{for all } t > t^*,\ t \in \hat{\mathcal{C}}_1,$$

where $\underline{\kappa}(\Delta)$ *is defined in* (7.3), η *is defined in* (3.5), *and* γ *is defined in* (3.6).

Proof. This follows immediately from Lemma 8.1 and (7.4). ∎

Another way to look at this corollary is to consider the step-length control parameters. By definition, $\underline{\kappa}(\Delta)$ returns the minimum number of contractions required to reduce $\Delta^{\min}$ to a given value Δ. Consider $\hat{t} > t^*$ with $\hat{t} \in \hat{\mathcal{C}}_1$. Corollary 8.2 then tells us that $\underline{\kappa}(\Delta_1^t)$ is at least $\eta + 2\gamma$. The importance of this connection with $\Delta^{\min}$ becomes clearer when we recall that (3.10) requires the Δ associated with any successful time step to satisfy $\Delta_i^t \geq \Delta^{\min}$. Therefore, we conclude that the minimum possible number of contractions since the last successful time step, at time step $\psi_1(\hat{t}\,)$, is $\eta + 2\gamma$. Since each contraction requires one function evaluation which, in turn, requires at least one time step, the situation illustrated in Figure 8.1 must hold.

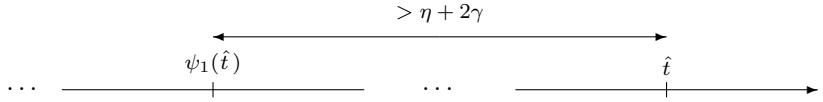


FIG. 8.1. *Relative order of events on process* 1 *when* $\hat{t} \in \hat{\mathcal{C}}_1$ *and* $\hat{t} > t^*$.

The situation illustrated in Figure 8.1 applies only to process 1. Now we show that for every $\hat{t} \in \hat{\mathcal{C}}_1$, $\hat{t} > t^*$, on each of the other processes there is a corresponding nonempty block of contiguous time steps that is devoid of successes. In particular, the situation shown in Figure 8.2 holds. The relative order between the time steps $\psi_1(\hat{t}\,) + \gamma$ and $\hat{t} - \gamma$ follows from Corollary 8.2. In the next lemma, we show that the relative order of the time steps $\psi_i(\hat{t} - \gamma)$ and $\psi_1(t) + \gamma$, as well as that of the time steps $\hat{t} - \gamma$ and $\phi_i(\psi_1(\hat{t}\,) + \gamma)$, also must hold for any $i \in \mathcal{P}$, $i \neq 1$, when $\hat{t} \in \hat{\mathcal{C}}_1$ and $\hat{t} > t^*$. The result we want then follows immediately.

LEMMA 8.3. *Let* $\hat{\mathcal{C}}_1$ *be as defined in Lemma* 8.1 *and let* t^* *be as defined in Corollary* 8.2. *Then for any* $\hat{t} \in \hat{\mathcal{C}}_1$ *with* $\hat{t} > t^*$ *and any* $i \in \mathcal{P}$, $i \neq 1$, *we have*

$$\psi_i(\hat{t} - \gamma) \leq \psi_1(\hat{t}\,) + \gamma \qquad \text{and} \tag{8.2}$$

$$\hat{t} - \gamma \leq \phi_i(\psi_1(\hat{t}\,) + \gamma), \tag{8.3}$$

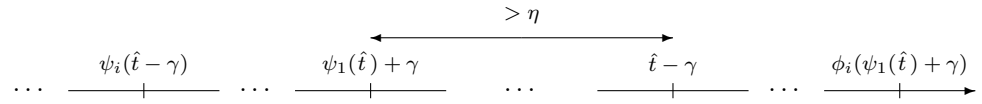


FIG. 8.2. *Relative order of events for any process $i \in \mathcal{P}$, $i \neq 1$, when $\hat{t} \in \hat{\mathcal{C}}_1$ and $\hat{t} > t^*$.*

where γ is defined in (3.6)*, $\psi_i(\cdot)$ is defined in* (7.5)*, and $\phi_i(\cdot)$ is defined in* (7.6)*. Further,*

$$\{\, t \in \mathcal{T} \,:\, \psi_1(\hat{t}\,) + \gamma < t < \hat{t} - \gamma \,\} \;\subseteq\; \mathcal{U}_i, \tag{8.4}$$

where $\mathcal{U}_i$ is defined in (7.2)*.*

Proof. Suppose not. First consider the proof for (8.2). Since the point $x_1^{\psi_1(\hat{t}\,)}$ is guaranteed to have been considered by process i by time step $\psi_1(\hat{t}\,)+\gamma$ and $\psi_1(\hat{t}\,)+\gamma < \psi_i(\hat{t}-\gamma)$ (from the null hypothesis), it must be true that

$$f(x_i^{\psi_i(\hat{t}-\gamma)}) < f(x_1^{\psi_1(\hat{t}\,)}), \tag{8.5}$$

or, equivalently for our purposes, that the tie-breaking condition in (3.11) is satisfied. Likewise, the point $x_i^{\psi_i(\hat{t}-\gamma)}$ will be considered by process 1 at some time step $t_1 \geq \psi_i(\hat{t}-\gamma)$. By the null hypothesis, we have $\psi_1(\hat{t}\,) < \psi_i(\hat{t}-\gamma)-\gamma$, so $\psi_1(\hat{t}\,) < t_1$. On the other hand, since the point $x_i^{\psi_i(\hat{t}-\gamma)}$ must be considered within γ time steps of $\psi_i(\hat{t}-\gamma)$, we have $t_1 \leq \psi_i(\hat{t}-\gamma)+\gamma$. By the definition of ψ, we conclude $t_1 \leq \hat{t}$. So we then have

$$\psi_1(\hat{t}\,) < t_1 \leq \hat{t}.$$

From (8.5), either $t_1 \in \mathcal{S}_1$, or there exists $t_2 \in \mathcal{S}_1$ with $\psi_1(\hat{t}\,) < t_2 < t_1$. In either case, we have a contradiction to the fact that $\psi_1(\hat{t}\,)$ is the most recent successful time step before $\hat{t}$ on process 1.

We follow the same line of reasoning for (8.3). Since $\phi_i(\psi_1(\hat{t}\,)+\gamma) \in \mathcal{S}_i$ (note that it is finite by the null hypothesis) and the point $x_1^{\psi_1(\hat{t}\,)}$ must have been considered by time step $\psi_1(\hat{t}\,)+\gamma$, it must be true that

$$f(x_i^{\phi_i(\psi_1(\hat{t}\,)+\gamma)}) < f(x_1^{\psi_1(\hat{t}\,)}), \tag{8.6}$$

or, equivalently for our purposes, that the tie-breaking condition in (3.11) is satisfied. Likewise, the point $x_i^{\phi_i(\psi_1(\hat{t}\,)+\gamma)}$ will be considered by process 1 by some time step t_1 satisfying

$$\psi_1(\hat{t}\,) < \phi_i(\psi_1(\hat{t}\,)+\gamma)-\gamma \leq t_1 \leq \phi_i(\psi_1(\hat{t}\,)+\gamma)+\gamma < \hat{t},$$

where the last part is from the null hypothesis and the first part is from the definition of ϕ. From (8.6), either $t_1 \in \mathcal{S}_1$ or there exists $t_2 \in \mathcal{S}_1$ with $\psi_1(\hat{t}\,) < t_2 < t_1$. In either case, we once again have a contradiction.

The proof for (8.4) follows immediately. ∎

Using the previous lemma, we can construct a set of time steps $\hat{\mathcal{C}}_i$ such that the corresponding sequence of step-length control parameters converges to zero.

LEMMA 8.4. *Consider any $i \in \mathcal{P}$, $i \neq 1$. Let $\hat{\mathcal{C}}_1$ be as defined in Lemma* 8.1 *and let t^* be as defined in Corollary* 8.2*. For any $\hat{t} \in \hat{\mathcal{C}}_1$ with $\hat{t} > t^*$ define*

$$\chi_i(\hat{t}\,) \;=\; \max\{\, t \in \mathcal{C}_i \,:\, t < \hat{t} - \gamma \,\} \tag{8.7}$$

and

$$\hat{\mathcal{C}}_i = \{\, \chi_i(\hat{t}\,) \,:\, \hat{t} > t^*,\ \hat{t} \in \hat{\mathcal{C}}_1 \,\}.$$

Then

$$\lim_{\substack{t \to +\infty \\ t \in \hat{\mathcal{C}}_i}} \Delta_i^t = 0. \tag{8.8}$$

Proof. First, we are guaranteed that

$$\chi_i(\hat{t}\,) > \psi_1(\hat{t}\,) + \gamma \quad \text{for each } \hat{t} \in \hat{\mathcal{C}}_1 \text{ with } \hat{t} > t^*$$

for the following reasons. Appealing to Corollary 8.2, we know $\underline{\kappa}(\Delta_1^{\hat{t}}) > \eta + 2\gamma$ and so the interval contains at least η time steps. Thus, one function evaluation must finish and another start on process i during that interval. Since, by Lemma 8.3, there are no successes on i between $\psi_1(\hat{t}\,) + \gamma$ and $\hat{t} - \gamma$, there must be at least one contraction on i in that interval, i.e., a $t \in \mathcal{C}_i$.

Next, from Lemma 8.1 and (7.4), we know that

$$\lim_{\substack{t \to +\infty \\ t \in \hat{\mathcal{C}}_1}} \underline{\kappa}(\Delta_1^t) = +\infty,$$

so it must also be the case that

$$\lim_{\substack{t \to +\infty \\ t \in \hat{\mathcal{C}}_1}} \frac{(t - \gamma) - (\psi_1(t) + \gamma)}{\eta} = +\infty.$$

In other words, the number of contractions in the interval defined by (8.4) is tending towards infinity. Therefore, (8.8) holds. ∎

Finally, we conclude that all processes share a common accumulation point and that such a point is a stationary point of f. This argument follows the same basic lines as those seen in [3, 9] (for the case that the search directions are restricted to the set $\mathcal{D} = \{\pm e_i, i = 1, \ldots, n\}$) and [11] (for the general case that $\mathcal{D}$ is a positive spanning set). Similar arguments have been used more recently in [8, 1, 4].

THEOREM 8.5. *Assume the function f in* (1.1) *is continuously differentiable on the closure of $\mathcal{L}(x^0)$. Then there exists $\hat{x} \in \mathbb{R}^n$ and, for each $i \in \mathcal{P}$, there exists $\hat{\mathcal{C}}_i \subset \mathcal{C}_i$ such that*

$$\lim_{\substack{t \to +\infty \\ t \in \hat{\mathcal{C}}_i}} \Delta_i^t = 0 \qquad \textit{and} \qquad \lim_{\substack{t \to +\infty \\ t \in \hat{\mathcal{C}}_i}} x_i^t = \hat{x}. \tag{8.9}$$

Furthermore,

$$\lim_{\substack{t \to +\infty \\ t \in \hat{\mathcal{C}}_i}} \nabla f(x_i^t) = 0.$$

Proof. By Lemma 8.1, we know that (8.9) holds for $i = 1$. By Lemma 8.4, we know that for each $i \in \mathcal{P}$, $i \neq 1$, we can construct $\hat{\mathcal{C}}_i$ such that the limit on Δ_i^t in (8.9) holds. Further, note that for every $\hat{t} \in \hat{\mathcal{C}}_i$, we have

$$x_i^{\chi_i(\hat{t}\,)} = x_1^{\hat{t}},$$

where $\chi_i(\hat{t}\,)$ is defined in (8.7). Thus,

$$\{\, x_i^t \,:\, t \in \hat{\mathcal{C}}_i \,\} \subseteq \{\, x_1^t \,:\, t \in \hat{\mathcal{C}}_1 \,\}.$$

So, the limit on x_i^t given in (8.9) holds as well. Hence, the claim.

Now, for any $t \in \mathcal{C}_i$, (3.15) and (3.16) give us

$$x_i^t = x_i^{t-1} \qquad \text{and} \qquad \Delta_i^t = \theta_i^t \Delta_i^{t-1}.$$

Define the set $\hat{\mathcal{B}}_i = \{\, t = \hat{t} - 1 \,:\, \hat{t} \in \hat{\mathcal{C}}_i \,\}$. Since θ_i^t is bounded below by $\theta^{\min}$, (8.9) ensures that

$$\lim_{\substack{t \to +\infty \\ t \in \hat{\mathcal{B}}_i}} \Delta_i^t = 0.$$

If $\hat{t} \in \hat{\mathcal{C}}_i$ this means that

$$f(x_i^{\hat{t}-1}) \leq f(x_i^{\hat{t}-1} + \Delta_i^{\hat{t}-1} d_i). \tag{8.10}$$

We rely here on the fact that even though the function evaluation that led to the conclusion that $\hat{t} \in \hat{\mathcal{C}}_i$ may have been initiated at some $t < \hat{t} - 1$, the update rules (3.15) and (3.16) ensure that $x_i^t = x_i^{t-1}$ and $\Delta_i^t = \Delta_i^{t-1}$ for any $t \in \mathcal{T} \setminus \mathcal{T}_i$. Since (8.10) holds for any $\hat{t} \in \hat{\mathcal{C}}_i$, this is equivalent to saying that for any $t \in \hat{\mathcal{B}}_i$

$$f(x_i^t) \leq f(x_i^t + \Delta_i^t d_i).$$

The mean value theorem then gives us

$$f(x_i^t) \leq f(x_i^t) + \Delta_i^t \nabla f(x_i^t + \sigma_i^t \Delta_i^t d_i)^T d_i$$

for some $\sigma_i^t \in [0, 1]$. Therefore,

$$0 \leq \nabla f(x_i^t + \sigma_i^t \Delta_i^t d_i)^T d_i, \quad t \in \hat{\mathcal{B}}_i.$$

Taking the limits as $t \to \infty$, we get

$$0 \leq \nabla f(\hat{x})^T d_i \quad \text{for all } i \in \mathcal{P}. \tag{8.11}$$

Since the vectors in $\mathcal{D}$ are assumed to form a positive spanning set for $\mathbb{R}^n$, (8.11) implies that $\nabla f(\hat{x}) = 0$. ∎

9. Conclusions. When developing this analysis, we tried to keep the number of assumptions made to a minimum. Our first priority was to assure that under standard assumptions, the version of APPS that we had implemented could be shown to be globally convergent. That said, there are some further relaxations we could have made. For instance, in (3.2) we assumed, for convenience, that all processes started with the same initial iterate x^0 and the same initial value Δ^0 for the step-length control parameter. While we could relax (3.2), to do so would introduce a level of complication to the analysis that does not appear to add appreciably to the fundamental result.

An extension of more obvious practical import is to allow the set of search directions to change over time. In this paper, we assume that the set of search directions is fixed. Earlier pattern search results [10] make clear that this condition can be

relaxed to allow a more general notion of *exploratory moves*. Experience with sequential implementations of pattern search has demonstrated that there certainly can be algorithmic advantage to doing so. For instance, the exploratory moves enable more aggressive or speculative steps that may either accelerate the progress of the search or move the iterates away from a local minimizer, without compromising global convergence. In the parallel setting, one of the motivations for APPS was to devise algorithms that could recover from the failure of a process. Since all we require, in the end, is that (8.11) holds for enough vectors in $\mathcal{D}$ to form a positive basis for $\mathbb{R}^n$, we have some flexibility in both the implementation and the analysis. In particular, exploratory moves are included implicitly. The exploratory moves play an active role in the search only if they produce a success, but our analysis focuses on the contractions. As long as we can express any point produced by an exploratory move as in Theorem 5.2 (i.e., any success produced by an exploratory move lies on an appropriate lattice), the analysis accommodates this extension in a straightforward fashion.

Another possible extension to the analysis is to examine the robustness of the search in the presence of process failures either because the processor on which the process resides fails or because on that particular process the evaluation of the function f at a given x fails. In the current implementation of APPS, we ignore the failure of a process so long as the search directions contained on the active processes continue to form a positive spanning set. If we experience enough process failures that this condition no longer holds, we restart enough processes so that the condition is once again satisfied. If we assume a finite number of failures for evaluation at a given point—an extension to (3.5), our assumption that the maximum number of time steps for evaluating f at a given x is finite—then the modifications required to the analysis seem straightforward enough that we simply note them here.

A more ambitious option, along the lines of related ideas proposed in [11, 8, 4], would be to actually change the set of search directions during the course of the search, rather than working with some subset of a fixed set of directions chosen at the start of the search. To do so requires some modification of the mechanism used to control the length of the step. Our analysis relies on the algebraic structure of the iterates. This can be relaxed, either by requiring Δ to go to zero in the limit [11, 4] or by introducing a sufficient decrease condition to determine the success of a step [8], in lieu of the simple decrease conditions in (3.7) and (3.11) that we use here.

We close with the observation that we can reduce the general framework presented here to a special case that looks more like traditional (sequential) pattern search. (This is what motivated us to allow $0 \leq \gamma$ so that communication can be "instantaneous," as it would be in the sequential case.) The difference here is that we have introduced the bounds given in (3.10) for $t \in \mathcal{S}_i$. These bounds are necessary for our analysis (e.g., in the proofs of Lemma 7.7 and Corollary 7.9 or for the definition of $\underline{\kappa}(\Delta)$ in (7.3), which plays a role in the proofs of Lemma 7.6, Corollary 8.2, and Lemma 8.4). Prior definitions of pattern search did not require the enforcement of (3.10) since the synchronization of the updates to Δ suffices without the imposition of these bounds on updates made after a successful step.

Acknowledgments. We thank both referees for their thoughtful comments. Among their many useful suggestions, they recommended that we provide the summaries in what is now section 4 and in the opening of what is now section 8. The presentation has been much improved by their recommendations. We also thank Margaret Wright, the associate editor, for her helpful suggestions and for her handling of the manuscript.

REFERENCES

[1] C. Audet and J. E. Dennis, Jr., *Pattern search algorithms for mixed variable programming*, SIAM J. Optim., 11 (2000), pp. 573–594.

[2] D. Bertsekas and J. Tsitsiklis, *Parallel and Distributed Computation: Numerical Methods*, Prentice-Hall, Englewood Cliffs, NJ, 1989.

[3] J. Céa, *Optimisation: Théorie et algorithmes*, Dunod, Paris, 1971.

[4] I. D. Coope and C. J. Price, *On the convergence of grid-based methods for unconstrained optimization*, SIAM J. Optim., 11 (2001), pp. 859–869.

[5] P. D. Hough, T. G. Kolda, and V. J. Torczon, *Asynchronous parallel pattern search for nonlinear optimization*, SIAM J. Sci. Comput., 23 (2001), pp. 134–156.

[6] T. G. Kolda and V. J. Torczon, *Understanding asynchronous parallel pattern search*, in High Performance Algorithms and Software for Nonlinear Optimization, G. DiPillo and A. Murli, eds., Kluwer Academic, Boston, 2003, pp. 316–335.

[7] R. M. Lewis and V. Torczon, *Rank Ordering and Positive Bases in Pattern Search Algorithms*, Tech. Rep. TR 96–71, Institute for Computer Applications in Science and Engineering (ICASE), NASA Langley Research Center, Hampton, VA, 1996.

[8] S. Lucidi and M. Sciandrone, *On the global convergence of derivative-free methods for unconstrained optimization*, SIAM J. Optim., 13 (2002), pp. 97–116.

[9] E. Polak, *Computational Methods in Optimization: A Unified Approach*, Academic Press, New York, 1971.

[10] V. Torczon, *On the convergence of pattern search algorithms*, SIAM J. Optim., 7 (1997), pp. 1–25.

[11] W. Yu, *Positive basis and a class of direct search techniques*, Sci. Sinica, Special Issue I on Math., 1 (1979), pp. 53–67 (in Chinese).

SIAM J. OPTIM.
Vol. 14, No. 4, pp. 965–979

CALCULATION OF UNIVERSAL BARRIER FUNCTIONS FOR CONES GENERATED BY CHEBYSHEV SYSTEMS OVER FINITE SETS*

LEONID FAYBUSOVICH† AND MICHAEL GEKHTMAN†

Abstract. We explicitly calculate universal barrier functions for cones generated by (weak) Chebyshev systems over finite sets. We show that universal barrier functions corresponding to Chebyshev systems on intervals are obtained as limits of universal barrier functions of their discretizations. The results rely heavily upon the classical work of Krein, Nudelman, and Schoenberg.

Key words. interior-point methods, Chebyshev systems, semi-infinite programming

AMS subject classifications. 90C51, 44A60, 90C34

DOI. 10.1137/S1052623403429585

1. Introduction. In [1] we calculated the universal barrier function for a broad class of cones generated by Chebyshev systems on intervals of the real line and the circle. In general, given a closed, convex, pointed cone K in $\mathbf{R}^n$, the universal barrier function (up to a multiplication by a positive constant) has the form [5]

$$\Phi(x) = \ln \int_{K^*} e^{-\langle x,y\rangle} d\mu(y),$$

where $x \in \mathrm{int}(K), K^*$ is the cone dual to K, and μ is the Lebesgue measure. The expression we obtained in the case of the cone generated by a Chebyshev system is of the following form:

$$\Phi(x) = \frac{1}{2} \ln \det(\bar{D}(x)),$$

where $\bar{D}(x)$ is a skew-symmetric matrix. The only complication, say, in comparison with the semidefinite programming case, is that the entries of $\bar{D}(x)$ are one-dimensional definite integrals. While there exist interesting cases in which these integrals can be explicitly calculated, in general it is important to understand what is the right way to approximate these integrals to preserve important properties of modern interior-point algorithms (e.g., complexity estimates). On the other hand, the class of optimization problems involving Chebyshev cones is a subclass of semi-infinite programming problems. Many natural procedures for finding approximate solutions to semi-infinite programming problems rely upon discretizations of semi-infinite constraints. Discretization procedures applied to Chebyshev cones lead to Chebyshev systems over finite (more generally, countable) sets. Thus, it is quite natural to try to calculate universal barrier functions for such systems. It follows from general properties of the Lebesgue integral that such barriers should converge to universal barriers of Chebyshev cones on intervals with the refinement of a discretization and, consequently, should provide a natural way to approximate those universal barriers. In

*Received by the editors June 6, 2003; accepted for publication (in revised form) November 13, 2003; published electronically May 25, 2004. This work was supported in part by NSF grant DMS01-02628.

http://www.siam.org/journals/siopt/14-4/42958.html

†Department of Mathematics, University of Notre Dame, Notre Dame, IN 46556 (leonid.faybusovich.1@nd.edu, michael.gekhtman.1@nd.edu).

particular, this approach allows us to provide a new and simpler proof of the formula for a universal barrier of a Chebyshev cone on an interval that was initially derived in [1].

The cones generated by Chebyshev systems over finite sets are defined by a finite number (equal to the cardinality of the set) of linear inequality constraints. Obviously, one can easily construct a self-concordant barrier for such cones (as the minus sum of logarithms of linear forms defining these inequalities). However, the so-called barrier parameter for this barrier will be equal to the number of inequality constraints (i.e., the cardinality of the set) and will grow rapidly with the refinement of a discretization. On the contrary, due to a deep result of Nesterov and Nemirovskii [5] the barrier parameter of the universal barrier function is of the order of dimension of the cone ($O(n)$), and hence in our case does not depend on the refinement of a discretization! Since the barrier parameter determines complexity estimates for practically all modern interior algorithms (the smaller the parameter, the better the estimates), the use of a universal barrier function for polyhedral cones is potentially quite beneficial. The problem, however, is that so far (within the class of polyhedral cones) such a universal barrier function has been calculated only for the positive orthant. More precisely, it is not known how to decompose in an efficient way an arbitrary polyhedral cone into cones linearly isomorphic to the positive orthant.

In the present paper we calculate the universal barrier function for a broad class of polyhedral cones generated by (weak) Chebyshev systems over finite sets in the computable form, as in semidefinite programming. More precisely,

$$\Phi(x) = \frac{1}{2} \ln \det \bar{D}_1(x),$$

where $\bar{D}_1(x)$ is again a skew-symmetric matrix. The entries of $\bar{D}_1(x)$ are essentially Riemann sums for entries of $\bar{D}(x)$ in the case where the corresponding finite set appears as a result of a discretization of an interval. Observe that polynomial splines belong to the class of weak Chebyshev systems and thus are covered by the results of the present paper.

We rely heavily upon the classical results of Krein, Nudelman, and Schoenberg. In our opinion, these results provide additional evidence that the modern theory of interior-point algorithms in the form developed by Nesterov and Nemirovskii has a deep mathematical structure which is currently only partially understood.

2. Chebyshev systems over finite sets. Let $\Delta = \{t_1 < t_2 < \cdots < t_m\}$ be an ordered finite set of real numbers. We say that the functions $u_i : \Delta \to \mathbf{R}, i = 0, \ldots, n$, form a Chebyshev system over Δ if

$$\det(u_i(t_{j_k})) > 0 \tag{2.1}$$

for any $1 \le j_1 < j_2 < \cdots < j_{n+1} \le m$. Introduce vectors $v_i \in \mathbf{R}^{n+1}, i = 1, \ldots, m$, where

$$v_i = (u_0(t_i), u_1(t_i), \ldots, u_n(t_i))^T.$$

Condition (2.1) can be rewritten in the form

$$\det[v_{j_1} v_{j_2} \ldots v_{j_{n+1}}] > 0 \tag{2.2}$$

for any $1 \le j_1 < j_2 < \cdots < j_{n+1} \le m$. Many examples of Chebyshev systems can be found, e.g., in [4]. In particular, a fundamental system of solutions of a linear

differential equation with constant coefficients is a Chebyshev system, provided all the roots of the characteristic equation are real.

One can naturally associate a cone with a given Chebyshev system as follows:

$$K = \left\{ x \in \mathbf{R}^{n+1} : x = (x_0, \dots, x_n)^T, \sum_{i=0}^{n} x_i u_i(t_j) \geq 0 \ \forall j \in [1, m] \right\}.$$

It is obvious that

$$K = \{x \in \mathbf{R}^{n+1} : \langle x, v_j \rangle \geq 0 \ \forall j \in [1, m]\}.$$

Here $\langle , \rangle$ is the standard scalar product in $\mathbf{R}^{n+1}$. Hence, the dual cone K^* can be described as

$$K^* = \left\{ \sum_{j=1}^{m} \rho_j v_j : \rho_j \geq 0, j = 1, \dots, m \right\}.$$

Our next goal is to describe a combinatorial concept of an index for the finite increasing sequence of integers. It will be used to parameterize the dual cone K^* using the so-called principal representations.

Let $\theta = \{1 \leq j_1 < j_2 < \cdots < j_p \leq m\}$ be a subset of the set $[1, m]$. This subset clearly can be partitioned, in a unique fashion, into subsequent blocks, each comprised of one or more integers without lacunae between them, while different blocks are separated from each other by a lacuna, as in the following example: $\theta = \{1, 2, 4, 6, 7, 9\} = \{1, 2\} \cup \{4\} \cup \{6, 7\} \cup \{9\}$. The block starting with 1 (ending with m), if any, is called *adjacent to* 1 (respectively, adjacent to m); all other blocks are called *interior*. According to the parity of k, a block is said to be even or odd. The index of an interior block θ' containing k elements is defined as

$$\epsilon(\theta') = k$$

if k is even, and

$$\epsilon(\theta') = k + 1$$

if k is odd. If a block θ' containing k elements adjoins either 1 or m, then $\epsilon(\theta') = k$. Finally, it is obvious that every θ can be partitioned into blocks $\theta_s, s = 1, 2, \dots, t$, for some t. We define the index of θ (notation: $\epsilon(\theta)$) as

$$\epsilon(\theta) = \sum_{i=1}^{t} \epsilon(\theta_s). \tag{2.3}$$

The next proposition immediately follows from definitions.

PROPOSITION 1. *We have*

$$card(\theta) \leq \epsilon(\theta).$$

Moreover, $card(\theta) = \epsilon(\theta)$ if and only if all interior blocks are even.

Here $card(\theta)$ is the number of elements in θ.

DEFINITION 1. *A sequence $\theta \subset [1, m]$ is called full if $card(\theta) = \epsilon(\theta)$.*

Let $x \in K^*$,

$$x = \sum_{i=1}^{p} \rho_i v_{j_i}, \tag{2.4}$$

$1 \le j_1 < j_2 \dots j_p \le m, \rho_i > 0, i = 1, \dots, p$. We say that (2.4) is a principal representation of x if

$$\epsilon\{j_1, \dots, j_p\} = n + 1.$$

For a proof of the following theorem see, e.g., [4, Theorem 5.1] or [6].

THEOREM 1. *Every $x \in \operatorname{int}(K^*)$ admits exactly two principal representations as follows:*

- *The block adjacent to m is even (in particular, empty). This is the so-called lower principal representation.*
- *The block adjacent to m is odd. This is the so-called upper principal representation.*

Given $\theta = \{j_1 < \dots < j_p\}$, where $1 \le j_1, j_p \le m$, denote by K^*_θ the following cone:

$$K^*_\theta = \left\{ \sum_{i=1}^{p} \rho_i v_{j_i} : \rho_i > 0 \right\}.$$

PROPOSITION 2. *Let Θ_u (respectively, Θ_l) be the set of all full subsets of $[1, m]$ of the index $n + 1$ such that the block adjacent to m is odd (respectively, even). Then*

$$S_u = \cup_{\theta \in \Theta_u} K^*_\theta \subset \operatorname{int}(K^*),$$

$$S_l = \cup_{\theta \in \Theta_l} K^*_\theta \subset \operatorname{int}(K^*).$$

Moreover, the Lebesgue measure of $\operatorname{int}(K^*) \setminus S_u$ (respectively, $\operatorname{int}(K^*) \setminus S_l$) is equal to zero. Besides,

$$K^*_\theta \cap K^*_{\theta'} = \emptyset$$

for $\theta, \theta' \in \Theta_u$ (respectively, Θ_l).

Proof. Let $\theta \subset [1, m], \theta \ne [1, m]$. Then $\epsilon(\theta) \ge card(\theta)$, and equality occurs if and only if θ is full. Let $\epsilon(\theta) = n + 1$. If $card(\theta) < n + 1$, then $\dim K^*_\theta < n + 1$ and, hence, $\mu(K^*_\theta) = 0$, where μ is the Lebesgue measure on $\mathbf{R}^{n+1}$. On the other hand, if $card(\theta) = n + 1$, then θ is full, $\dim K^*_\theta = n + 1$ (since the vectors $v_j,\ j \in \theta$, form a basis in $\mathbf{R}^{n+1}$). But then $K^*_\theta \subset \operatorname{int}(K^*)$. Using the uniqueness of upper and lower principal representations guaranteed by Theorem 1, we immediately conclude that $K^*_\theta \cap K^*_{\theta'} = \emptyset, \theta, \theta' \in \Theta_u$ (respectively, Θ_l), $\theta \ne \theta'$. The result follows.

EXAMPLE 1. *Let $n = 2, m = 5$. We have*

$$\Theta_u = \{\theta_1, \theta_2, \theta_3\},$$

$$\theta_1 = \{1, 2, 5\},\ \theta_2 = \{2, 3, 5\},\ \theta_3 = \{3, 4, 5\};$$

$$\Theta_l = \{\theta_4, \theta_5, \theta_6\},$$

$$\theta_4 = \{1,2,3\},\ \theta_5 = \{1,4,5\},\ \theta_6 = \{1,3,4\}.$$

We are now in a position to calculate the characteristic function of K.

THEOREM 2. *Let $x \in \operatorname{int}(K)$. Then*

$$I(x) = \int_{K^*} e^{-\langle x,y\rangle} d\mu(y) = \sum_{\theta=\{j_1,\dots,j_{n+1}\}\in\Theta} \frac{\det[v_{j_1},\dots,v_{j_{n+1}}]}{\prod_{l=1}^{n+1}\langle v_{j_l}, x\rangle},$$

where $\Theta = \Theta_u$ or $\Theta = \Theta_l$.

Remark 1. Since $x \in \operatorname{int}(K)$, we have $\langle x, v_j\rangle > 0\, \forall j \in [1,m]$.

Proof. Consider the case $\Theta = \Theta_u$ (the case $\Theta = \Theta_l$ is absolutely similar). By Proposition 2,

$$I(x) = \sum_{\theta\in\Theta_u} \int_{K^*_\theta} e^{-\langle x,y\rangle} d\mu(y).$$

If $y \in K^*_\theta, \theta = \{j_1,\dots,j_{n+1}\}$, we have

$$y = \sum_{k=1}^{n+1} \rho_k v_{j_k}, \qquad \rho_k > 0.$$

Hence,

$$\langle x, y\rangle = \sum_{k=1}^{n+1} \rho_k \langle x, v_{j_k}\rangle,$$

$$d\mu(y) = \det[v_{j_1},\dots,v_{j_{n+1}}] d\rho_1 d\rho_2 \dots d\rho_{n+1}.$$

We used the fact that $\det[v_{j_1},\dots,v_{j_{n+1}}] > 0$ while changing variables. Hence,

$$\begin{aligned}\int_{K^*_\theta} e^{-\langle x,y\rangle} d\mu(y) &= \int_0^{+\infty} \dots \int_0^{+\infty} \prod_{k=1}^{n+1} e^{-\rho_k\langle x, v_{j_k}\rangle} \det[v_{j_1},\dots,v_{j_{n+1}}] d\rho_1 \dots d\rho_{n+1} \\ &= \frac{\det[v_{j_1},\dots,v_{j_{n+1}}]}{\prod_{k=1}^{n+1}\langle x, v_{j_k}\rangle}.\end{aligned}$$

The result follows.

3. "Pfaffian" form of universal barrier functions. In principle, Theorem 2 provides an explicit description of the universal barrier function for the cone K. However, it is difficult to compare this description with the results of [1]. Thus, we need to find a more "computable" form of our universal barrier function that we can compare with results of [1].

Given $x \in \operatorname{int}(K)$, introduce vectors

$$a_i = \frac{v_i}{\langle x, v_i\rangle} - \frac{v_{i+1}}{\langle x, v_{i+1}\rangle}, \qquad i = 1,2,\dots,m,$$

where by definition, $v_{m+1} = 0$.

THEOREM 3. *We have*

$$\sum_{\{j_1<j_2<\dots<j_{n+1}\}\in\Theta_l} \frac{\det[v_{j_1},\dots,v_{j_{n+1}}]}{\prod_{k=1}^{n+1}\langle x, v_{j_k}\rangle} = \sum_{1\le l_1<l_2<\dots<l_{n+1}\le m} \det[a_{l_1}, a_{l_2},\dots,a_{l_{n+1}}].$$

The proof of this theorem will be given in the appendix. Here we illustrate this result by an example.

EXAMPLE 2. *Let $m = 6, n = 3$. We have*

$$\Theta_l = \{\theta_1, \theta_2, \theta_3, \theta_4, \theta_5, \theta_6\},$$

$$\theta_1 = \{1,2,3,4\},\ \theta_2 = \{1,2,4,5\},\ \theta_3 = \{1,2,5,6\},$$

$$\theta_4 = \{2,3,4,5\},\ \theta_5 = \{2,3,5,6\},\ \theta_6 = \{3,4,5,6\}.$$

With the notation introduced above, we have

$$\sum_{\{j_1<j_2<j_3<j_4\}\in\Theta_l} \frac{\det[v_{j_1}, v_{j_2}, v_{j_3}, v_{j_4}]}{\prod_{k=1}^4 \langle x, v_{j_k}\rangle}$$

$$= \det[a_1, a_2, a_3, a_4 + a_5 + a_6] + \det[a_1, a_2 + a_3, a_4, a_5 + a_6] + \det[a_1, a_2 + a_3 + a_4, a_5, a_6]$$

$$+ \det[a_2, a_3, a_4, a_5 + a_6] + \det[a_2, a_3 + a_4, a_5, a_6] + \det[a_3, a_4, a_5, a_6]$$

$$= \sum_{1\le l_1<l_2<l_3<l_4\le 6} \det[a_{l_1}, a_{l_2}, a_{l_3}, a_{l_4}].$$

The analogue of Theorem 4 for Θ_u can be formulated as follows.

THEOREM 4. *We have*

$$\sum_{\{j_1<j_2<\cdots<j_{n+1}\}\in\Theta_u} \frac{\det[v_{j_1}, \ldots, v_{j_{n+1}}]}{\prod_{k=1}^{n+1}\langle x, v_{j_k}\rangle} = \sum_{1\le l_1<l_2<\cdots<l_n<m} \frac{\det[a_{l_1}, a_{l_2}, \ldots, a_{l_n}, v_m]}{\langle x, v_m\rangle}.$$

The proof of Theorem 4 is completely analogous to the proof of Theorem 3 given in the appendix.

EXAMPLE 3. *Let $m = 6, n = 2$. We have*

$$\Theta_u = \{\theta_1, \theta_2, \theta_3, \theta_4\},$$

$$\theta_1 = \{1,2,6\},\ \theta_2 = \{2,3,6\},\ \theta_3 = \{3,4,6\},$$

$$\theta_4 = \{4,5,6\}.$$

With the notation introduced above, we have

$$\sum_{\{j_1<j_2<j_3\}\in\Theta_u} \frac{\det[v_{j_1}, v_{j_2}, v_{j_3}]}{\prod_{k=1}^3 \langle x, v_{j_k}\rangle}$$

$$= \frac{1}{\langle x, v_6\rangle}(\det[a_1, a_2 + a_3 + a_4 + a_5, v_6] + \det[a_2, a_3 + a_4 + a_5, v_6]$$

$$+\det[a_3, a_4 + a_5, v_6] + \det[a_4, a_5, v_6])$$

$$= \sum_{1 \le j_1 < j_2 < j_3 \le 5} \frac{\det[a_{j_1}, a_{j_2}, a_{j_3}, v_6]}{\langle x, v_6 \rangle}.$$

Let $b_1, b_2, \ldots, b_m$ be vectors in an even-dimensional vector space $\mathbf{R^{2r}}, m \ge 2r$. Let, further,

$$b_i = (b_i(1), b_i(2), \ldots, b_i(2r))^T,$$

$$d(\alpha, \beta) = \sum_{1 \le i < j \le m} \det \begin{bmatrix} b_i(\alpha) & b_j(\alpha) \\ b_i(\beta) & b_j(\beta) \end{bmatrix},$$

$\alpha, \beta = 1, 2, \ldots, 2r$. Let D be a skew-symmetric $2r \times 2r$ matrix such that

$$D(\alpha, \beta) = d(\alpha, \beta), \qquad \alpha, \beta = 1, 2, \ldots, 2r.$$

The next proposition is due to Okada [7, Theorem 3].

PROPOSITION 3. *We have*

$$Pf(D) = \sum_{1 \le j_1 < j_2 < \cdots < j_{2r} \le m} \det[b_{j_1}, b_{j_2}, \ldots, b_{j_{2r}}].$$

Here $Pf(D)$ stands for the Pfaffian of an even-dimensional skew-symmetric matrix D.

Remark 2. One substantial property of Pfaffians is that

$$Pf(D)^2 = \det(D).$$

Hence, $\ln Pf(D) = \frac{1}{2} \ln \det(D)$. For a good introductory discussion of major properties of Pfaffians, we recommend [2].

LEMMA 1. *Let $(z(i,j)), i, j = 1, 2, \ldots, N+1$, be a skew-symmetric matrix such that $z(i, N+1) = 0, i = 1, \ldots, N$. Then*

$$S = \sum_{1 \le i < j \le N} (z(i,j) + z(i+1, j+1) - z(i, j+1) - z(i+1, j)) = \sum_{i=1}^{N-1} z(i, i+1).$$

Proof. We have

$$S = \sum_{1 \le i < j \le N} z(i,j) + \sum_{2 \le i < j \le N+1} z(i,j) - \sum_{2 \le i \le j \le N} z(i,j) - \sum_{1 \le i < j-1 \le N} z(i,j).$$

Combining the first and the third (respectively, the second and the fourth) terms and taking into account that $z(i,i) = 0$, we obtain

$$S = \sum_{j=2}^{N} z(1,j) + \sum_{i=2}^{N} z(i, i+1) - \sum_{j=3}^{N+1} z(1,j)$$

$$= \sum_{i=1}^{N} z(i, i+1) - z(1, N+1).$$

The result follows.

We are now in a position to describe the characteristic function of a cone generated by a Chebyshev system over a finite set in the form similar to [1]. Recall that such a system is determined by a finite set of vectors $v_i, i = 1, \dots, m$, in $\mathbf{R}^{n+1}$ satisfying (2.1), where

$$v_i = (v_i(0), v_i(1), \dots, v_i(n))^T,$$

$$v_i(k) = u_k(t_i), \qquad k = 0, 1, \dots, n.$$

THEOREM 5. *Let $x \in \operatorname{int}(K)$ and n be odd. Then*

$$I(x) = \int_{K^*} e^{-\langle x,y\rangle} d\mu(y) = Pf(D(x)),$$

where $D(x) = (d(\alpha,\beta)), \alpha, \beta = 0, 1, \dots, n,$

$$d(\alpha,\beta) = \sum_{i=1}^{m-1} \frac{\det \begin{bmatrix} v_i(\alpha) & v_{i+1}(\alpha) \\ v_i(\beta) & v_{i+1}(\beta) \end{bmatrix}}{\langle x, v_i\rangle \langle x, v_{i+1}\rangle}$$

$$= \sum_{i=1}^{m-1} \frac{\det \begin{bmatrix} u_\alpha(t_i) & u_\alpha(t_{i+1}) - u_\alpha(t_i) \\ u_\beta(t_i) & u_\beta(t_{i+1}) - u_\beta(t_i) \end{bmatrix}}{\langle x, v_i\rangle \langle x, v_{i+1}\rangle}.$$

Proof. By Theorems 2, 3, and Proposition 3, we have

$$I(x) = Pf(D), \qquad D = (d(\alpha,\beta)),$$

$$d(\alpha,\beta) = \sum_{1 \le i < j \le m} \det \begin{bmatrix} a_i(\alpha) & a_j(\alpha) \\ a_i(\beta) & a_j(\beta) \end{bmatrix},$$

where

$$a_i = \frac{v_i}{\langle x, v_i\rangle} - \frac{v_{i+1}}{\langle x, v_{i+1}\rangle}, \qquad i = 1, 2, \dots, m.$$

Let

$$z(i,j) = \det \begin{bmatrix} \tilde{v_i}(\alpha) & \tilde{v_j}(\alpha) \\ \tilde{v_i}(\beta) & \tilde{v_j}(\beta) \end{bmatrix},$$

where

$$\tilde{v_i} = \frac{v_i}{\langle x, v_i\rangle}.$$

Then

$$d(\alpha,\beta) = \sum_{1 \le i < j \le m} (z(i,j) + z(i+1,j+1) - z(i,j+1) - z(i+1,j)).$$

The result follows from Lemma 1.

Let $v_1, v_2, \ldots, v_m$ be vectors in $\mathbf{R}^{n+1}$ associated with a Chebyshev system $u_0, \ldots, u_n$ via (2.2), and let L be a linear isomorphism of $\mathbf{R}^{n+1}$ such that $\det L > 0$. Then $Lv_1, Lv_2, \ldots, Lv_m$ is also a Chebyshev system in $\mathbf{R}^{n+1}$ since

$$\det[Lv_{j_1}, \ldots, Lv_{j_{n+1}}] = \det L \det[v_{j_1}, \ldots, v_{j_{n+1}}].$$

Moreover, if we denote by K_L the corresponding cone generated by $Lv_1, \ldots, Lv_m$, i.e.,

$$K_L = \{x \in \mathbf{R}^{n+1} : \langle x, Lv_i \rangle \geq 0 \ \forall i = 1, \ldots, m\},$$

then

$$K_L = L^{-T}K, \qquad K_L^* = LK^*.$$

In other words, the cones K and K_L are linearly isomorphic. Hence, their characteristic functions coincide up to a multiplicative constant. One can always find such an L so that

$$Lv_m = (0, \ldots, 0, 1)^T.$$

THEOREM 6. *Let $x \in \operatorname{int}(K), n$ be even, and $v_m = (0, \ldots, 0, 1)^T$. Then*

$$I(x) = \int_{K^*} e^{-\langle x,y \rangle} d\mu(y) = \frac{Pf(D_1(x))}{\langle x, v_m \rangle},$$

where

$$D_1(x) = (d_1(\alpha, \beta)), \qquad \alpha, \beta = 0, 1, \ldots, n-1,$$

$$d_1(\alpha, \beta) = \sum_{i=1}^{m-2} \frac{\det \begin{bmatrix} v_i(\alpha) & v_{i+1}(\alpha) \\ v_i(\beta) & v_{i+1}(\beta) \end{bmatrix}}{\langle x, v_i \rangle \langle x, v_{i+1} \rangle}$$

$$= \sum_{i=1}^{m-2} \frac{\det \begin{bmatrix} u_\alpha(t_i) & u_\alpha(t_{i+1}) - u_\alpha(t_i) \\ u_\beta(t_i) & u_\beta(t_{i+1}) - u_\beta(t_i) \end{bmatrix}}{\langle x, v_i \rangle \langle x, v_{i+1} \rangle}.$$

Proof. By Theorems 2 and 4, we have

$$I(x) = \sum_{1 \leq l_1 < l_2 < \cdots < l_n \leq m-1} \frac{\det[a_{l_1}, a_{l_2}, \ldots, a_{l_n}, v_m]}{\langle x, v_m \rangle}.$$

Taking into account $v_m = (0, \ldots, 0, 1)^T$ and expanding determinants over the last column, we obtain

$$I(x) = \sum_{1 \leq l_1 < l_2 < \cdots < l_n \leq m-1} \frac{\det[\tilde{a}_{l_1}, \ldots, \tilde{a}_{l_n}]}{\langle x, v_m \rangle}.$$

Here $\tilde{a}_i \in \mathbf{R}^n$,

$$\tilde{a}_i(j) = a_i(j) = \frac{v_i(j)}{\langle x, v_i \rangle} - \frac{v_{i+1}(j)}{\langle x, v_i \rangle},$$

$j = 0, 1, \ldots, n-1$. We can now finish the proof exactly as in Theorem 5.

DEFINITION 2. *We say that the functions $u_i : \Delta \to \mathbf{R}, i = 0, \dots, n$, form a periodic Chebyshev system if the corresponding vectors $v_1, \dots, v_m \in \mathbf{R}^{n+1}$ satisfy $v_1 = v_m$ and*

$$\det[v_{j_1}, \dots, v_{j_{n+1}}] > 0$$

for all subsets $\{1 \le j_1 < j_2 < \cdots < j_{n+1} \le m\}$, except for the subsets with $j_1 = 1$ and $j_{n+1} = m$.

LEMMA 2. *If $v_1, \dots, v_m \in \mathbf{R}^{n+1}$ correspond to a periodic Chebyshev system, then n is even.*

Proof. Take $j_{n+1} = m, j_n = m-1, \dots, j_1 = m-n$. Then

$$\det[v_{j_1}, \dots, v_{j_{n+1}}] = (-1)^n \det[v_{j_{n+1}}, v_{j_1}, v_{j_2}, \dots, v_{j_n}]$$

$$= (-1)^n \det[v_1, v_{j_1}, \dots, v_{j_n}].$$

Both the first and last determinants should be positive. Hence, n is even.

If $v_1, \dots, v_m$ correspond to a periodic Chebyshev system, then $v_2, v_3, \dots, v_m$ form a usual Chebyshev system, and cones generated by these systems obviously coincide. Hence, we can apply Theorem 6 to calculate the universal barrier function for the cone generated by a periodic Chebyshev system.

Now let $u_0, \dots, u_n$ be a Chebyshev system of continuously differentiable functions on the interval $[a, b]$. Let

$$C = \left\{ (x_0, \dots, x_n) : \sum_{i=0}^{n} x_i u_i(t) \ge 0 \ \forall t \in [a, b] \right\}.$$

In [1] we calculated the universal barrier function for C. Suppose that n is odd. Then, given $x \in \operatorname{int}(C)$,

$$I_C(x) = Pf(D_2(x)), D_2(x) = (d_2(\alpha, \beta)),$$

$\alpha, \beta = 0, 1, \dots, n$, where

$$d_2(\alpha, \beta) = \int_a^b \frac{\det \begin{bmatrix} u_\alpha(t) & \dot{u}_\alpha(t) \\ u_\beta(t) & \dot{u}_\beta(t)) \end{bmatrix}}{x(t)^2} dt,$$

$$x(t) = \sum_{i=0}^{n} x_i u_i(t).$$

Choose $a \le t_1 < t_2 < \cdots < t_m \le b$ for some $m \ge n+1$. It is clear that $u_0, \dots, u_n$ form a Chebyshev system over the finite set $\Delta = \{t_1, \dots, t_m\}$. It is quite obvious that $C \subset K = K_\Delta$. Hence, we can compare $I_C(x)$ and $I_K(x)$. By Theorem 5 we see that $d(\alpha, \beta)$ is essentially a Riemann sum for $d_2(\alpha, \beta)$ (observe that $\langle x, v_i \rangle = x(t_i)\ \forall i$). The difference between the corresponding formulas for the case n even is explained by the fact that in Theorem 6 we used the upper principal representation, whereas in Theorem 5 of [1] we used the lower principal representation. More precisely, let us divide the interval $[a, b]$ into 2^k equal parts by the points

$$t_\nu = a + \frac{(b-a)(\nu - 1)}{2^k}, \qquad \nu = 1, \dots, m = 2^k + 1.$$

Let, further, $\Delta_k = \{t_0 < t_1 < \cdots < t_{2^k+1}\}$. Denote by K_{Δ_k} the cone generated by the corresponding Chebyshev system. It is clear that $K_{\Delta_{K_1}} \supset K_{\Delta_2} \supset \cdots \supset C$ and hence, $K^*_{\Delta_1} \subset K^*_{\Delta_2} \subset \cdots \subset C^*$.

THEOREM 7. *Let $x \in \operatorname{int}(C)$. Then the sequence $I_{K_{\Delta_k}}(x), k = 1, 2, \ldots$, is monotonically increasing and*

$$I_{K_{\Delta_k}}(x) \to I_C(x), \qquad k \to \infty.$$

Moreover, the convergence is uniform on any compact subset in $\operatorname{int}(C)$.

The result easily follows from the fact that entries of the corresponding skew-symmetric matrices as described in Theorems 5 and 6 converge to the corresponding integrals $d_2(\alpha, \beta)$.

We conclude this section with a unifying result that combines formulae for universal barrier functions obtained in [1] and in Theorems 5 and 6. Consider a union of disjoint closed intervals $I = \cup_{i=1}^N [a_i, b_i]$ $(a_1 < b_1 < a_2 < b_2 < \cdots < a_N < b_N)$ and a collection of (possibly empty) finite sets Δ_i $(i = 0, \ldots, N)$ such that $\Delta_0 \subset (-\infty, a_1), \Delta_i \subset (b_i, a_{i+1})$ $(i = 1, \ldots, N-1)$, $\Delta_n \subset (b_N, \infty)$. If Δ_i is nonempty, we describe it by $\Delta_i = \{t_{i1} < t_2 < \cdots < t_{im_i}\}$. It will also be convenient for us to denote $t_{im_{i+1}} = a_{i+1}$ $(i = 0, \ldots, N-1)$ and $t_{i0} = b_i$ $(i = 1, \ldots, N)$.

Define a set $\Delta = I \cup (\cup_{i=0}^N \Delta_i)$. Assume that the functions $u_i : \Delta \to \mathbf{R}, i = 0, \ldots, n$, form a Chebyshev system over Δ and that K is the corresponding Chebyshev cone. As in (2.2), we introduce vectors $v(t) \in \mathbf{R}^{n+1}$, where

$$v(t) = (u_0(t), u_1(t), \ldots, u_n(t))^T.$$

THEOREM 8. *Let $x \in \operatorname{int}(K)$ and n be odd. Then*

$$I(x) = \int_{K^*} e^{-\langle x, y\rangle} d\mu(y) = Pf(D(x)),$$

where $D(x) = \|d(\alpha, \beta)\|$, $\alpha, \beta = 0, 1, \ldots, n$,

$$d(\alpha, \beta) = \sum_{i=0}^{N} \sum_{j=0}^{m_i} \frac{\det \begin{bmatrix} u_\alpha(t_{ij}) & u_\alpha(t_{i,j+1}) \\ u_\beta(t_{ij}) & u_\beta(t_{i,j+1}) \end{bmatrix}}{\langle x, v(t_{ij})\rangle \langle x, v(t_{i,j+1})\rangle} + \sum_{i=1}^{N} \int_{a_i}^{b_i} \frac{\det \begin{bmatrix} u_\alpha(t) & \dot{u}_\alpha(t) \\ u_\beta(t) & \dot{u}_\beta(t)) \end{bmatrix}}{\langle x, v(t)\rangle^2} dt.$$

The proof of this statement is completely analogous to that of Theorem 7. A similar result holds for n even, in which case the system has to be normalized so that $v(\sup \Delta) = (0, \ldots, 0, 1)^T$ and the right-hand side of the formula for $I(x)$ has to be divided by $\langle x, v(\sup \Delta)\rangle$.

4. Extensions. Suppose that, instead of (2.1), the following weaker condition is satisfied:

$$\det(u_i(t_{j_k})) \geq 0 \quad \forall 1 \leq j_1 < j_2 < \cdots < j_{n+1} \leq m.$$

If the corresponding $(n+1) \times m$ matrix

$$[v_1, \ldots, v_m]$$

has the maximal rank $n+1$ (which is equivalent to saying that $\det[v_{j_1}, \ldots, v_{j_{n+1}}] > 0$ for at least one ordered set $1 \leq j_1 < j_2 < \cdots < l_{j_{n+1}} \leq m$), then $v_1, \ldots, v_m$ is called

a *weak Chebyshev system* (using terminology from [3]). We will assume that the cone K,

$$K = \{x \in \mathbf{R}^{n+1} : \langle x, v_i \rangle \geq 0 \ \forall i\},$$

is pointed, i.e., $\text{int}(K) \neq \emptyset$ and K does not contain straight lines (observe that if $v_1, \dots, v_m$ is a Chebyshev system, then this is always the case [4]). Our last assumption is that $v_i \neq 0 \ \forall i$. One can easily see that, under these assumptions,

$$\text{int}(K) = \{x \in \mathbf{R}^{n+1} : \langle x, v_i \rangle > 0 \ \forall i\}.$$

As is well known, the dual cone K^* is also pointed. Following a classical argument of Schoenberg [9], consider for a given $0 < q < 1$ an $m \times m$ matrix

$$X(q) = (x_{ij}(q)), \ x_{ij}(q) = q^{(i-j)^2}, \qquad i, j = 1, 2, \dots, m.$$

The major property of $X(q)$, which we are going to use, is that all minors of $X(q)$ are positive. This easily follows from the identity

$$X(q) = \text{diag}(q, q^{2^2}, \dots, q^{i^2}, \dots, q^{m^2}) W(q) \text{diag}(q, q^{2^2}, \dots, q^{m^2}),$$

where $W(q) = (w_{ij}(q)), w_{ij}(q) = (q^{-2i})^j, ij, = 1, 2, \dots m$, is a Vandermonde matrix. Consider the $(n+1) \times m$ matrix

$$[v_1(q), \dots, v_m(q)] = [v_1, \dots, v_m] X(q).$$

Observe that

$$v_i(q) = \sum_{j=1}^{m} x_{ij}(q) v_j. \tag{4.1}$$

Denote by $\delta(i_1, \dots, i_{n+1}; j_1, \dots, j_{n+1})(q)$ the minor of $X(q)$ corresponding to rows $i_1 < i_2 < \cdots < i_{n+1}$ and columns $j_1 < j_2 < \cdots < j_{n+1}$. By the Binet–Cauchy formula,

$$\begin{aligned} &\det[v_{i_1}(q), \dots, v_{i_{n+1}}(q)] \\ &\quad = \sum_{1 \leq j_1 < j_2 < \cdots < j_{n+1} \leq m} \delta(j_1, \dots, j_{n+1}; i_1, \dots, i_{n+1})(q)) \det[v_{j_1}, \dots, v_{j_{n+1}}], \\ &1 \leq i_1 < i_2 < \cdots < i_{n+1} \leq m. \end{aligned}$$

Hence, if at least one $\det[v_{j_1}, \dots, v_{j_{n+1}}] > 0$ (which is the case if K is pointed), then $v_1(q), \dots, v_m(q)$ form a Chebyshev system for any $0 < q < 1$. Since $X(q) \to I$, when $q \to 0$ (I is the identity matrix), we can use this construction to calculate the characteristic function for the cone K. Denote by $K(q)$ the cone generated by the Chebyshev system $v_1(q), \dots, v_m(q)$. It is obvious from (4.1) that $K \subset K(q), \text{int}(K) \subset \text{int}(K(q)), 0 < q < 1$. Thus, $K(q)^* \subset K^*, \text{int}(K(q)^*) \subset \text{int}(K^*)$. Given $A \subset \mathbf{R}^{n+1}$, denote by $\chi_A : \mathbf{R}^{n+1} \to \mathbf{R}$ the function such that $\chi_A(y) = 1$ if $y \in A$; $\chi_A(y) = 0$ otherwise.

We will need the following geometrically evident lemma.

LEMMA 3. *Let $M = cone(w_1, w_2, \dots, w_l)$ be a finitely generated cone with a nonempty interior in a finite-dimensional vector space V. Then $x \in \text{int}(M)$ if and only if it admits a representation of the form*

$$x = \sum_{i=1}^{l} \lambda_i w_i,$$

where all λ_i are strictly positive.

Proof. Let $e_1, \dots, e_l$ be a canonical basis in $\mathbf{R}^l$. Consider a linear map $B : \mathbf{R}^l \to V, Be_i = w_i, i = 1, \dots, l$. Then $M = B(\mathbf{R}^l_+)$, and hence $\operatorname{int}(M) = B(\operatorname{int}(\mathbf{R}^l_+))$ (see, e.g., [8]). The result follows.

LEMMA 4. *For any sequence $q_1 > q_2 > \cdots$, converging to zero, we have*

$$\lim \chi_{\operatorname{int}(K(q_i)^*)}(y) = \chi_{\operatorname{int}(K^*)}(y), \qquad i \to \infty \ \forall y \in \mathbf{R}^{n+1}. \tag{4.2}$$

Proof. Let $y \in \operatorname{int}(K^*)$. Then by Lemma 2, y admits a representation of the form

$$y = \sum_{i=1}^{m} \lambda_i v_i, \tag{4.3}$$

where all λ_i are strictly positive. Let us show that $y \in \operatorname{int}(K(q)^*)$ for all sufficiently small q. By Lemma 2 it suffices to indicate a representation

$$y = \sum_{i=1}^{m} \lambda_i(q) v_i(q) \tag{4.4}$$

with $\lambda_i(q) > 0 \ \forall i$. Comparing (4.2) with (4.3), we see that one can take $\lambda_i(q)$ to be solutions of the system of linear equations

$$X(q)\mu = \lambda, \tag{4.5}$$

where $\mu = (\mu_1, \dots, \mu_m)^T, \lambda = (\lambda_1, \dots, \lambda_m)^T$. But $X(q)$ tends to the identity matrix when q tends to zero. Hence, $X(q)^{-1}$ tends to the identity matrix when q tends to zero. But then all components of the solution to (4.4) will be positive for sufficiently small q (since all λ_i are positive). This provides a representation in the form of (4.4) for all sufficiently small q. Thus, $y \in \operatorname{int}(K(q)^*)$ for all sufficiently small q. The result follows.

THEOREM 9. *Let $x \in \operatorname{int}(K)$. Then (in the notation of the previous lemma)*

$$I_{K_{q_i}}(x) \to I_K(x), \qquad i \to \infty.$$

Proof. We have

$$I_{K_{q_i}}(x) = \int_{K^*_{q_i}} e^{-\langle x,y\rangle} d\mu(y) = \int_{\operatorname{int}(K^*_{q_i})} e^{-\langle x,y\rangle} d\mu(y)$$

$$= \int_{\mathbf{R}^{n+1}} \chi_{\operatorname{int}(K^*_{q_i})}(y) e^{-\langle x,y\rangle} d\mu(y) \to \int_{\mathbf{R}^{n+1}} \chi_{\operatorname{int}(K^*)}(y) e^{-\langle x,y\rangle} d\mu(y)$$

$$= \int_{K^*} e^{-\langle x,y\rangle} d\mu(y).$$

The convergence follows by Lemma 4, the Lebesgue dominated convergence theorem, and by (obvious) inequalities

$$0 \le \chi_{\operatorname{int}(K^*_{q_i})} \le \chi_{\operatorname{int}(K^*)}.$$

COROLLARY 1. *Theorems 2–6 hold for a weak Chebyshev system $v_1, \dots, v_m$ satisfying the following two additional conditions:*

- *The corresponding cone K is pointed.*
- *All vectors v_i are nonzero.*

Proof. Since Theorem 2 is true for each cone K_q, it is also true for K: it suffices to take limit $q \to 0$ and apply the previous theorem. The remaining theorems are derived from Theorem 2 exactly as in the case of a Chebyshev system.

EXAMPLE 4. *Consider the following system of functions:*

$$t^l, t^{l-1}, \dots, t, 1, (t-x_1)_+^l, (t-x_2)_+^l, \dots, (t-x_r)_+^l$$

on the interval $[-1,1]$. *Here* $-1 < x_1 < x_2 < \cdots < x_r < 1$ *and* $x_+ = \max\{x, 0\}$. *The linear combinations of these functions are called spline polynomials of degree l with knots $x_1, \dots, x_r$. These functions form a weak Chebyshev system* (*see, e.g.,* [3]). *Thus, we can apply our results to its discretizations.*

Appendix.

Proof of Theorem 3 (compare with [6]). For $i = 1, \dots, m$, denote $\frac{v_i}{\langle x, v_i\rangle}$ by v_i' and observe that, for $i < j$, $v_i' - v_j' = a_i + a_{i+1} + \cdots + a_{j-1}$. Therefore,

$$\begin{aligned}\frac{\det[v_{j_1}, \dots, v_{j_{n+1}}]}{\prod_{k=1}^{n+1}\langle x, v_{j_k}\rangle} &= \det[v'_{j_1}, \dots, v'_{j_{n+1}}] = \det[v'_{j_1} - v'_{j_2}, \dots, v'_{j_{n+1}} - v'_{m+1}] \\ &= \sum_{j_1 \le l_1 < j_2 \le l_2 < j_3 \le \cdots \le j_{n+1} \le l_{n+1} \le m} \det[a_{l_1}, a_{l_2}, \dots, a_{l_{n+1}}].\end{aligned}$$

To complete the proof, one needs to show that, for every collection of indices $I := \{1 \le l_1 < l_2 < \cdots < l_{n+1} \le m\}$, there exists exactly one $(n+1)$-tuple $\theta(I) = \{j_1 < j_2 < \cdots < j_{n+1}\} \in \Theta_l$ such that $j_1 \le l_1 < j_2 \le l_2 < j_3 \le \cdots \le j_{n+1} \le l_{n+1} \le m$. This can be done by induction on m and n. Indeed, let k be the smallest integer such that $l_{k+1} > k+1$. Then $l_i = i$ for $i = 1, \dots, k$ and, therefore, $\theta(I)$ must also satisfy $j_i = i$ for $i = 1, \dots, k$. If $\theta(I) \in \Theta_l$, then the parity of the size of the block that contains 1 is equal to the parity of $n+1$. Thus, if $k \equiv n+1 (\mathrm{mod}\ 2)$, we must have $k+1 < j_{k+1} \le l_{k+1}$; otherwise $j_{k+1} = k+1$ and $l_{k+1} < j_{k+2} \le l_{k+2}$.

Let k' be equal to $k+1$ if $k \equiv n+1 (\mathrm{mod}\ 2)$ and to $k+2$ otherwise. We see that $j_{k'}$ is the smallest index of the second block in $\theta(I)$. Since this block must be of even size, and since $j_{k'} \le l_{k'} < j_{k'+1}$, we conclude that $j_{k'} = l_{k'}$ and $j_{k'+1} = l_{k'} + 1 \le l_{k'+1}$. Now if we define m', n', and l_i', j_i' $(i = 1, \dots, n-k')$ by setting $m' = m - l_{k'+1}$, $n' = n - k' - 1$, $l_i' = l_{k'+i+1} - l_{k'+1}$, and $j_i' = j_{k'+i+1} - l_{k'+1}$, then $1 \le l_1' < \cdots < l_{n'+1}' \le m'$ and $\theta(I)$ satisfies needed properties if and only if the collection $j_1' < \cdots < j_{n+1}'$ belongs to Θ_l and $j_1' \le l_1' < j_2' \le l_2' < j_3' \le \cdots \le j_{n'+1}' \le l_{n'+1}' \le m'$. The statement now follows from the induction assumption.

Remark 3. The proof of Theorem 4 is completely analogous to the proof of Theorem 3.

REFERENCES

[1] L. FAYBUSOVICH, *Self-concordant barriers for cones generated by Chebyshev systems*, SIAM J. Optim., 12 (2002), pp. 770–781.

[2] J. FELDMAN, H. KNORRER, AND E. TRUBOWITZ, *Fermionic Functional Integrals and the Renormalization Group*, AMS, Providence, RI, 2002.

[3] S. KARLIN AND W. J. STUDDEN, *Tchebycheff systems: With applications in analysis and statistics*, Interscience, New York, 1966.

[4] M. G. KREIN AND A. A. NUDELMAN, *The Markov Moment Problem and Extremal Problems. Ideas and Problems of P. L. Chebyshev and A. A. Markov and Their Further Development*, AMS, Providence, RI, 1977.

[5] Y. Nesterov and A. Nemirovskii, *Interior-Point Polynomial Algorithms in Convex Programming*, SIAM Stud. Appl. Math. 13, SIAM, Philadelphia, 1994.
[6] A. A. Nudelman, *Isoperimetric problems for the convex hulls of polygonal lines and curves in higher-dimensional spaces*, Mat. Sb., 96 (1975), pp. 294–313.
[7] S. Okada, *On the generating functions for certain classes of plane partitions*, J. Combin. Theory Ser. A, 51 (1989), pp. 1–23.
[8] R. T. Rockafellar, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.
[9] I. J. Schoenberg, *An isoperimetric inequality for closed curves in even-dimensional Euclidean spaces*, Acta Math., 91 (1954), pp. 143–164.

SIAM J. OPTIM.
Vol. 14, No. 4, pp. 980–1010

A PATTERN SEARCH FILTER METHOD FOR NONLINEAR PROGRAMMING WITHOUT DERIVATIVES*

CHARLES AUDET† AND J. E. DENNIS, JR.‡

Abstract. This paper formulates and analyzes a pattern search method for general constrained optimization based on filter methods for step acceptance. Roughly, a filter method accepts a step that improves either the objective function value or the value of some function that measures the constraint violation. The new algorithm does not compute or approximate any derivatives, penalty constants, or Lagrange multipliers. A key feature of the new algorithm is that it preserves the division into SEARCH and local POLL steps, which allows the explicit use of inexpensive surrogates or random search heuristics in the SEARCH step. It is shown here that the algorithm identifies limit points at which optimality conditions depend on local smoothness of the functions and, to a greater extent, on the choice of a certain set of directions. Stronger optimality conditions are guaranteed for smoother functions and, in the constrained case, for a fortunate choice of the directions on which the algorithm depends. These directional conditions generalize those given previously for linear constraints, but they do not require a feasible starting point. In the absence of general constraints, the proposed algorithm and its convergence analysis generalize previous work on unconstrained, bound constrained, and linearly constrained generalized pattern search. The algorithm is illustrated on some test examples and on an industrial wing planform engineering design application.

Key words. pattern search algorithm, filter algorithm, surrogate-based optimization, derivative-free convergence analysis, constrained optimization, nonlinear programming

AMS subject classifications. 90C30, 90C56, 65K05

DOI. 10.1137/S105262340138983X

1. Introduction.

The optimization problem considered in this paper is

$$\begin{aligned} \min_{x\in X} \quad & f(x) \\ \text{s.t.} \quad & C(x) \le 0, \end{aligned} \tag{1.1}$$

where $f : X \to \mathbb{R} \cup \{\infty\}$ and $C : X \to (\mathbb{R} \cup \{\infty\})^m$ are functions with $C = (c_1, \ldots, c_m)^T$, and X is a full dimensional polyhedron in $\mathbb{R}^n$ defined by finitely many nondegenerate explicit bound and linear constraints. It is possible, for instance when the functions are provided as "black box" subroutine calls, that some constraints might be linear without the knowledge of the user. In that case, these linear constraints are incorporated in $C(x) \le 0$. The region defined by feasibility with respect to the constraints $C(x) \le 0$ is denoted by Ω. The combined feasible region with respect to both sets of constraints is $X \cap \Omega$. The proposed approach combines aspects of filter algorithms to handle Ω and a "barrier" approach to maintain feasibility with respect to X.

*Received by the editors May 25, 2001; accepted for publication (in revised form) November 6, 2003; published electronically May 25, 2004. The work of the first author was supported by FCAR grant NC72792, NSERC grant 239436-01, and fellowship PDF-207432-1998 in part during a post-doctoral stay at Rice University, and both authors were supported by AFOSR F49620-01-1-0013, the Boeing Company, Sandia LG-4253, ExxonMobil, LANL Computer Science Institute (LACSI) contract 03891-99-23, and the Sandia Computer Science Research Institute (CSRI).

http://www.siam.org/journals/siopt/14-4/38983.html

†GERAD and Département de Mathématiques et de Génie Industriel, École Polytechnique de Montréal, C.P. 6079, Succ. Centreville, Montréal, Québec, H3C 3A7 Canada (Charles.Audet@gerad.ca, www.gerad.ca/Charles.Audet).

‡Computational and Applied Mathematics Department, Rice University, MS 134, 6100 South Main Street, Houston, TX 77005-1892 (dennis@caam.rice.edu, www.caam.rice.edu/∼dennis).

Filter algorithms were introduced by Fletcher and Leyffer [14] as a way to globalize sequential linear programming (SLP) and sequential quadratic programming (SQP) without using any merit function that would require a troublesome parameter to be provided by the user for weighting the relative merits of improving feasibility and optimality. A filter algorithm introduces a function that aggregates constraint violations and then treats the resulting biobjective problem. A step is accepted if it reduces the value either of the objective function or of the constraint violation. Although this clearly is less parameter dependent than a penalty function, or an augmented Lagrangian, still we acknowledge that specifying a constraint violation function implies assigning relative weights to reducing each constraint. The algorithm maintains feasibility with respect to X by modifying the aggregate constraint violation for Ω to $+\infty$ outside of X.

Fletcher et al. [15, 16] show convergence of the filter method that uses SQP or SLP to suggest steps. Thus, previous filter algorithms require explicit use of the derivatives of both the objective and the constraints. They also require more than a simple decrease of the objective and constraint violation functions to accept a step. Numerical results for their filter methods are very promising.

The generalized pattern search (GPS) algorithm class designed by Torczon [28] unifies a wide class of useful derivative-free algorithms for unconstrained optimization. Lewis and Torczon extended the GPS framework to bound constrained optimization [21] and, more generally [23], for problems with a finite number of linear constraints. Audet and Dennis [2] allow extended valued functions, which arise often in practice (see, e.g., [4, 5]), and provide an analysis that, among other things, identifies a specific set of limit points allowing the application of Clarke's [8] generalized derivatives under local Lipschitz continuity to unify, strengthen, and simplify the unconstrained and simply constrained Lewis–Torczon theory. In our opinion, a significant feature of our nonsmooth analysis is its ability to highlight the dependence of GPS even with simple constraints on the choice of directions.

Under the assumption that f is continuously differentiable, Torczon [28] showed that GPS methods for unconstrained optimization produce some limit point for which the gradient of the objective function is zero, and Lewis and Torczon showed that their adaptations produce a Karush–Kuhn–Tucker (KKT) point for bound constrained [21] and linearly constrained [23] problems. These adaptations require that the set of directions given to the GPS method include the tangent cone generators of the feasible region at every feasible point.

We identified specific subsequences of trial points in [2]. These "refining subsequences" will be discussed in section 5.3 and, even without any assumptions on the smoothness of f, limit points of these subsequences are shown to exist under standard assumptions. It is also shown that the following intermediate results hold: If it turns out that f is Lipschitz near a strictly feasible limit point, then Clarke's derivatives are nonnegative on a set of positive spanning directions. A positive spanning set is a set of directions in $\mathbb{R}^n$ whose nonnegative linear combinations span the whole space $\mathbb{R}^n$. Moreover, if f is strictly differentiable (defined in section 5.2) at that strictly feasible point, then the gradient is guaranteed to be zero. Similar results are shown when the limit point is on the boundary of the linearly constrained domain—again under the assumption that the tangent cone generators at every point are available to the GPS method.

Assuming that the functions f and C are twice continuously differentiable and that the constraint Jacobian has full rank, Lewis and Torczon [24] propose and analyze a derivative-free procedure to handle general constraints. In their procedure,

GPS for bound constraints is used to carry out the specified inexact minimizations of the sequence of augmented Lagrangian subproblems formulated by Conn, Gould, and Toint [9]. Our algorithm is an alternative to their method, for the cases when one would prefer not to assume continuous second derivatives, or when one wishes to directly engage the original problem and avoid estimating penalty parameters and Lagrange multipliers, but is willing to settle for weaker optimality conditions. Thus, we do not claim that our method is to be preferred for every problem, but our algorithm does have the advantage that it reduces to the GPS method for linear constraints when all the constraints are known to be linear. In fact, we also allow infeasible starting points, but that is a minor point for linear constraints.

One of our objectives is to construct an algorithm that can be used on applications in which the objective and constraints are not given analytically but as "black boxes." For such applications, a value $x \in \mathbb{R}^n$ will be used as an input to a nontrivial simulation to evaluate $f(x)$ and $C(x)$. The subroutine call may fail to return a value a significant percentage of times it is invoked [4, 5, 6, 7, 18] and, even when it succeeds, several factors (e.g., noise, numerical instability, modeling inaccuracy, etc.) may mean that one cannot construct accurate approximate derivatives. Under these circumstances, i.e., the structure of the functions is unknown to the optimizer, we will settle for weakened local optimality convergence results. The pattern search filter algorithm presented here has the following features:

- It is completely derivative free. It neither uses or approximates any derivative information nor attempts to linearize any constraint.
- It transparently generalizes GPS for unconstrained problems and for bound or linear constraints when they are treated, as in [2, 21, 23], by rejecting points infeasible for those constraints, and by selecting polling directions that conform to the boundary of the domain (see Definition 4.1).
- It uses a step acceptance rule based on filter methods, so there is no need for any penalty constants or Lagrange multiplier estimates.
- It makes assumptions on the problem functions f and C that conform to the practical instances that interest us most [4, 3, 5]. They may be discontinuous and even take on infinite values. Therefore, no global smoothness assumption is justified; however, the strength of the optimality conditions guaranteed by the algorithm depends on local smoothness of the functions at the limit point and, most strongly, on properties of the GPS directions at that point.
- It preserves the desirable GPS property of requiring only simple decrease, which is expressed in the present context with respect to the objective and constraint violation functions.
- It does not require any constraint qualifications on the nonlinear constraints.

The Boeing Design Explorer software uses the GPS filter algorithm given here as a meta algorithm in the surrogate management framework for general nonlinear programming [3]. We will give representative results for Design Explorer applied to such an industrial design problem, and so we will mention specific implementation along the way to show how certain aspects of the algorithm have been implemented successfully.

A key issue for implementations like Design Explorer, the one presented in [25], and our own NOMAD software, is that the division into SEARCH and POLL steps is preserved. This is crucial to the use of the surrogate management framework with or without constraints [4]. This division of steps is not used by all researchers, but we do need it for use with implementations we prefer [4, 5, 25, 3] and for our own NOMAD. Abramson's MatLab 6 implementation NOMADm of a suite of algorithms including

the filter algorithm given here can be accessed online from http://en.afit.edu/ENC/Faculty/MAbramson/abramson.html.

The SEARCH steps we prefer make a global exploration of the variable space, and they might use inexpensive surrogate objective and constraints to predict points that constitute improvements to the real problem. We believe that this is crucial for the application of direct search methods to a large class of engineering design problems because applying these methods directly to the actual problem would be in many cases prohibitively expensive. Needing minutes, hours, or even days to compute a real function value is common.

The POLL step is a local exploration near an identified incumbent filter point, and its properties enable the theory to guarantee convergence. Both the SEARCH and POLL steps are detailed in section 2. Another use for our algorithm, with properly chosen SEARCH steps, is that in which one would rather not find only a nearby local optimizer, but instead is willing to use some function evaluations to explore the domain more thoroughly. For example, the implementation in [25] uses an evolutionary algorithm on the surrogate problem for just this reason. We do not guarantee a global optimum; after all, it is important to keep in mind that global optimization of black box functions is impossible, even to the point that if one had the global optimum, one could not be certain of it [27].

The paper is organized as follows. Sections 2 and 3 give brief descriptions of pattern search and filter algorithms. In section 4, we present and begin the analysis of a new algorithm that combines their features. Specifically, without any smoothness assumptions on the problem, we show the existence of some promising limit points. Our optimality results rely on Clarke's calculus with respect to both the constraint violation and objective functions, and on the notion of contingent cones, and so we provide the necessary background in section 5. Section 6 shows that if the constraint violation or objective function is locally smooth at such a limit point, then some first order optimality conditions are satisfied. In the absence of general constraints, the convergence results reduce to those presented in [2]. In fact, Proposition 6.7 and the remark following it show that the results here generalize to nonlinear constraints the results for linear constraints if the choice of GPS directions satisfies conditions easily enforced for the linear case. Finally, in section 7, we make important points through three examples. First, we show the value of a filter method as opposed to a "barrier" method, which rejects trial points that violate the linear constraints [23, 2]. Second, we show the advantages for our algorithm of a squared ℓ_2 over an ℓ_1 measure of constraint violations. Third, we show that our main convergence result concerning the objective function cannot be improved without removing some of the flexibility of the algorithm or adding more assumptions about the problem, including knowledge of the geometry of the constraints at a limit point. We conclude this section by applying our method to a real engineering problem. These examples show the strength and limitations of our approach. The paper concludes with a discussion of the significance of the convergence results, especially as they relate to the dependence of GPS algorithms on the finite set of polling directions to which they are restricted.

2. Pattern search algorithms for unconstrained optimization. The reader is referred to [23, 2] for a thorough description of linearly constrained pattern search algorithms. In the present paper, the same notation as in [2] is used.

2.1. Search and poll steps. Pattern search algorithms for unconstrained minimization generate a sequence of iterates $\{x_k\}$ in $\mathbb{R}^n$ with nonincreasing objective function values. At each iteration, the objective function is evaluated at a finite

number of points on a mesh (a discrete subset of $\mathbb{R}^n$ defined below) to try to find one that yields a lower objective function value than the incumbent. An incumbent solution is a trial point in $\mathbb{R}^n$ where the algorithm evaluated the objective function f and has the lowest value found so far. Any strategy may be used to select mesh points that are to be candidates for the next iteration, provided that only a finite number of points (possibly none) is selected. This is called the SEARCH step. We favor SEARCH procedures like those used in Design Explorer and NOMAD that choose candidate points independent of the incumbent and whose commonality is that their global reach is independent of the mesh size. We feel that such SEARCH procedures are more likely to discover different basins for the function than the one in which the initial point lies.

When the SEARCH fails in finding an improved mesh point, the POLL step must be invoked. In that "fall back" step the function value is evaluated at neighboring mesh points around x_k. If the POLL step also fails in finding an improved mesh point, then x_k is said to be a mesh local optimizer. The mesh is then refined and x_{k+1} is set to x_k. The situation for our constrained version is going to be a bit more complex but is consistent in spirit.

If either the SEARCH or POLL step succeeds in finding an improved mesh point $x_{k+1} \neq x_k$ with a strictly lower objective function value, then the mesh size parameter is kept the same or increased, and the process is reiterated. Indeed, as long as improved mesh points are found, one would likely choose trial points on coarser meshes. With surrogate-based SEARCH steps [4], a great deal of progress can often be made with few function values, and $\mathcal{O}(n)$ function values are needed only when the POLL step detects a mesh local optimizer, which indicates that the mesh needs to be refined. We warn the reader that there is only a cursory discussion of SEARCH strategies in the present paper. The reason is that since the SEARCH is free of any rule, except finiteness and being on the mesh, it cannot be used to enhance the convergence theory. Indeed, some examples in [1] exploit perverse SEARCH strategies to show negative results. However, we are willing to pay this "theoretical" price for the practical reasons given above.

The formal definition of the mesh requires the following. Let D be a finite matrix whose columns in $\mathbb{R}^n$ form a positive spanning set. We use the notation $d \in D$ to indicate that d is a column of the matrix D. It is also required that each column $d \in D$ is the product of a nonsingular generating matrix $G \in \mathbb{R}^{n\times n}$ by some integer vector in $\mathbb{Z}^n$. The same generating matrix G is used for all directions d. See [1, 2] for further insight on this set D, or see [22] for the original equivalent formulation. The set valued function $M(\cdot,\cdot)$ defines the current mesh through the lattices spanned by the columns of D, centered around the current iterate x_k:

$$M(x_k, \Delta_k) = \{x_k + \Delta_k D z : z \in \mathbb{N}^{n_D}\}, \tag{2.1}$$

where $\Delta_k \in \mathbb{R}_+$ is the mesh size parameter, and n_D is the number of columns of the matrix D. Note that in section 4.2, on the filter GPS algorithm we will use a more general definition of the mesh.

When the SEARCH fails in providing an improved mesh point, the objective function must be evaluated at the mesh points that neighbor the current iterate x_k, the current incumbent solution. In the unconstrained case the POLL set is *centered* at x_k, the current iterate. This defines the POLL set $P_k = \{x_k\} \cup \{x_k + \Delta_k d : d \in D_k\}$ for some positive spanning matrix $D_k \subseteq D$. This notation means that the columns of D_k are chosen from those of D. We will refer to evaluating $f(x_k + \Delta_k d)$ as polling in the direction d.

Since *iterate* has little meaning for the filter algorithm presented in section 4, we will poll about a choice of poll centers to be defined later. This will give a different definition for the current mesh.

2.2. Parameter update. At any iteration, there are two possible outcomes, which lead to two sets of rules to update the parameters.

If the iteration fails to produce an improved mesh point, then the POLL step guarantees that x_k is a mesh local optimizer. The mesh is then refined. More precisely,

$$\Delta_{k+1} = \tau^{w_k}\Delta_k \ < \ \Delta_k \tag{2.2}$$

with $0 < \tau^{w_k} < 1$, where $\tau > 1$ is a rational number that remains constant over all iterations, and $w_k \leq -1$ is an integer bounded below by the constant $w^- \leq -1$.

If the iteration produces an improved mesh point, then the mesh size parameter is kept the same or is increased, and the process is reiterated. The coarsening of the mesh follows the rule

$$\Delta_{k+1} = \tau^{w_k}\Delta_k \ \geq \ \Delta_k, \tag{2.3}$$

where $\tau > 1$ is defined above and $w_k \geq 0$ is an integer bounded above by $w^+ \geq 0$. By modifying the mesh size parameters this way, it follows that for any $k \geq 0$, there exists an integer $r_k \in \mathbb{Z}$ such that $\Delta_k = \tau^{r_k}\Delta_0$.

Typical values for the mesh parameter update are $\tau = 2$ and $w_k = -1$ when the poll center is shown to be a local mesh optimizer, and $w_k = 1$ when an improved mesh point is found. This leads to setting $\Delta_{k+1} = \frac{1}{2}\Delta_k$ when the mesh needs to be refined, and $\Delta_{k+1} = 2\Delta_k$ when the mesh is coarsened. An example of the direction matrix might be $D = [I_n \ -I_n]$, where I_n is the $n \times n$ identity matrix. The mesh would then be $M(x_k, \Delta_k) = \{x_k + \Delta_k z : z \in \mathbb{Z}^n\}$ and the POLL set would be $P_k = \{x_k\} \cup \{x_k \pm \Delta_k e_i : i = 1, 2, \ldots, n\}$, where e_i is the ith column of the identity matrix. In the case where D is constructed from all the columns of the set $\{-1, 0, 1\}^n$, the mesh is the same as the previous one, but the POLL set may differ because any set of mesh points, whose directions from the poll center form a positive basis, may be chosen. In $\mathbb{R}^2$ for instance, the POLL set could be $P_k = \{x_k,\ x_k + \Delta_k(1,0)^T,\ x_k + \Delta_k(0,1)^T,\ x_k + \Delta_k(-1,-1)^T\}$.

We borrow Coope and Price's [10] terminology for the following final remark on GPS. These methods are said to be *opportunistic* in the sense that as soon as an improved mesh point is found, the current iteration may stop without completing the function evaluations in the SEARCH and POLL steps.

3. Filter algorithms for constrained optimization. Filter algorithms treat the optimization problem as biobjective: one wishes to minimize both the objective function f and a nonnegative aggregate constraint violation function h. Filter algorithms attempt to minimize both functions but, clearly, priority must be given to h, at least until a feasible iterate is found. This priority appears also in our algorithm in the definition of the poll centers and the poll set. Fletcher et al. [14, 15, 16] do this via *restoration* steps. Another difference in our algorithm is that in keeping with pattern search algorithms for less general problems, we only require improvement in either f or h, while Fletcher et al. have a sufficient decrease condition in the form of an *envelope* over the filter that constitutes a "sufficiently unfiltered" condition.

The terminology used in this paper differs slightly from that used by Fletcher et al. Our notation is more compact for our class of algorithms, and so it simplifies the presentation of our results. In addition, since our plan is to provide a truly multiobjective

GPS algorithm in later work, and since it is likely to involve a version of the filter, it is best to conform to standard terminology in multiobjective optimization [13].

Fletcher et al.'s definition of *dominance* makes it a reflexive relation, which simplifies the definition of a filter, but we will forgo that convenience to adhere to standard multiobjective terminology. The point is that the reader familiar with the filter literature should read this section carefully. We will end up with almost the standard notion of a filter, but we will define it slightly differently using the standard multiobjective notion of dominance: For a pair of vectors w, w', with finite components, w *dominates* w', written $w \prec w'$, if and only if for all i, $w_i \leq w'_i$, and $w \neq w'$. We will use $w \preceq w'$ to indicate that either $w \prec w'$ or $w = w'$, which is the notion of dominance used in earlier filter papers.

The constraint violation function is defined to satisfy the following properties: $h(x) \geq 0$, $h(x) = 0$ if and only if $C(x) \leq 0$, thus $h(x) > 0$ if and only if $C(x) \not\leq 0$, and $h(x) = +\infty$ whenever any component of $C(x)$ is infinite. For example, we could set $h(x) = \|C(x)_+\|$, where $\|\cdot\|$ is a vector norm and where $(C(x)_+)_i$ is set to zero if $c_i(x) \leq 0$ and to $c_i(x)$ otherwise, $i = 1, 2, \ldots, n$. We show in section 6.1 that the more locally smooth h is, the better the algorithm is able to exploit the positive spanning sets used. Our analysis and the examples in sections 5.2 and 7.2 indicate that $h(x) = \|C(x)_+\|_2^2$ is a sound choice.

Recall that the feasible region of the optimization problem (1.1) is defined to be the intersection of a polyhedron X and Ω. Since it is simple to remain feasible with respect to X, we define a second constraint violation function

$$h_X = h + \psi_X, \tag{3.1}$$

where ψ_X is the indicator function for X. It is zero on X and $+\infty$ elsewhere. We will see in the next section that by applying our pattern search filter algorithm to h_X and f, the convergence results with respect to feasibility will depend on local smoothness of h, and not of h_X, which is obviously discontinuous on the boundary of X.

There should be no confusion in defining a special meaning of dominance for the vector arguments of our problem functions h_X, f. This will simplify our terminology rather than use some other symbol such as $\prec_{(h_X,f)}$. Thus, a point $x_k \in \mathbb{R}^n$ is said to *dominate* $x \in \mathbb{R}^n$, $x_k \prec x$ if and only if $(h_X(x_k), f(x_k))^T \prec (h_X(x), f(x))^T$. Two points are equivalent if they generate an identical pair of h_X and f values. As above, $x \preceq x'$ indicates that either $x \prec x'$ or x and x' are equivalent.

A filter $\mathcal{F}$ is a finite set of infeasible points in $\mathbb{R}^n$ such that no pair x, x' in the filter is in the relation $x \prec x'$. A point x' is said to be filtered either if $x' \succeq x$ for some $x \in \mathcal{F}$, or if $h_X(x') \geq h_{max}$ for some positive finite upper bound h_{max} on allowable aggregate infeasibility, or if x' is feasible and $f(x) \geq f^F$ (i.e., the least function value found so far at a feasible point). The point x' is unfiltered otherwise. The set of filtered points $\overline{\mathcal{F}}$ is denoted in standard notation as

$$\overline{\mathcal{F}} = \bigcup_{x \in \mathcal{F}} \{x' : x' \succeq x\} \cup \{x' : h_X(x') \geq h_{max}\} \cup \{x' : h_X(x') = 0,\ f(x') \geq f^F\}. \tag{3.2}$$

For our version of the filter, *iterate* is not a necessary concept. Instead, the role of iterate in the nonfilter versions of GPS is played here by the set of *incumbents* consisting of the best feasible point in $X \cap \Omega$ found so far (if any have been found), and the least infeasible point with the best function value found so far.

Unfiltered points are added to $\mathcal{F}$ as they are generated, and filtered ones are rejected. Whether a point is filtered can depend on when it is generated. This

temporal property causes "blocking entries" [14]. In order to avoid the problem of blocking entries, the filter contains only infeasible points. The incumbent best feasible point is treated separately. The reason for this separation is to encourage moving over an infeasible function ridge and approaching a different part of the feasible region.

4. A pattern search filter algorithm for constrained optimization. In the previous sections we presented the filter framework for general constraints, and the GPS algorithm for unconstrained optimization. We now present a GPS filter method for the optimization problem (1.1). When some of the constraints are known to be linear, i.e., when X is not trivial, it is frequently advantageous to treat them separately from the others and to ask that every point at which f is evaluated belong to X. This is especially true of linear equality or bound constraints. For any derivative-free algorithm, one should surely use linear equality constraints to eliminate variables, and this is desirable for nonlinear equality constraints if it is practical. A reference dealing with various formulations based on eliminating some variables in this way is [4].

4.1. Bound and linear constraints. By applying the algorithm to h_X, defined in (3.1), instead of h, any trial point outside of X is rejected since its constraint violation function value is larger than h_{max}. We called this the "barrier" approach for X. In the absence of general constraints [2], the indicator function is added to f instead of h. In the present work, we cannot add it to f because the trial points $x \in X$ for which $-\infty < f(x) \leq \infty$ and $0 < h(x) < h(x')$ for all $x' \in \mathcal{F}$ are unfiltered.

In addition, the fact that some linear constraints are explicitly known must be used to select mesh directions that take into account the geometry of the region X, just as suggested in [21, 23, 2]: When the poll center is within a given tolerance $\epsilon > 0$ of the boundary of X, then the positive spanning directions D_k that define the poll set are chosen to contain the ones that span the tangent cone $T_X(y)$ to X at all boundary points y within the tolerance ϵ. The formal definition is as follows.

DEFINITION 4.1. *A rule for selecting the positive spanning sets $D_k = D(k, x_k) \subseteq D$* conforms to *$X$ for some $\epsilon > 0$ if, at each iteration k and for each y in the boundary of X for which $\|y - x_k\| < \epsilon$, $T_X(y)$ is generated by nonnegative linear combinations of the columns of a subset D_k^y of D_k.*

These tangent cone directions should be added to D_k before getting too close to the boundary, i.e., it is best not to take the tolerance ϵ too small. The reader will see that a finite set D cannot conform to $\Omega \cap X$ when Ω is defined by nonlinear constraints. Nonetheless, it is interesting that in the case when $\Omega \cap X$ is defined by linear constraints, the above is less than we assume in Proposition 6.7. This is the substance of the remark following that proposition.

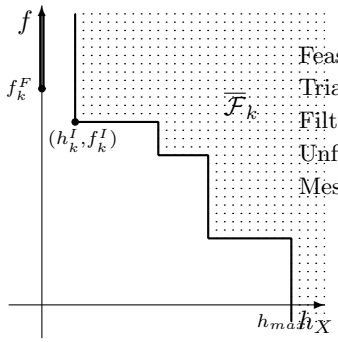
4.2. Meshes and poll centers. In our proposed pattern search filter algorithm, the test for accepting a better mesh point is not based solely on the decrease of the objective function value when there are constraints. Therefore, the terminology *improved mesh point* (used in the unconstrained case) is not suitable in a biobjective context. Instead, we will use the terminology *unfiltered mesh point* when either the SEARCH or POLL step finds a mesh point that is not filtered. If both steps fail in finding an unfiltered mesh point, then we cannot say that the poll center is a *mesh local optimizer* (as in the unconstrained case); instead we will say that the poll center, which will be chosen to be one of two special points that we call incumbents, or chosen to be a point that ties one of them, is a *mesh isolated filter point* since its mesh neighbors (the points in the poll set) are all filtered.

As in the pattern search algorithms presented in section 2, the SEARCH and POLL steps are opportunistic and may be terminated without any more function evaluations when an unfiltered mesh point is found. The mesh size parameter is then either increased or kept constant according to rule (2.3). When no such point is found, the poll center is a mesh isolated filter point and the filter remains unmodified. The mesh size parameter is decreased according to rule (2.2). Unlike Fletcher et al.'s filter algorithms, there is no "envelope" added to the filter to guarantee a form of sufficient decrease.

We define two types of incumbents: the feasible ones, and the infeasible ones with minimal constraint violation. Let f_k^F represent the feasible incumbent value, i.e., the smallest objective function value (for feasible points) found by the algorithm up to iteration k. If no feasible point has been found, f_k^F is set at $+\infty$. Let $h_k^I > 0$ be the least positive constraint violation function value found up to iteration k, and let f_k^I denote the smallest objective function value of the points found whose constraint violation function values are equal to h_k^I. If no such point exists, or if $h_k^I > h_{max}$, then h_k^I is fixed at h_{max} and f_k^I at $-\infty$. The superscript F stands for *feasible* and I for *infeasible*. We denote by S_k the set of points at which f and h were evaluated by the start of iteration k. This notation leads to the definition of incumbent solutions.

DEFINITION 4.2. *At iteration k of a pattern search filter algorithm, any $x \in X \cap \Omega \cap S_k$ such that $f(x) = f_k^F$ is a* feasible incumbent *point, and any $x \in X \cap S_k$ such that $0 < h(x) = h_k^I < h_{max}$ and $f(x) = f_k^I$ is an* infeasible incumbent.

Figure 1 shows an example of a filter and illustrates the incumbent solutions.



Feasible region:	$X \cap \Omega = \{x \in \mathbb{R}^n : h_X(x) = 0\}$	
Trial point:	$t_k \in M(x_k, \Delta_k)$	equation (4.1)
Filtered trial point:	$t_k \in \overline{\mathcal{F}}_k$	
Unfiltered mesh point:	$t_k \notin \overline{\mathcal{F}}_k$	equation (3.2)
Mesh isolated filter point:	$P_k \subset \overline{\mathcal{F}}_k$	equation (4.2)

FIG. 1. *Feasible incumbent set: $\{x \in X \cap \Omega \cap S_k : f(x) = f_k^F\}$. Infeasible incumbent set $\{x \in X \cap S_k : h(x) = h_k^I,\ f(x) = f_k^I\}$. ($S_k$ is the set of points at which f and h were evaluated by the start of iteration k.)*

In the unconstrained case, there was a single type of incumbent solution. Therefore the mesh was necessarily defined around it. We redefine the current mesh so that it contains more points, therefore allowing more flexibility to the algorithm. Recall that the mesh is conceptual in the sense that it is never actually constructed, and its sole purpose is to allow us to capture some structure about the trial points in order to derive some convergence results.

Let $S_0 \subset X$ be the set of initial trial points provided by the user at which the function values are computed. We assume that

- at least one point of S_0 has a constraint violation function value less than h_{max}. This ensures that there is at least one incumbent solution.
- every element of S_0 lies on $M_0(x, \Delta_0)$ (see (2.1)) for some $x \in S_0$.

The mesh is now defined as the following union:

$$M(S_k, \Delta_k) = \bigcup_{x \in S_k} M(x, \Delta_k), \tag{4.1}$$

and $M(x, \Delta_k)$ is defined in (2.1). This new definition allows more flexibility for the user. For example, the SEARCH step is now allowed to explore, in a way similar to the poll step, around any trial point x in S_k, i.e., to evaluate the functions at points from the set $\{x + \Delta_k d : d \in D\}$.

The poll center p_k, the point around which the poll set is constructed, is chosen in the sets of either feasible or infeasible incumbents. Note that when these sets of incumbents are nonempty, each will usually be composed of a single element. Thus, the poll center satisfies either

$$(h_X(p_k), f(p_k)) \;=\; (0, f_k^F) \quad \text{or} \quad (h_X(p_k), f(p_k)) \;=\; (h_k^I, f_k^I). \tag{4.2}$$

The poll set is the poll center p_k together with its mesh neighbors:

$$P_k = \{p_k\} \cup \{p_k + \Delta_k d : d \in D_k\}. \tag{4.3}$$

The positive spanning matrix D_k is composed of columns of D and conforms to the boundary of the linear constraints for an $\epsilon > 0$ (see Definition 4.1) when the poll center is within ϵ of the boundary of X.

Our class of algorithms and their analysis are completely flexible about the choice between these two types of poll centers. The user may supply a strategy to select a poll center. In section 7, we give what we hope are convincing arguments to not always make one choice or the other. Indeed, always choosing an unchanging feasible incumbent poll center makes the filter algorithm essentially reduce to the barrier approach, which is not indicated for constraints where the polling directions do not conform to the feasible region. Still, we prefer to neither prescribe nor proscribe any choice rule for poll centers. This flexibility may seem tedious to the reader, but the user may have a clear preference based on the results of the SEARCH, which may have involved some strategy that makes it unlikely that one or the other choice would be successful. For example, polling on the surrogate function around several filter points in the SEARCH step seems useful in [25], and the results on the surrogate would be likely to influence the choice of poll center.

Even if we already have a feasible incumbent point, we may wish to poll around one of the least infeasible points, which might have a lower objective function value, in order to try to find and explore a different part of the feasible region Ω. Also, this is what allows our filter algorithm to avoid stalling in the Lewis–Torczon [21] example when those linear constraints are treated by the filter. This is illustrated in section 7.1.

4.3. Description of the algorithm. At any iteration, three types of unfiltered mesh points $x_{k+1} \in M(S_k, \Delta_k)$ can be generated by the algorithm. The most useful ones are the unfiltered feasible mesh points. They improve the feasible incumbent value to $f_{k+1}^F = f(x_{k+1}) < f_k^F$. Next are the infeasible ones that improve the infeasible incumbent with minimal constraint violation: $0 < h_{k+1}^I = h_X(x_{k+1}) < h_k^I$ and $f_{k+1}^I = f(x_{k+1})$. Finally, there are the other infeasible ones that add some elements to the filter but leave the incumbents unchanged. In all three cases, the mesh size parameter is updated according to rule (2.3) with, possibly, some different values of w_k. A typical

way to update the mesh size parameter is to double it when a new incumbent solution is found; otherwise keep it constant when only an unfiltered mesh point is found, and cut it in half when the poll center is shown to be a mesh isolated filter point.

To check whether a trial point x is filtered or not, the following strategy is used in order to avoid wasting expensive function evaluations of f and C. First, $\psi_X(x)$ is evaluated by determining if x belongs to X. If not, then $h_X(x) > h_{max}$ and x is filtered, and the evaluation of $f(x)$ and $C(x)$ is avoided. Second, it is possible that partial information on $C(x)$ allows the algorithm to conclude that x is filtered. For example, if $h(x) = \|C(x)_+\|_2^2$, and if it is known that $\sum_{i=1}^{p} |c_i(x)_+|^2 \geq h_{max}$ for some index $p < m$, then the evaluation of $f(x)$ and $c_i(x)$ for $i = p+1, p+2, \ldots, m$ is not necessary. Similar observations hold if $f(x)$ and partial information on $C(x)$ are known, though this situation is more complicated since the value of $f(x)$ alone cannot allow us to conclude that x is filtered without at least partial knowledge of $h(x)$.

When all trial points are filtered, then the poll center p_k is a mesh isolated filter point, and the mesh size parameter is decreased according to rule (2.2). The next poll center p_{k+1} need not be fixed to p_k. These iterations usually require more function evaluations than when an unfiltered mesh point is found. A useful strategy is to poll around both incumbents before decreasing the mesh size parameter. Logically, one can declare that the first POLL step was actually a part of the SEARCH.

Our algorithm for constrained optimization is formally stated in Figure 2. We allow for the fact that in some applications, a set S_0 of initial points may be available from solving similar problems and can be used to seed the filter. Without any loss of generality, we assume that any such points, or at least the undominated ones, are on the initial mesh and have been "filtered" to be consistent with our initialization step in the sense that x_0 will not be filtered by the other seed points. An easy way to assure this would be to take x_0 to be the seed point with the smallest value of h_X, to break ties by taking one with the smallest objective function value, and to make sure that the necessary directions are in D in order that all the initial filter points are on the mesh. Of course, one must ensure that the directions satisfy the conditions of section 2.

In pattern search algorithms, one role of the POLL step is to guarantee convergence. This is why it is rigidly defined through the positive spanning sets $D_k \subset D$. In practice, the largest improvements in the incumbent points are obtained in the SEARCH step (e.g., see [3, 4, 5], where an inexpensive-to-evaluate surrogate of an expensive function is constructed). The SEARCH step is usually the one that drives the iterates away from a local optimum. In a SEARCH implementation, it might be a good idea to try some points that are near points of the filter. Frank [17] made a suggestion, which seems valuable in practice, that SEARCH might include polling the expensive function around the next most feasible filter point, i.e., $x \in \mathcal{F}_k$ with the least value of $h_X(x) > h^I$. The objective here again is to attempt to find and then explore a different part of the feasible region. This is illustrated by the example in section 7.4.

In the next section, we discuss the reduction of the algorithm proposed here in the absence of nonlinear constraints to those given earlier for unconstrained and linearly constrained problems.

4.4. Reduction of the GPS-filter method to linearly constrained optimization. Consider the case where $m = 0$, i.e., when there are no nonlinear constraints. In [2], linear constraints defining X were handled by adding the indicator function to f and, in the present paper, it is added to h. The effect is the same, since in both cases the indicator function simply eliminates from consideration the

- INITIALIZATION: Let $\mathcal{F}_0$ be the filter associated with a set of initial points S_0. Let x_0 be an undominated point of $\mathcal{F}_0$. Fix $\Delta_0 > 0$ and set the iteration counter k to 0.
- DEFINITION OF INCUMBENT POINTS: Define (if possible)
  f_k^F: the least objective function value for all feasible points found so far;
  $h_k^I > 0$: the least positive constraint violation function value found so far;
  f_k^I: the least objective function value of the points found so far whose constraint violation function value is equal to h_k^I.
- SEARCH AND POLL ON CURRENT MESH $M(S_k, \Delta_k)$ (see (4.1)):
  Perform the SEARCH and possibly the POLL steps (or only part of the steps) until an unfiltered trial point x_{k+1} is found, or until it is shown that all trial points are filtered by $\mathcal{F}_k$.
  - SEARCH STEP: Evaluate h_X and f on a set of trial points on the current mesh $M(S_k, \Delta_k)$ (the strategy that gives the set of points is usually provided by the user).
  - POLL STEP: Evaluate h_X and f on the POLL set P_k (see (4.3)) for a poll center p_k that satisfies (4.2).
- PARAMETER UPDATE: Let $S_{k+1} = S_k \cup \{$the set of all trial points visited in the SEARCH and POLL steps$\}$. If the SEARCH or the POLL step produced an unfiltered mesh point not in $\overline{\mathcal{F}}_k$, then update $\Delta_{k+1} \geq \Delta_k$ according to rule (2.3), and go to the next step to update the filter.
  Otherwise, update $\Delta_{k+1} < \Delta_k$ according to rule (2.2), and set $\mathcal{F}_{k+1} = \mathcal{F}_k$; increase $k \leftarrow k + 1$ and go back to the definition of the incumbents.
- FILTER UPDATE: Let $\mathcal{F}_{k+1}$ be the union of $\mathcal{F}_k$ with all infeasible unfiltered points (with respect to $\mathcal{F}_k$) found during the SEARCH and POLL steps. Remove dominated points from $\mathcal{F}_{k+1}$. Increase $k \leftarrow k + 1$ and return to the definition of the incumbents.

FIG. 2. *A pattern search filter algorithm.*

infeasible points with respect to X. In both cases, the convergence results are relative to the smoothness of h and f and not of h_X and f_X.

The main result of [2] was to identify a convergent subsequence of poll centers (called a refining subsequence) such that the Clarke derivatives at the limit point $\hat{x}$ are nonnegative in all the unsuccessful polling directions used infinitely often in the subsequence. The analysis below generalizes this by identifying a large set of directions for which Clarke derivatives are nonnegative.

4.5. Infinite refinement of the mesh. In this section, we identify a set of limit points of poll centers. Each such limit point satisfies optimality conditions whose strength depends on the smoothness of the problem and the choice of directions.

The convergence analysis of our algorithm is based on the standard (see [9, 14, 15, 16]) assumption that all trial points produced by the algorithm lie in a compact set. A consequence of this is that since the mesh size parameter does not decrease when an unfiltered mesh point is found ($\Delta_{k+1} \geq \Delta_k$), then it follows that only finitely many *consecutive* unfiltered mesh points can be generated.

We will be mainly concerned with the poll centers p_k that are mesh isolated filter points (i.e., the mesh neighbors of p_k are filtered) and for which the mesh size parameter is reduced ($\Delta_{k+1} < \Delta_k$). The proofs of the results in this subsection

are omitted, even if the definition of the mesh is slightly different. The key element required is not the mesh but the fact that any mesh point $x \in M(S_k, \Delta_k)$ can be written as $x + \sum_{i=0}^{k} \Delta_i D z_i$ for some $x \in S_0$ and $z_i \in \mathbb{N}^{n_D}$ for $i = 0, 1, \ldots, k$.

Our first result is that there is a subsequence of iterations for which the mesh size parameter goes to zero. In order to prove it we require the following lemma from Torczon [28] or Audet and Dennis [2]. We omit the proof because it simply involves incorporating our definition (4.1) of the mesh into the same proof.

LEMMA 4.3. *The mesh size parameters Δ_k are bounded above by a positive constant independent of the iteration number k.*

Combining this lemma with the assumption that all incumbents lie in a compact set implies the following result. Its proof is omitted since it is identical to that of the same result in [2]. The original proof of this, using slightly different notation, can be found in Torczon [28].

LEMMA 4.4. *The mesh size parameters satisfy* $\liminf_{k \to +\infty} \Delta_k = 0$.

Coope and Price [10] analyze mesh-based algorithms for the unconstrained and linearly constrained problems in which, instead of requiring that the SEARCH be performed on the mesh, they assume that the limit inferior of the mesh size parameter goes to zero. This shifts the burden from the algorithm specification to the implementation.

Since the mesh size parameter shrinks only at mesh isolated filter points, Lemma 4.4 guarantees that there are infinitely many iterations for which the poll centers are mesh isolated filter points. Thus by compactness, the mesh isolated filter points have limit points. Moreover, all these limit points belong to the polyhedron X. At such an iteration the entire trial set, and in particular the poll set P_k, is filtered. Therefore, for each direction $d \in D_k$ either $h_X(p_k + \Delta_k d) \geq h_{max}$ or there exists some element x in the filter $\mathcal{F}_k$ such that both $f(p_k + \Delta_k d) \geq f(x)$ and $h_X(p_k + \Delta_k d) \geq h_X(x)$ or $h_X(p_k + \Delta_k d) = 0$ and $f_X(p_k + \Delta_k d) \geq f_k^F$.

5. Background for optimality results. As in [2], Clarke's [8] generalized derivatives are the key to our convergence analysis. To use this powerful tool, we analyze the case where the function is Lipschitz in a neighborhood of the limit point in question. Of course, there are some optimization problems on which we would apply our algorithm where the functions are not Lipschitz, or optimization problems where we cannot show that the functions are Lipschitz. But this is beside the point. We show how the algorithm behaves on problems with Lipschitz functions. Another ingredient needed for optimality conditions is the contingent cone, which generalizes the notion of tangent cone to more general constraints. The following material is adapted from [19, 26].

DEFINITION 5.1. *Let $S \subset \mathbb{R}^n$ be nonempty. The cone generated by S is*

$$\mathit{cone}(S) \;=\; \{\lambda s : \lambda \geq 0 \text{ and } s \in S\}.$$

A tangent vector to S at x in the closure of S is $v \in \mathbb{R}^n$ such that there exists a sequence $\{y_k\}$ of elements of S that converges to x and a sequence of positive real numbers $\{\lambda_k\}$ for which $v = \lim_k \lambda_k(y_k - x)$. The set $T(S, x)$ of all tangent vectors to S at x is called the contingent cone (or sequential Bouligand tangent cone) to S at x. The polar cone of a cone $K \subset \mathbb{R}^n$ is $K^\circ = \{x \in \mathbb{R}^n : x^T v \leq 0 \text{ for all } v \in K\}$.

For X, the contingent cone is the same as the tangent cone. The normal cone, used to define KKT points, is less useful here than the polar cone since the normal cone in our context may have little to do with optimality given its usual definition

as the convex conic hull of the gradients of the constraints. The polar cone of the contingent cone is more useful in this context.

Optimality conditions for a differentiable function can be stated in terms of the cone generated by the convex hull of a set S, i.e., the set of nonnegative linear combinations of elements of S. We will use the standard notation $co(S)$ for the convex hull of S but, rather than use the induced but somewhat unwieldy notation $cone(co(S))$, we will use the notation $cc(S)$ for the convex conic hull of S. Thus, for example, to say that a set S is a positive spanning set is to say that $cc(S) = \mathbb{R}^n$.

DEFINITION 5.2 (see [8]). *Let $g : \mathbb{R}^n \to \mathbb{R}$ be Lipschitz near $\bar{x} \in \mathbb{R}^n$. Clarke's generalized derivative at $\bar{x}$ in the direction $v \in \mathbb{R}^n$ is*

$$g^\circ(\bar{x}; v) := \limsup_{y \to \bar{x},\ t \downarrow 0} \frac{g(y + tv) - g(y)}{t}.$$

The generalized gradient of g at $\bar{x}$ is the set

$$\partial g(\bar{x}) := \{s \in \mathbb{R}^n : g^\circ(\bar{x}; v) \geq v^T s \text{ for all } v \in \mathbb{R}^n\}.$$

The generalized derivative may be obtained from the generalized gradient as follows: $g^\circ(\bar{x}; v) = \max\{v^T s : s \in \partial g(\bar{x})\}$.

The following alternate definition of directional derivative will be useful.

LEMMA 5.3. *Let $g : \mathbb{R}^n \to \mathbb{R}$ be Lipschitz near $\bar{x} \in \mathbb{R}^n$. Then,*

$$g^\circ(\bar{x}; v) = \limsup_{y \to \bar{x},\ w \to v,\ t \downarrow 0} \frac{g(y + tw) - g(y)}{t}.$$

Proof. Let L be a Lipschitz constant for g near $\bar{x}$. Then

$$\begin{aligned}\limsup_{y \to \bar{x},\ w \to v,\ t \downarrow 0} \frac{g(y + tw) - g(y)}{t} &= \limsup_{y \to \bar{x},\ w \to v,\ t \downarrow 0} \frac{g(y + tv) - g(y)}{t} + \frac{g(y + tw) - g(y + tv)}{t} \\ &\leq \limsup_{y \to \bar{x},\ w \to v,\ t \downarrow 0} \frac{g(y + tv) - g(y)}{t} + L\|w - v\| = g^\circ(\bar{x}; v).\end{aligned}$$

On the other hand, setting $w = v$ gives a lower bound on the limit supremum:

$$\limsup_{y \to \bar{x},\ w \to v,\ t \downarrow 0} \frac{g(y + tw) - g(y)}{t} \geq \limsup_{y \to \bar{x},\ t \downarrow 0} \frac{g(y + tv) - g(y)}{t} = g^\circ(\bar{x}; v). \qquad \square$$

In order to show that the Clarke generalized directional derivative is nonnegative at a point $\bar{x} \in \mathbb{R}^n$ in the direction $v \in \mathbb{R}^n$, it suffices to generate three subsequences: $\{y_k\}$ converging to $\bar{x}$, $\{w_k\}$ converging to v, and $\{t_k\}$ converging to zero from above in such a way that $g(y_k) \leq g(y_k + t_k w_k)$ for infinitely many k's.

5.1. Clarke's derivatives at limit points. In this subsection, we develop some results about the directions in which the Clarke derivatives indicate optimality. To save space, we prove our preliminary results for a function g, which can be either h or f. Also, for any $\bar{x} \in \mathbb{R}^n$, we define $\Gamma_g(\bar{x})$ to be the closure of $\{x \in \mathbb{R}^n : g(x) \geq g(\bar{x})\}$.

PROPOSITION 5.4. *Let $S \subset \mathbb{R}^n$ be nonempty, let g be defined on an open superset of S, and let g be Lipschitz near $\bar{x} \in S$. Necessary conditions for $\bar{x}$ to be a local minimizer of g on S are as follows:*

- *$g^\circ(\bar{x}; v) \geq 0$ for every $v \in T(S, \bar{x})$.*

- *If g has a Fréchet derivative $\nabla g(\bar{x})$ at $\bar{x}$, then $\nabla g(\bar{x})^T v \geq 0$ for every $v \in co(T(S,\bar{x}))$, and so $-\nabla g(\bar{x}) \in co(T(S,\bar{x}))^\circ$. Thus, if $T(S,\bar{x})$ contains a positive spanning set, then $co(T(S,\bar{x}))^\circ = \{0\}$ and $\nabla g(\hat{x}) = 0$.*

Proof. Let S, g, and $\bar{x}$ be as in the statement, and let v be in $T(S;\bar{x})$. Then, there exists a sequence $\{x_k\}$ of elements of S converging to the local minimizer $\bar{x}$ of g on S, and there exists some positive sequence $\{\lambda_k\}$ such that $v = \lim_k \lambda_k(x_k - \bar{x})$. If $v = 0$, the result is trivial. If $v \neq 0$, then $\lim_k \frac{1}{\lambda_k} = 0$.

Now, take $y_k = \bar{x}$, $w_k = \lambda_k(x_k - \bar{x})$, and $t_k = \frac{1}{\lambda_k}$; we see that

$$g^\circ(\bar{x};v) \;\geq\; \limsup_k \frac{g(y_k + t_k w_k) - g(y_k)}{t_k} = \limsup_k \lambda_k[g(x_k) - g(\bar{x})].$$

But, since $x_k \in S$, $\{x_k\}$ converges to $\bar{x}$, and $\bar{x}$ is a local minimizer of g on S, we have that for sufficiently large k, $\lambda_k[g(x_k) - g(\bar{x})]$ is nonnegative and the first result follows.

Now assume that $\nabla g(\bar{x})$ is the Fréchet derivative at $\bar{x}$. Then by Theorem 4.14 of [19], $\nabla g(\bar{x})^T v \geq 0$ for every $v \in T(S,x)$. Let $v \in co(T(S,\bar{x}))$. Then, there is a nonnegative coefficient vector α such that $v = \sum_i \alpha_i s_i$ for some $s_i \in T(S,\bar{x})$. The second result follows from the linearity of the inner product and the definition of polar cone. If $T(S,\bar{x})$ contains a positive spanning set, then $co(T(S,\bar{x})) = \mathbb{R}^n$ and, therefore, for every $v \neq 0$, we have that $v, -v \in T(S,\bar{x})$, and so $\nabla g(\bar{x})^T v \geq 0$ and $\nabla g(\bar{x})^T(-v) \geq 0$, which completes the proof. ∎

The approach we now give for generating directions in which Clarke derivatives are nonnegative generalizes the one presented in [2]. Indeed, the following result will be useful in enlarging the set of directions and, in addition, it relates the generalized directional derivative to the class of iterative methods that require a decrease in some merit function at each iteration. We prove this more general result first.

LEMMA 5.5. *Let g be Lipschitz near the limit $\bar{x}$ of a sequence $\{y_k\}$ for which the corresponding values $g(y_k)$ are monotone nonincreasing and for which $y_k \neq \bar{x}$ for all k. If v is any limit point of the sequence $\{\frac{y_k - \bar{x}}{\|y_k - \bar{x}\|}\}$, then $v \in T(\Gamma_g(\bar{x}),\bar{x})$ and $g^\circ(\bar{x};v) \geq 0$.*

Proof. Let $\{y_k\}$ and $\bar{x}$ be as in the above statement. There is at least one limit point v of $\{\frac{y_k - \bar{x}}{\|y_k - \bar{x}\|}\}$ since the unit ball is compact. Setting $\lambda_k = \frac{1}{\|y_k - \bar{x}\|}$ in Definition 5.1 yields trivially that $v \in T(\Gamma_g(\bar{x}),\bar{x})$. Moreover, Lemma 5.3 implies

$$\begin{aligned} g^\circ(\bar{x};v) &= \limsup_{y \to \bar{x},\; w \to v,\; t \downarrow 0} \frac{g(y + tw) - g(y)}{t} \\ &\geq \limsup_{k \to \infty} \frac{g\left(\bar{x} + \|y_k - \bar{x}\| \frac{y_k - \bar{x}}{\|y_k - \bar{x}\|}\right) - g(\bar{x})}{\|y_k - \bar{x}\|} \;=\; \limsup_{k \to \infty} \frac{g(y_k) - g(\bar{x})}{\|y_k - \bar{x}\|} \geq 0. \end{aligned}$$ ∎

5.2. Choice of the constraint violation norm. The ℓ_2 constraint violation function $h_2(x) = \|C(x)_+\|_2^2$ will give our best results since it is continuously differentiable whenever C is (see Dennis, El-Alem, and Williamson [11] for a compact formulation of ∇h_2). The constraint violation function $h_1(x) = \|C(x)_+\|_1$ is another common choice, at least for SQP. Thus, the question arises as to the differentiability of h_1. The answer, which implies that it is rarely strictly differentiable at $\hat{x}$, is given by the following result. Recall that a function g is said to be *strictly differentiable* [20, 8] at x if $\lim_{y \to \hat{x},\, t \downarrow 0} \frac{g(y+tv) - g(y)}{t} = \nabla g(\hat{x})^T v$ for all $v \in \mathbb{R}^n$, and g is said to be *regular* [8] at x if for all $v \in \mathbb{R}^n$, the one-sided directional derivative $g'(x,v)$ in the

direction v exists and coincides with $g^\circ(x;v)$. In section 7.2 we will see an example showing the cost of this lack of smoothness.

PROPOSITION 5.6. *If C is regular at every x, then so is h_1. Let $I(x) = \{i : c_i(x) > 0\}$ and $A(x) = \{i : c_i(x) = 0\}$ be the inactive and active set at x, respectively. Then the generalized gradients are related by*

$$\partial h_1(x) \;=\; \sum_{i\in I(x)} \partial c_i(x) + \left\{ \sum_{i\in A(x)} \gamma_i\zeta_i : \gamma_i \in [0,1],\, \zeta_i \in \partial c_i(x),\, i \in A(x) \right\}.$$

The generalized directional derivatives of h_1 and C in a direction v at x are related by

$$h_1^\circ(x;v) \;=\; \sum_{i\in I(x)} c_i^\circ(x;v) + \sum_{i\in A(x)} (c_i^\circ(x;v))_+.$$

Thus, if C is strictly differentiable at x, then

$$h_1^\circ(x;v) \;=\; \sum_{i\in I(x)} \nabla c_i(x)^T v + \sum_{i\in A(x)} (\nabla c_i(x)^T v)_+.$$

Proof. The proof follows from various results in [8] and from some simple observations. Clarke's Propositions 2.3.12 and 2.3.6 guarantee that both $c_i(x)_+$ and $h_1(x)$ are regular at x whenever $c_i(x)$ is.

The third corollary to Clarke's Proposition 2.3.3 implies that $\partial h_1(x) = \sum_i \partial c_i(x)_+$, where this means all possible sums of an element from each $\partial c_i(x)_+$. Clarke's Proposition 2.3.12 implies that

$$\partial c_i(x)_+ \;=\; \begin{cases} \partial c_i(x) & \text{if } c_i(x) > 0, \\ co\{\partial c_i(x), \partial 0(x)\} = \{\gamma_i\zeta_i : \gamma_i \in [0,1],\, \zeta_i \in \partial c_i(x)\} & \text{if } c_i(x) = 0, \\ \partial 0(x) = \{0\} & \text{if } c_i(x) < 0. \end{cases}$$

The generalized directional derivative in any direction v can be written as $h_1^\circ(x;v) = \sum_i (c_i(x;v)_+)^\circ$. If $c_i(x) > 0$, then $(c_i(x;v)_+)^\circ = \max\{v^T\zeta : \zeta \in \partial c_i(x)\} = c_i^\circ(x;v)$. If $c_i(x) < 0$, then $(c_i(x;v)_+)^\circ = \max\{v^T\zeta : \zeta \in \partial 0(x)\} = 0$. Finally, if $c_i(x) = 0$, then

$$\begin{aligned} (c_i(x;v)_+)^\circ &= \max\{v^T\eta : \eta \in \partial c_i(x)_+\} \\ &= \max\{v^T\eta : \eta \in \{\gamma_i\zeta_i : \gamma_i \in [0,1],\, \zeta_i \in \partial c_i(x)\}\} \\ &= \begin{cases} 0 & \text{if } \max\{v^T\zeta_i \in \partial c_i(x)\} \le 0, \\ \max\{v^T\zeta_i \in \partial c_i(x)\} & \text{otherwise} \end{cases} \\ &= (\max\{v^T\zeta_i \in \partial c_i(x)\})_+ \quad = \quad (c_i^\circ(x;v))_+. \end{aligned}$$

The last part of the result follows by definition of strict differentiability. ∎

Note that the above result could be slightly rewritten by using one-sided directional derivatives instead of generalized directional derivatives. Indeed, if C is regular at x, then $c_i^\circ(x;v)$ coincides with the one-sided directional derivative $c_i'(x;v)$, and $h_1^\circ(x;v)$ coincides with $h_1'(x;v)$.

5.3. Refining sequences. The purpose of the following definition is to identify a limit of trial points and as many directions as possible for which Clarke's derivatives are nonnegative at that limit. We will make the convention, which is implied by the algorithm, that p_k being an *active poll center* implies that polling around p_k was at

least initiated at iteration k, although function values at all the corresponding poll set may not have been computed because polling is allowed to stop if some poll step yields an improved mesh point.

DEFINITION 5.7. *A convergent subsequence of active poll centers $\{p_k\}_{k\in K}$ (for some subset of indices K) is said to be a* refining subsequence *if $\lim_{k\in K}\Delta_k = 0$. The set of* refining directions for g *associated with a refining subsequence $\{p_k\}_{k\in K}$ is*

$$R_g(K) = \{v \in \mathbb{R}^n : v = \zeta - \xi \neq 0 \text{ and } -\infty < g(p_k + \Delta_k\xi) \leq g(p_k + \Delta_k\zeta) < \infty \\ \text{and } p_k + \Delta_k\zeta, p_k + \Delta_k\xi \in V_k \text{ for infinitely many } k \in K\},$$

where $V_k \subset P_k$ are the members of the poll set visited by GPS. The set of limit directions *for g associated with the limit $\hat{x}$ of a refining subsequence $\{p_k\}_{k\in K}$ is*

$$L_g(K) = \left\{ v \in \mathbb{R}^n : \exists \{y_k\}_{k\in K} \subset V_k \setminus \{\hat{x}\} \text{ such that } \lim_{k\in K} y_k = \hat{x}, \text{ and } g(y_k) \geq g(y_{k'}) \right. \\ \left. \forall k' > k \in K, \text{ and } v \text{ is a limit point of } \left\{ \frac{y_k - \hat{x}}{\|y_k - \hat{x}\|} \right\}_{k\in K} \right\}.$$

Figure 3 illustrates an example of refining directions for a refining subsequence p_{k_i}: The subsequence $\{p_k\}_{k\in K}$ converges to $\hat{x}$ and has six associated directions, represented by vectors, and four limit directions, represented by dotted lines. Note that in the above definition, if $\xi = 0$, then the refining direction $v = \zeta - \xi$ belongs to D_k. Thus, all the directions in infinitely many D_k, where polling was unsuccessful in finding a better point (for g), are refining directions. Also, if the function is constant in the refining direction $v \in R_g(K)$, then $-v$ also will be a refining direction.

We now show the existence of refining subsequences and directions, but because

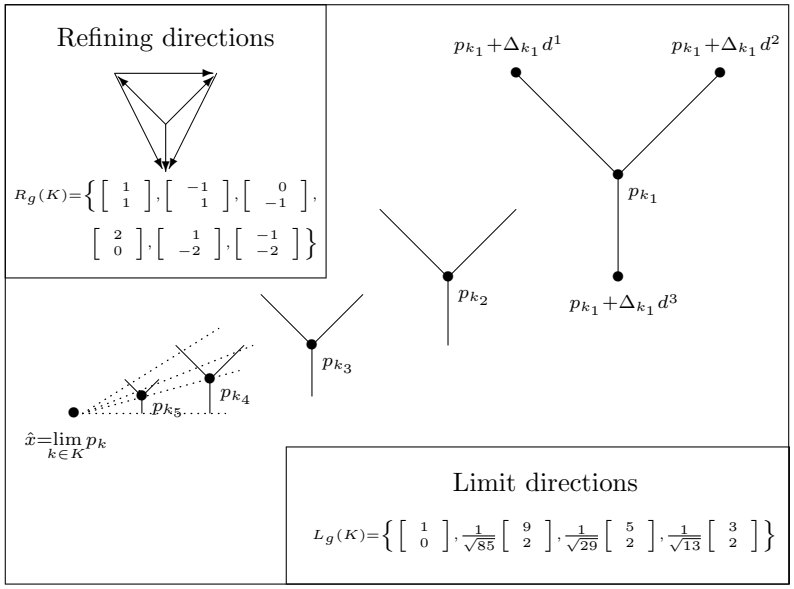


FIG. 3. *An example of refining and limit directions. The six refining directions (represented by vectors) use the fact that $g(p_{k_i} + \Delta_{k_i}d^1) > g(p_{k_i} + \Delta_{k_i}d^2) > g(p_{k_i} + \Delta_{k_i}d^3) > g(p_{k_i})$. The four limit directions are represented by dotted lines.*

the limit directions depend on whether g is f or h, we postpone their existence results until the next section.

LEMMA 5.8. *There exists at least one refining subsequence composed of mesh isolated filter points. Let $\{p_k\}_{k\in K}$ be a refining subsequence. If there exists some $t_k \in V_k$ with $-\infty < g(t_k) < \infty$ for infinitely many k in K, then the set of refining directions $R_g(K)$ is nonempty.*

Proof. Lemma 4.4 guarantees that there exists a subsequence of iterations whose mesh size parameter goes to zero. The mesh size parameter Δ_k decreases only when the POLL step shows that all trial points in P_k are filtered. Moreover, the assumption that all trial points (thus all active poll centers) are in a compact set implies that one such subsequence has a limit point. Thus, there exists a refining subsequence consisting exclusively of mesh isolated poll centers.

Let p_k and t_k be as in the above statement. The second result follows from the fact that the directions $\frac{t_k - p_k}{\Delta_k}$ belong to the finite set D, and therefore there is a direction $d \in D$ used infinitely often. By definition, either d or $-d$ (or both) belongs to $R_g(K)$. ∎

We now show that the generalized directional derivative is nonnegative at limit points of refining subsequences for all associated refining and limit directions for g.

THEOREM 5.9. *Let g be Lipschitz near the limit point $\hat{x}$ of a refining subsequence $\{p_k\}_{k\in K}$. Then g satisfies optimality conditions at $\hat{x}$ on cone$(R_g(K) \cup L_g(K))$ in the sense that if $v \in$ cone$(R_g(K) \cup L_g(K))$, then $g^\circ(\hat{x}; v) \geq 0$. Moreover, $L_g(K) \subset T(\Gamma_g(\hat{x}), \hat{x})$.*

Proof. Let g be Lipschitz near the limit point $\hat{x}$ of a refining subsequence $\{p_k\}_{k\in K}$. If $v = \zeta - \xi \neq 0$ belongs to $R_g(K)$ for some $\zeta, \xi \in \mathbb{R}^n$, then by the definition of the generalized directional derivative we have that

$$g^\circ(\hat{x}; v) \;\geq\; \limsup_{k\in K} \frac{g((p_k + \Delta_k \xi) + \Delta_k d) - g(p_k + \Delta_k \xi)}{\Delta_k} \;\geq\; 0.$$

The same result on cone$(R_g(K) \cup L_g(K))$ follows from the positive homogeneity of Clarke's generalized directional derivative. If v belongs to $L_g(K)$ for some subsequence $\{y_k\}_{k\in K} \subset V_k \setminus \{\hat{x}\}$ converging to $\hat{x}$, then Lemma 5.5 completes the proof. ∎

The previous theorem implies that one of the advantages of using a large number of positive spanning directions in the algorithm is that the set of directions for which Clarke's generalized derivatives are shown to be nonnegative will be larger.

The following corollary strengthens Theorem 5.9 when g is strictly differentiable at the limit point $\hat{x}$. Assuming that g is strictly differentiable at $\hat{x}$, as defined in section 5.2, is equivalent in finite dimensions to assuming that g is Lipschitz near $\hat{x}$, Fréchet differentiable, and regular at $\hat{x}$ [8].

COROLLARY 5.10. *If g is strictly differentiable at $\hat{x}$, then $\nabla g(\hat{x})^T v \geq 0$ for every $v \in cc(R_g(K) \cup L_g(K))$ and thus, if $R_g(K) \cup L_g(K)$ contains a positive spanning set, then $\nabla g(\hat{x}) = 0$.*

Proof. Assume that $\nabla g(\hat{x})$ is the Fréchet derivative at $\hat{x}$. Then Theorem 5.9 ensures that $\nabla g(\hat{x})^T v \geq 0$ for every $v \in R_g(K) \cup L_g(K)$. Let $v \in cc(R_g(K) \cup L_g(K))$. Then, there is a nonnegative coefficient vector α such that $v = \sum_i \alpha_i s_i$ for some $s_i \in R_g(K) \cup L_g(K)$. The first result follows from the linearity of the inner product.

If $R_g(K) \cup L_g(K)$ contains a positive spanning set, then $cc(R_g(K) \cup L_g(K)) = \mathbb{R}^n$ and, therefore, for every $v \neq 0$, we have that $v, -v \in R_g(K) \cup L_g(K)$, and so $\nabla g(\bar{x})^T v \geq 0$ and $\nabla g(\bar{x})^T (-v) \geq 0$, which completes the proof. ∎

6. Optimality conditions for the GPS-filter method. We will continue with results that consider only the behavior of h and then complete our results by analyzing the effect of the filter on the objective function f.

6.1. Results for the constraint violation function. The algorithm definition gives priority to feasibility, as expressed by the constraint violation function, over minimizing the objective function. A consequence of this is that the optimality conditions guaranteed by the algorithm are stronger for h, i.e., achieving feasibility, than they are for achieving constrained optimality. Indeed, in the absence of the assumption of linearly independent constraint gradients, our feasibility results are what one would prove for standard SQP methods—that we obtain a stationary point of the ℓ_2 norm of the constraint violations.

An obvious initial comment is that $h(p_k) = h_X(p_k)$ for every poll center p_k since trial points violating some linear constraints are rejected. Therefore, any limit of poll centers belongs to X. So the analysis can be done in terms of h instead of h_X. Another observation is that, by definition, if $h(\hat{x}) = 0$, then $\hat{x}$ is a global minimizer for h. Furthermore, any limit point of a sequence of feasible points would be feasible if h were lower semicontinuous there or if the feasible region were closed. However, it is possible for a sequence of least infeasible poll centers to converge to an infeasible point. We will therefore concentrate on limit points of infeasible mesh isolated poll centers.

Before presenting results that assume local Lipschitz continuity, we prove the following result, which shows in particular that if any limit point of least infeasible poll centers is feasible and if h is continuous there, then all limit points at which h is lower semicontinuous are also feasible. It also provides a way to identify some limit directions in $L_h(K)$.

THEOREM 6.1. *Let $\{p_k^I\}_{k\in K}$ be a convergent subsequence of least infeasible poll centers. Then $\lim_k h(p_k^I)$ exists and, if h is lower semicontinuous at any limit point $\bar{x}$ of $\{p_k^I\}$, then $\lim_k h(p_k^I) \geq h(\bar{x}) \geq 0$. Every limit point of least infeasible poll centers at which h is continuous has the same constraint violation function value. Furthermore, if h is Lipschitz near any $\bar{x}$, then $h^\circ(\bar{x}; v) \geq 0$ for any limit direction v of $\{\frac{p_k^I - \bar{x}}{\|p_k^I - \bar{x}\|}\}$. In addition, each limit direction satisfies $v \in T(\Gamma_h(\bar{x}), \bar{x})$.*

Proof. The sequence $\{h(p_k^I)\}$ is convergent because it is a nonincreasing sequence of positive numbers. Of course, for any subsequence of $\{p_k^I\}$, the corresponding h values have the same limit. Thus, if h is lower semicontinuous at $\bar{x}$, we know that, for any subsequence $\{p_k\}_{k\in K}$ of the iteration sequence that converges to $\bar{x}$,

$$\lim_k h(p_k^I) \;=\; \lim_{k\in K} h(p_k^I) \;=\; \liminf_{k\in K} h(p_k^I) \;\geq\; h(\bar{x}) \;\geq\; 0.$$

If h is continuous at some limit points of infeasible poll centers, then the same argument shows that all such limit points have the same value of the constraint violation function. Thus, if any such limit point is feasible, they all are feasible.

The rest of the proof follows by noticing that $v \in L_h(K)$ and by applying Theorem 5.9. ∎

The next result guarantees some nonsmooth first order optimality conditions. It shows that Clarke's derivatives for h are nonnegative in a subset of refining directions whose convex conic hull is the tangent cone to X.

PROPOSITION 6.2. *Let $\hat{x}$ be the limit of a refining subsequence composed of mesh isolated filter points $\{p_k\}_{k\in K}$, and assume that the rule for selecting D_k conforms*

to X for an $\epsilon > 0$. If h is Lipschitz near $\hat{x}$, then $h^\circ(\hat{x}; v) \geq 0$ for any direction v in a set of directions $D' \subset R_h(K)$ satisfying $cc(D') = T(X, \hat{x})$.

Proof. Let $\{p_k\}_{k \in K}$, ϵ, and $\hat{x}$ be as in the statement of the result. When $h(\hat{x}) = 0$, $h^\circ(\hat{x}; v) \geq 0$ for any $v \in \mathbb{R}^n$, so assume that $h(\hat{x}) > 0$. Since the rule for selecting D_k conforms to X for an $\epsilon > 0$, there exists a subset of directions D' of D such that $cc(D') = T(X, \hat{x})$, and for any $v \in D'$ and sufficiently large $k \in K$, $p_k + \Delta_k v \in P_k \cup X$ and $h(p_k + \Delta_k v) > 0$. However, since the poll centers are mesh isolated filter points, it follows that $h(p_k + \Delta_k v) \geq h(p_k)$, and therefore v belongs to $R_h(K)$. Theorem 5.9 completes the proof. ∎

A consequence of this result is that if h is strictly differentiable at the limit point of a refining subsequence composed of mesh isolated filter points, then standard first order optimality conditions for h are satisfied.

COROLLARY 6.3. *Let $\hat{x}$ be the limit of a refining subsequence composed of mesh isolated filter points $\{p_k\}_{k \in K}$, and assume that the rule for selecting D_k conforms to X for an $\epsilon > 0$. If h is strictly differentiable at $\hat{x}$, then $\nabla h(\hat{x})^T v \geq 0$ for every v in $T(X, \hat{x})$.*

Proof. The result is a direct consequence of Corollary 5.10 and Proposition 6.2. ∎

6.2. Results for the objective function. We have shown above that the limit point for a refining subsequence generated by the algorithm satisfies local optimality conditions for the constraint violation function. We now derive some results for the objective function. The first result proposes a way to identify some limit directions in $L_f(K)$.

PROPOSITION 6.4. *Let $\{p_k^F\}_{k \in K}$ be a subsequence of feasible poll centers convergent to a point $\bar{x}$. If f is lower semicontinuous at $\bar{x}$, then $\lim_k f(p_k^k)$ exists and is greater than or equal to $f(\bar{x})$. The set of such limit points at which f is continuous all have the same objective function value. Furthermore, if f is Lipschitz near any $\bar{x}$, then any limit direction v of $\{\frac{p_k^F - \bar{x}}{\|p_k^F - \bar{x}\|}\}$ is such that $f^\circ(\bar{x}; v) \geq 0$ and $v \in T(\Omega, \bar{x}) \cap T(\Gamma_f(\bar{x}), \bar{x})$.*

Proof. Let $\{p_k^F\}_{k \in K}$ and $\bar{x}$ be as in the above statement. The subsequence $f(p_k^F)_{k \in K}$ is monotone nonincreasing and bounded below by the finite value $f(\bar{x})$, and therefore it converges.

Since $p_k^F \in \Omega$ for every $k \in K$, it follows by the definition of the contingent cone that $v \in T(\Omega, \bar{x})$. The rest of the proof follows by noticing that $v \in L_f(K)$ and by applying Theorem 5.9. ∎

The next pair of results guarantees some nonsmooth first order optimality conditions related to the cone tangent to X. They are similar to the first order optimality results for h, Proposition 6.2, and Corollary 6.3, except that they require the limit point to be strictly feasible with respect to Ω. Basically, these results show that when the nonlinear constraints are not binding, the use of the filter does not interfere with the linearly constrained results.

PROPOSITION 6.5. *Let $\hat{x}$ be the limit of a refining subsequence composed of mesh isolated filter points $\{p_k\}_{k \in K}$, and assume that the rule for selecting D_k conforms to X for an $\epsilon > 0$. If f is Lipschitz near $\hat{x}$, and if $\hat{x}$ is strictly feasible with respect to Ω, then $f^\circ(\hat{x}; v) \geq 0$ for any direction v in a set of directions $D' \subset R_h(K)$ satisfying $cc(D') = T(X, \hat{x})$.*

Proof. Let $\{p_k\}_{k \in K}$, ϵ, and $\hat{x}$ be as in the statement of the result. Since the rule for selecting D_k conforms to X for an $\epsilon > 0$, then there exists a subset of directions D'

of D such that $cc(D') = T(X, \hat{x})$, and for any $v \in D'$ and sufficiently large $k \in K$, $p_k + \Delta_k v \in P_k \cup X \cup \Omega$. However, since the poll centers are mesh isolated filter points, it follows that $f(p_k + \Delta_k v) \geq f(p_k)$, and therefore v belongs to $R_f(K)$. Theorem 5.9 completes the proof. ∎

The following corollary to this result shows standard first order optimality conditions for f on X under the additional assumption of strict differentiability.

COROLLARY 6.6. *Let $\hat{x}$ be the limit of a refining subsequence composed of mesh isolated filter points $\{p_k\}_{k\in K}$, and assume that the rule for selecting D_k conforms to X for an $\epsilon > 0$. If f has a strict derivative $\nabla f(\hat{x})$ at $\hat{x}$, and if $\hat{x}$ is strictly feasible with respect to Ω, then $\nabla f(\hat{x})^T v \geq 0$ for every v in $T(X, \hat{x})$.*

Proof. The result is a direct consequence of Corollary 5.10 and Proposition 6.5. ∎

Note that since it is assumed in Corollary 6.6 that $\hat{x}$ is feasible with respect to Ω, and since the algorithm reduces to the one in [23, 2] in the absence of general constraints, the proof of the corollary also follows from a result in [2].

Our next result does not assume strict feasibility of the limit point. It is a corollary of Corollary 5.10. It gives conditions for the limit point of a refining sequence to satisfy optimality conditions on problem (1.1). It is that the convex conic hull of the union of the refining and the limit directions contains the contingent cone for the feasible region at $\hat{x}$. It is interesting that this condition can be met without any feasible descent directions in any poll set. In the simple linearly constrained case, this is implied by ensuring that the polling directions conform to the boundary of X but, here, the corresponding assumption that the polling directions for a refining sequence generate the contingent cone for the feasible region at $\hat{x}$ is not as constructive. This result is illustrated on the three examples of section 7.

PROPOSITION 6.7. *Let $\hat{x}$ be the limit point of a refining subsequence $\{p_k\}_{k\in K}$. If f has a strict derivative $\nabla f(\hat{x})$ at $\hat{x}$, then $-\nabla f(\hat{x})$ belongs to the polar C_f° of $C_f = cc(R_f(K) \cup L_f(K))$, and so $\hat{x}$ satisfies the optimality conditions of Corollary* 5.10 *for f on C_f. Moreover, if $T(\Omega \cap X, \hat{x}) \subset C_f$, then $\hat{x}$ satisfies the optimality conditions of Corollary* 5.10 *for f on problem* (1.1).

Proof. Let $\hat{x}$, f, and C_f be as in the above statement. Corollary 5.10 guarantees that $\nabla f(\hat{x})^T v \geq 0$ for any vector $v \in C_f$. The results follow from the definition of polarity: in general, $-\nabla f(\hat{x}) \in C_f^\circ = \{u \in \mathbb{R}^n : u^T v \leq 0 \text{ for all } v \in C_f\}$. If $C_f \supseteq T(\Omega \cap X, \hat{x})$, then $C_f^\circ \subset T^\circ(\Omega \cap X, \hat{x})$, and the proof is complete. ∎

Remark. Notice that under the assumption that the contingent cone generators of the nonlinear constraints binding at $\hat{x}$ belong to the set of refining or limit directions (as will be the case for linear constraints and conforming directions; see Definition 4.1), then the preceding result reduces to the corresponding result from [2, 23]. This is because, in that case, the contingent cone is the tangent cone, and the polar of the contingent cone is the normal cone, so $\hat{x}$ is a KKT point.

By using a filter-based step acceptance criterion, we have overcome a difficulty in applying pattern search algorithms to constrained optimization. Specifically, we have that the objective function descent directions in the positive spanning set D may be infeasible. Lewis and Torczon [21] give an example in which a nonfilter version of the pattern search algorithm stalls (i.e., all subsequent iterates are the same mesh isolated filter point) at a point containing a strictly feasible descent direction.

The following result shows that, under assumptions on the smoothness of the functions but regardless of the choice of positive spanning set, our algorithm will eventually find an unfiltered mesh point, except when $\nabla f(p_k) = 0$. Thus, we will move away from a constrained minimizer, even a global solution, if doing so decreases

either h or f or produces a new least infeasible point. This is an essential ingredient of any method with ambitions of finding more than a single local constrained minimizer.

PROPOSITION 6.8. *If both h and f are strictly differentiable at the poll center p_k, and if $\nabla f(p_k) \neq 0$, then there cannot be infinitely many consecutive iterations in which p_k is a mesh isolated filter point.*

Proof. Let h and f be strictly differentiable at p_k, where $\nabla f(p_k) \neq 0$. Assume that there are infinitely many consecutive iterations in which p_k is a mesh isolated filter point. Let d be a direction used infinitely often in the (constant) subsequence of poll centers such that $\nabla f(p_k)^T d < 0$.

Since the function h is strictly differentiable at p_k, there exists an $\epsilon > 0$ such that one of the following two conditions is satisfied: either $h_X(p_k + \Delta d) \leq h_X(p_k) < h_{max}$ or $h_X(p_k + \Delta d) > h_X(p_k)$ for all $0 < \Delta < \epsilon$.

If the first condition is satisfied, then for $\Delta_k < \epsilon$ the POLL step will find an unfiltered mesh point. This is a contradiction. If the second condition is satisfied, then let $\tilde{h}$ be the smallest value of $\{h_X(x) : h_X(x) > h_X(p_k),\, x \in \mathcal{F}_k\} \cup \{h_{max}\}$, and let $\tilde{f}$ be the corresponding objective function value, i.e., either $\tilde{f} = f(\tilde{x})$ for some vector $\tilde{x} \in \mathcal{F}_k$ that satisfies $h_X(\tilde{x}) = \tilde{h}$, or $\tilde{f} = -\infty$ in the case that $\tilde{h} = h_{max}$. It follows that $\tilde{h} > h_X(p_k)$ and $\tilde{f} < f(p_k)$. Therefore, whenever $\Delta_k < \epsilon$ is small enough, the following inequalities hold: $h_X(p_k) < h_X(p_k + \Delta_k d) < \tilde{h}$ and $\tilde{f} < f(p_k + \Delta_k d) < f(p_k)$; thus the trial mesh point is unfiltered. This is a contradiction. ∎

7. Illustration of our results. We now illustrate the behavior of our algorithm on three test examples and on a real engineering problem. The first test example is due to Lewis and Torczon [21]. Unlike the barrier approach in [21], the filter approach can converge even with a badly chosen positive spanning set.

The second example justifies our choice of the squared ℓ_2 norm over the ℓ_1 norm in the definition of the constraint violation function. The nonsmoothness of the latter may not provide descent on h_1 in some of the poll directions for which h_2 does descend. The example shows that, since h_1 does not allow movement, using it can result in stalling at an infeasible point.

The third example shows the limitations of our results; there is more left to do. This example uses the algorithm's flexibility as a loophole to avoid a desirable outcome. Even with the squared ℓ_2 norm, it is still possible to choose the positive spanning sets, and to be unlucky, in a way that there is a polling direction which is a feasible descent direction for the objective function f from the limit point $\hat{x}$. This does not contradict our results, but it does show their limitations without a suitable SEARCH scheme.

The last example is a wing planform design problem from Boeing for an airplane different from the two airplanes used to generate the results reported in [3].

7.1. Example of Lewis and Torczon. Consider the linear program [21]

$$\begin{aligned} \min_{x=(a,b)^T} \quad & -a - 2b \\ \text{s.t.} \quad & 0 \leq a \leq 1, \\ & b \leq 0. \end{aligned}$$

The optimal solution is $\hat{x} = (1, 0)^T$. Let us apply our algorithm with initial point $x_0 = (0, 0)^T$, initial mesh size parameter $\Delta_0 = 1$, and a single positive spanning matrix $D_k = D$ constructed with the four directions $\pm(1, 1)^T$ and $\pm(1, -1)^T$. We will not use any SEARCH step for this example. It is pointed out in [21] that all iterations of a "barrier" pattern search algorithm that assigns an objective function value of $+\infty$

to infeasible points, but does not take into consideration the geometry of the feasible region, remain at the origin since the polling directions that yield decrease in the objective function are infeasible.

Suppose that the constraints are given as black boxes and that the algorithm is not aware that they are linear. Therefore $X = \mathbb{R}^2$ and $\Omega = \{(a, b) \in \mathbb{R}^2 : 0 \leq a \leq 1, b \leq 0\}$.

One might consider using the unconstrained GPS on an ℓ_1 exact penalty function for this problem. It turns out that for any penalty constant greater than or equal to 3, the algorithm with the same starting data never moves from the origin. The penalty constant must be greater than 2 for the minimizer of the ℓ_1 penalty function to be the solution to the original problem, so the penalty function approach is not useful here.

Our filter algorithm, using the above-mentioned spanning set, converges to the optimal solution. Mesh directions that conform to the boundary of the feasible region cannot be identified. Figure 4 displays the first few iterations. The shaded area is the feasible region. The poll centers are underlined, and the functions values are displayed between brackets: $[h(x), f(x)]$. The points in the poll set are joined to the poll center by dotted lines.

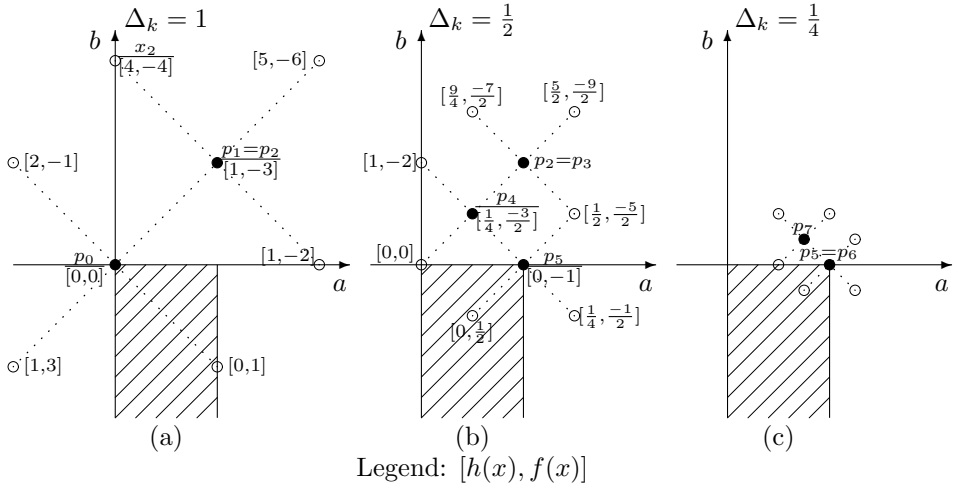


FIG. 4. *First iterations on example from Lewis and Torczon.*

Figure 4(a) illustrates the iterations for which $\Delta_k = 1$. Starting at $x_0 = p_0 = (0, 0)^T$ the algorithm evaluates both functions at $\pm(1, 1)^T$ and $\pm(1, -1)^T$. Only the trial point $(1, -1)^T$ is feasible; it is, however, dominated by x_0. The point $(1, 1)^T$ dominates the two other trial points and is unfiltered.

Let $x_1 = p_1 = (1, 1)^T$. The functions are evaluated at the four points around p_1, and two unfiltered points are found: $x_2 = (0, 2)^T$ and $(2, 2)^T$. Even if an unfiltered mesh point was found, the poll center p_2 would remain at p_1. Polling around p_2 yields filtered points; thus p_2 is a mesh isolated filter point. Figure 5 displays the filters corresponding to the poll centers in Figure 4. Figure 4(b) starts at iteration 3 with $p_3 = (1, 1)^T$ and $\Delta_3 = \frac{1}{2}$. Two consecutive iterations, in which an unfiltered mesh point is found, lead to $p_4 = (\frac{3}{2}, \frac{1}{2})^T$ then $p_5 = (1, 0)^T$, which is the optimal solution.

However, since the gradient is nonzero at this point, Proposition 6.8 ensures that polling around this point will eventually produce an unfiltered mesh point. Indeed, as shown in Figure 4(c), iteration 5 produces a mesh isolated filter point, but iteration 6

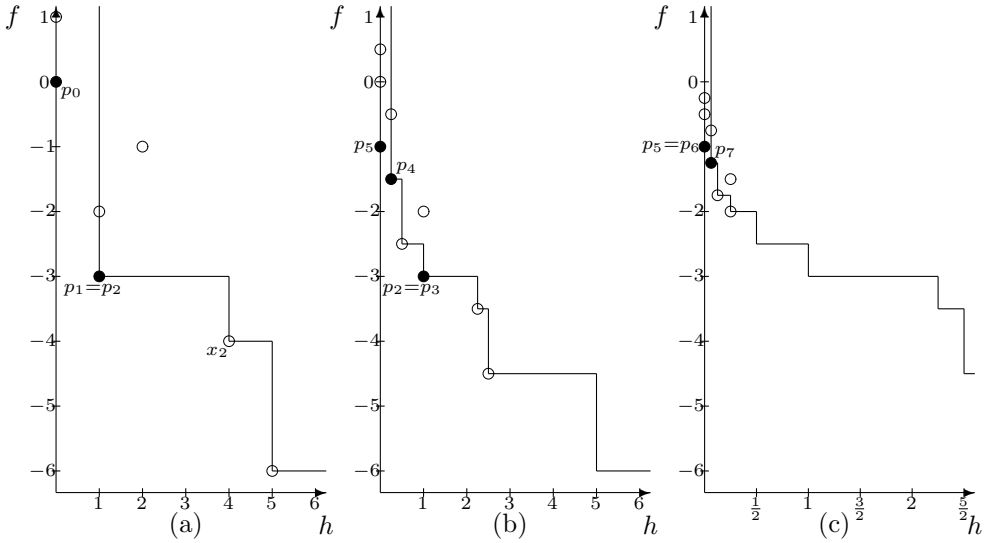


FIG. 5. *Filter for the example from Lewis and Torczon.*

generates an unfiltered mesh point, which is a new infeasible incumbent with a minimal constraint violation. For this example, the limit point is feasible, and so it is a global optimizer for the constraint violation function. The sets of refining and limit directions for f are

$$R_f(K) = \left\{ \begin{bmatrix} -1 \\ -1 \end{bmatrix}, \begin{bmatrix} 1 \\ -1 \end{bmatrix}, \begin{bmatrix} -2 \\ -2 \end{bmatrix}, \begin{bmatrix} 2 \\ -2 \end{bmatrix}, \begin{bmatrix} 0 \\ -2 \end{bmatrix}, \begin{bmatrix} -2 \\ 0 \end{bmatrix} \right\},$$
$$L_f(K) = \left\{ \frac{1}{\sqrt{2}} \begin{bmatrix} -1 \\ -1 \end{bmatrix}, \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ -1 \end{bmatrix} \right\}.$$

The polar of the cone spanned by the refining and limit directions of f is spanned by $\{(1,1)^T, (0,1)^T\}$ and indeed contains $-\nabla f(1,0) = (1,2)^T$ (as stated by Proposition 6.7). Moreover, nonnegative combinations of the last two directions of $R_f(K)$ span the contingent cone to the constraints, and therefore the solution satisfies the KKT conditions.

Remark. Observe that the choice of the poll centers is also important to the quality of the limit points the algorithm finds. Indeed, in this example, if one were to always take the current poll center to be the best feasible incumbent, then the refining sequence will have every term equal to x_0. C_f is again spanned by the same set of directions $R_f(K)$.

Of course, continuing to poll around an unchanging feasible incumbent is a bad idea since it ignores the flexibility of the filter method by reducing it to the barrier method. A better poll center selection strategy could be to alternate between the two incumbents every time the POLL step detects a mesh isolated filter point. Polling around the infeasible one with a minimal constraint violation is especially interesting when f^I is less than f^F since it might move the trial points away from a local optimum or toward a more interesting part of the feasible region. That way, there will be infinitely many poll steps around both types of incumbents. It is also worthwhile to change the positive spanning set to enrich the set of refining directions for both h and f. The flexibility of our theory ensures that such heuristics can be part of a rigorously convergent algorithm.

7.2. Choice of the constraint violation norm. The choice of the norm in the definition of the constraint violation function $h(x) = \|C(x)_+\|$ affects the convergence behavior. The example presented here complements the theoretical results of section 5.2. We prefer the squared ℓ_2 norm over the ℓ_1 norm since it is differentiable whenever the constraint function C is (see [11] for an explicit formulation of the gradient). This means that if there is a descent direction, then a positive spanning set will detect it with the ℓ_2 norm, but ℓ_1 might miss it (see Corollary 6.3). This is illustrated in the following simple linear program:

$$\begin{array}{rl} \min\limits_{x=(a,b)^T} & b \\ \text{s.t.} & -b \le 3a \le b, \\ & b \ge 1 \end{array}$$

with an ℓ_1 constraint violation function $h_1(a,b) = \max(3a-b,0) + \max(-3a-b,0) + \max(1-b,0)$.

Let the algorithm start at the infeasible point $x_0 = p_0 = (0,0)^T$, and let the positive spanning set be $D = \{(1,1)^T, (1,-1)^T, (0,-1)^T\}$. The poll centers and the filter are depicted in Figure 6.

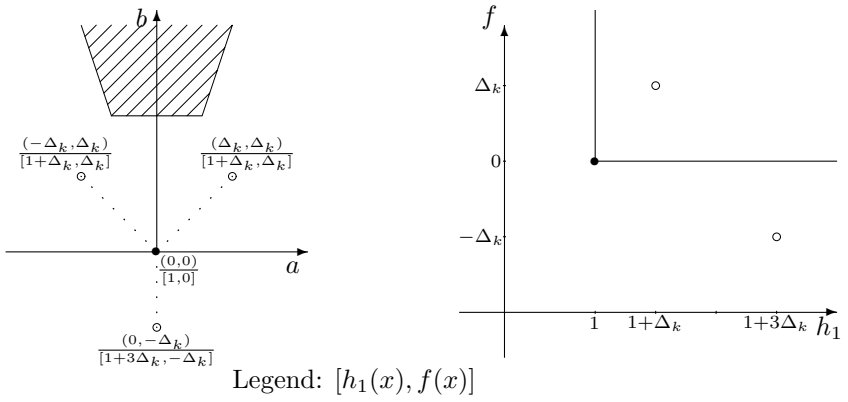


FIG. 6. *The algorithm stalls with the ℓ_1 norm.*

Every even iteration k produces an unfiltered mesh point that does not improve any of the incumbents: The trial point $x_{k+1} = p_k + \Delta_k(1,1)^T$ is unfiltered by $\mathcal{F}_k$. Every odd iteration confirms that the poll center is a mesh isolated filter point: The three trial points are filtered since $p_{k+1} = p_k = (0,0)^T$. Therefore, the mesh size parameter is reduced at each odd iteration.

The sequence of poll centers stalls at the infeasible point $\hat{x} = (0,0)^T$. This means that the nondifferentiability of h_1 hides the descent directions for the constraint violation function. It also means that this is another example in which the unconstrained GPS algorithm applied to the ℓ_1 exact penalty function fails as a solution approach. However, as guaranteed by Theorem 5.9, $h^\circ(\hat{x}, v)$ is nonnegative for the positive spanning directions in D as well as for other refining and limit directions for h:

$$R_h(K) = \left\{ \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \begin{bmatrix} -1 \\ 1 \end{bmatrix}, \begin{bmatrix} 0 \\ -1 \end{bmatrix}, \begin{bmatrix} 1 \\ -2 \end{bmatrix}, \begin{bmatrix} -1 \\ -2 \end{bmatrix}, \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \begin{bmatrix} -1 \\ 0 \end{bmatrix} \right\},$$

$$L_h(K) = \left\{ \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \frac{1}{\sqrt{2}} \begin{bmatrix} -1 \\ 1 \end{bmatrix} \right\}.$$

The sets of refining and limit directions for f are

$$R_f(K) = \left\{ \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \begin{bmatrix} -1 \\ 1 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \begin{bmatrix} -1 \\ 2 \end{bmatrix}, \begin{bmatrix} 1 \\ 2 \end{bmatrix}, \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \begin{bmatrix} -1 \\ 0 \end{bmatrix} \right\},$$
$$L_f(K) = \left\{ \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \frac{1}{\sqrt{2}} \begin{bmatrix} -1 \\ 1 \end{bmatrix} \right\}.$$

The polar of the cone spanned by the refining and limit directions for f reduces to the negative gradient of f: $\{(0,-1)^T\}$.

Remark. If the squared ℓ_2 norm is used for h instead of ℓ_1, then the poll center moves away from $(0,0)$ to $(\Delta_k, \Delta_k)^T$ as soon as the mesh size parameter drops below $\frac{2}{3}$ since $0 < h(\Delta_k, \Delta_k) < 1$ whenever $0 < \Delta_k < \frac{2}{3}$. The sets of refining and limit directions for f are the same as above, but the algorithm converges to a global optimal solution.

7.3. Illustration of the limitation of the results. Consider the problem

$$\begin{aligned} \min_{x=(a,b)^T} \quad & b \\ \text{s.t.} \quad & a(1-a) - b \leq 0. \end{aligned}$$

The algorithmic strategies described below are such that the algorithm goes through infinitely many consecutive cycles of three iterations, and the sequence of poll centers converges to a feasible limit point from which there is a feasible descent direction. This direction is used infinitely often in the POLL step. The first iteration of each cycle improves the feasible incumbent, the second one improves the least infeasible incumbent, and the last one produces a mesh isolated filter point. We admit that the flexibility in the choice of polling directions is exploited to lead to a weak result, but our point is that it can happen.

The trial points generated during cycle ℓ are summarized in Table 1. The algorithm does not perform any SEARCH, and complete polling is always performed. The table also displays the positive spanning directions used at each POLL step. The initial points in cycle $\ell = 1$ are $p_0 = x^I = (\frac{1}{4}, 0)^T$ and $x^F = (0,1)^T$, and the initial mesh size parameter is $\Delta_0 = \frac{1}{8}$.

Figure 7 displays the first cycle (polling around the poll centers p_0, p_1, and p_2) and the corresponding filter. Cycle 1 terminates with a mesh isolated filter point, and cycle 2 is initiated at $p_3 = (\frac{1}{8}, 0)^T$ with $\Delta_3 = \frac{1}{16}$. More generally, cycle ℓ terminates with a mesh isolated filter point. The mesh size parameter is divided by 2, and cycle $\ell + 1$ starts.

All trial points including the sequence of mesh isolated filtered points, i.e., the poll centers corresponding to the third step of each cycle, converge to the feasible point $\hat{x} = (0,0)^T$. There are no other limit points. The results of section 6.1 concerning the constraint violation function are clearly satisfied since the limit point is at the global minimum of h. However, there is a feasible direction from the limit point used infinitely often by the subsequence, which is also a descent direction for the objective function. The sets of refining and limit directions for f are

$$R_f(K) = \left\{ \begin{bmatrix} -2 \\ 1 \end{bmatrix}, \begin{bmatrix} 2 \\ 0 \end{bmatrix}, \begin{bmatrix} -2 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \begin{bmatrix} 0 \\ 2 \end{bmatrix}, \begin{bmatrix} -1 \\ 1 \end{bmatrix}, \begin{bmatrix} -1 \\ 2 \end{bmatrix}, \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \begin{bmatrix} -1 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 \\ 1 \end{bmatrix} \right\},$$
$$L_f(K) = \left\{ \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \frac{1}{\sqrt{10}} \begin{bmatrix} -1 \\ 3 \end{bmatrix} \right\}.$$

TABLE 1

Description of the three iterations of cycle ℓ with initial incumbents $x^F = (0, \frac{1}{2^{\ell+1}})^T$ and $x^I = (\frac{1}{2^{\ell+1}}, 0)^T$ and function values $[h(x^F), f(x^F)] = [0, \frac{1}{2^{\ell+1}}]$ and $[h(x^I), f(x^I)] = [\frac{2^{\ell+1}-1}{4^{\ell+1}}, 0]$.

Mesh size (Δ_k)	Poll center (p_k)	$[h, f]$	Poll dirs (D_k)	Trial points	$[h, f]$	Comments
$\frac{1}{2^{\ell+2}}$	$\left(\frac{1}{2^{\ell+1}}, 0\right)$	$\left[\frac{2^{\ell+1}-1}{4^{\ell+1}}, 0\right]$	$(-2, 1)$	$\left(0, \frac{1}{2^{\ell+2}}\right)$	$\left[0, \frac{1}{2^{\ell+2}}\right]$	x^F is improved
			$(2, 0)$	$\left(\frac{1}{2^{\ell}}, 0\right)$	$\left[\frac{2^{\ell}-1}{4^{\ell}}, 0\right]$	Filtered by poll center
			$(0, -1)$	$\left(\frac{1}{2^{\ell+1}}, \frac{-1}{2^{\ell+2}}\right)$	$\left[\frac{3\times 2^{\ell}-1}{4^{\ell+1}}, \frac{-1}{2^{\ell+2}}\right]$	Unfiltered
$\frac{1}{2^{\ell+2}}$	$\left(0, \frac{1}{2^{\ell+2}}\right)$	$\left[0, \frac{1}{2^{\ell+2}}\right]$	$(0, -2)$	$\left(0, \frac{-1}{2^{\ell+2}}\right)$	$\left[\frac{1}{2^{\ell+2}}, \frac{-1}{2^{\ell+2}}\right]$	Unfiltered
			$(1, -1)$	$\left(\frac{1}{2^{\ell+2}}, 0\right)$	$\left[\frac{2^{\ell+2}-1}{4^{\ell+2}}, 0\right]$	x^I is improved
			$(-1, 2)$	$\left(\frac{-1}{2^{\ell+2}}, \frac{3}{2^{\ell+2}}\right)$	$\left[0, \frac{3}{2^{\ell+2}}\right]$	Filtered by poll center
$\frac{1}{2^{\ell+2}}$	$\left(\frac{1}{2^{\ell+2}}, 0\right)$	$\left[\frac{2^{\ell+2}-1}{4^{\ell+2}}, 0\right]$	$(1, 0)$	$\left(\frac{1}{2^{\ell+1}}, 0\right)$	$\left[\frac{2^{\ell+1}-1}{4^{\ell+1}}, 0\right]$	Already in filter
			$(-1, 1)$	$\left(0, \frac{1}{2^{\ell+2}}\right)$	$\left[0, \frac{1}{2^{\ell+2}}\right]$	Already in filter
			$(-1, -1)$	$\left(0, \frac{-1}{2^{\ell+2}}\right)$	$\left[\frac{1}{2^{\ell+2}}, \frac{-1}{2^{\ell+2}}\right]$	Already in filter

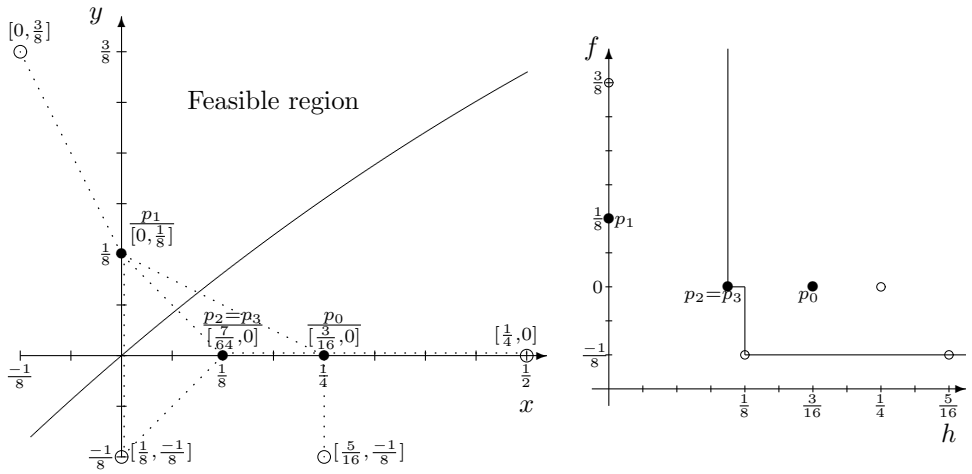


FIG. 7. *The first cycle.*

The polar of the cone spanned by these directions is the cone spanned by the single direction $(0, -1)^T$, which is the gradient at the origin. Thus, Proposition 6.7 is again sharp. The contingent cone at the origin is the half-space $a-b \leq 0$, and the intersection of the contingent cone at the origin with C_f, the cone generated by the convex conic hull of $R_f(K) \cup L_f(K)$, is the convex conic hull of $(-2, 0)^T$ and $(1, 1)^T$.

7.4. Filter results on a Boeing planform design application. The GPS filter algorithm, as implemented in Boeing's Design Explorer, has been applied to wing planform design for several different airplanes. Typically, each airplane requires several redesigns as the design project progresses. The wing planform is the two-dimensional, downward, vertical projection of the wing. The design variables are the line segment end point for the wing leading edges, trailing edges, and spars. Also there are variables related to wing thickness and aerodynamic loading [3]. A typical design problem is to minimize direct operating cost subject to several constraints. The constraints include required range, maximum approach velocity, maximum required runway length, and several others. The analysis code is a sophisticated combination of preliminary design tools from many disciplines. The disciplines include structures, aerodynamics weights, costing, and configuration management.

This problem has 17 variables, 13 nonlinear constraints, and no linear constraints. The best point in the initial surrogate (a kriging model that interpolates data from 200 points obtained from an orthogonal-array–based Latin hypercube) is the least infeasible point, which has a constraint violation of 0.426 and an objective of 9.845.

Figure 8 illustrates the progression of the filter for this application. In all plots, the symbol $\times$ represents the initial point, except in the bottom right plot due to the scale change. The top two plots correspond to the first 15 function evaluations, the middle to the first 50, and the bottom plots show the whole filter after completing 117 evaluations after the initial 200. The initial point gets filtered at the third function evaluation. The first feasible point is found at the 58th evaluation. The best feasible point is denoted on the two bottom plots by a star at $(0, 9.75)^T$.

The bottom left plot contains several trial points with an objective function value near 9.6. This suggests that the SEARCH strategy tried to find a feasible design with such a low f value, but was unsuccessful.

8. Discussion. Though the algorithm behaved very well on the real industrial design problem above (as well as on those in [3] and others), at first glance, one might be unimpressed by the behavior of the algorithm on the academic examples of the last section. Of course, they were designed to illustrate the tightness of our convergence results and the crucial directional dependence of GPS methods. Our interest is in optimization problems, such as the planform problem, where derivative-based methods are impractical. Our algorithm can only rely on function values, and sometimes even these values are not reliable. A design example is presented in [4, 5], where two times out of three the evaluation of the objective function failed to produce a value.

A consequence of this absence of structure is that the convergence results that are guaranteed depend strongly on the set of directions used in the POLL step. Indeed, the richer the set of directions, the stronger the convergence result, since adding directions can increase the number of refining and limit directions, and widen the cone C_f of Proposition 6.7, and hence narrow its polar cone, where the negative gradient of the objective is shown to reside. Intuitively, if a POLL step identifies a poll center that is a mesh isolated filter point, then, the next time a POLL is performed there (with a reduced mesh size parameter) it would be natural to use a different positive spanning set to increase the likelihood of detecting an eventual descent direction. However, essential to the convergence proof is a finite total number of polling directions. It follows that one cannot attempt to obtain a dense set of polling directions within the GPS class of algorithms.

In practice, one would never use a pattern search algorithm following the rules

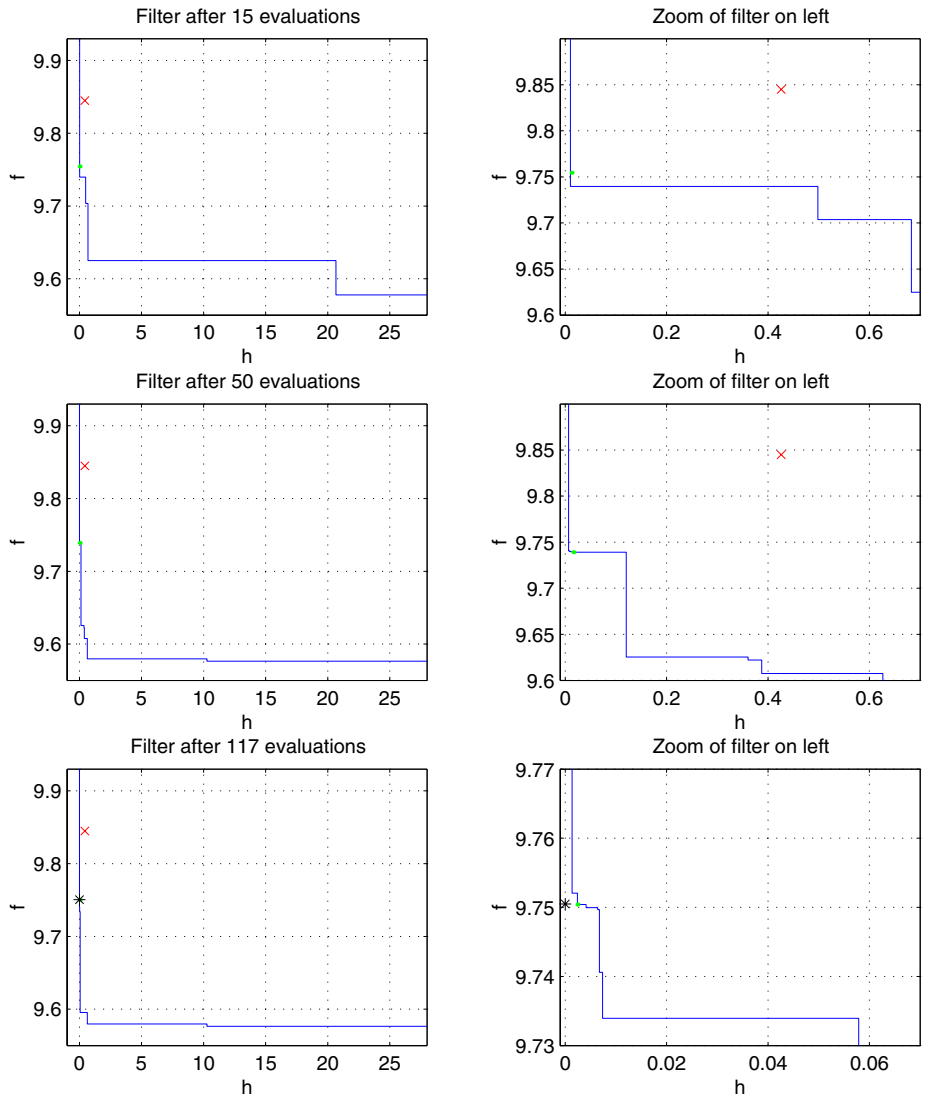


FIG. 8. *Filter progression on a Boeing planform design application.*

upon which these examples are based. First, we would use a SEARCH step such as a space-filling Latin hypercube sampling, a surrogate-based exploration, or a more local SEARCH such as the type suggested in [12]. Second, the set of polling directions would be enlarged in order to avoid large gaps in the directions explored. Finally, the polling centers would sometimes be the feasible incumbent, and sometimes the infeasible one with least constraint violation value, but when promising filter points are generated (such as ones with low f and h values), nothing stops the SEARCH step from including an unofficial POLL around these candidates as a part of the search. These simple algorithmic enhancements fit into the general description of the algorithm presented in section 4.3. Even with these improvements one could devise twisted examples with the behavior of the above examples. It is, however, unlikely that such behavior would be encountered in practice.

Acknowledgments. The authors would like to thank Lt. Col. Mark Abramson, USAF, for many useful suggestions which improved the presentation. We also appreciate the collaborations with Paul Frank of Mathematics and Engineering Analysis at the Boeing Phantom Works and with Alison Marsden of Stanford University. Paul's insightful use of the filter on real Boeing problems and the subsequent feedback he provided has helped us improve our work. Alison provided valuable feedback on her independent implementation of our algorithm applied to an expensive design problem. Finally, we thank both referees for helpful reports that improved the paper.

REFERENCES

[1] C. Audet, *Convergence results for pattern search algorithms are tight*, Optim. Eng., to appear.
[2] C. Audet and J. E. Dennis, Jr., *Analysis of generalized pattern searches*, SIAM J. Optim., 13 (2003), pp. 889–903.
[3] C. Audet, J. E. Dennis, Jr., D. W. Moore, A. J. Booker, and P. D. Frank, *A surrogate-model-based method for constrained optimization*, in Proceedings of the 8th AIAA/USAF/NASA/ISSMO Symposium on Multidisciplinary Analysis and Optimization, AIAA paper 4891, American Institute of Aeronautics and Astronautics, Reston, VA, 2000.
[4] A. J. Booker, J. E. Dennis, Jr., P. D. Frank, D. B. Serafini, V. Torczon, and M. W. Trosset, *A rigorous framework for optimization of expensive functions by surrogates*, Struct. Optim. 17 (1999), pp. 1–13.
[5] A. J. Booker, P. D. Frank, J. E. Dennis, Jr., D. W. Moore, and D. B. Serafini, *Managing surrogate objectives to optimize a helicopter rotor design—further experiments*, in Proceedings of the 7th AIAA/USAF/NASA/ISSMO Symposium on Multidisciplinary Analysis and Optimization, AIAA paper 4717, American Institute of Aeronautics and Astronautics, Reston, VA, 1998.
[6] T. D. Choi, O. J. Eslinger, C. T. Kelley, J. W. David, and M. Etheridge, *Optimization of automotive valve train components with implicit filtering*, Optim. Eng., 1 (2000), pp. 9–27.
[7] T. D. Choi and C. T. Kelley, *Superlinear convergence and implicit filtering*, SIAM J. Optim., 10 (2000), pp. 1149–1162.
[8] F. H. Clarke, *Optimization and Nonsmooth Analysis*, Classics Appl. Math. 5, SIAM, Philadelphia, 1990.
[9] A. R. Conn, N. I. M. Gould, and Ph. L. Toint, *A globally convergent augmented Lagrangian algorithm for optimization with general constraints and simple bounds*, SIAM J. Numer. Anal., 28 (1991), pp. 545–572.
[10] I. D. Coope and C. J. Price, *On the convergence of grid-based methods for unconstrained optimization*, SIAM J. Optim., 11 (2001), pp. 859–869.
[11] J. E. Dennis, Jr., M. El-Alem, and K. Williamson, *A trust-region approach to nonlinear systems of equalities and inequalities*, SIAM J. Optim., 9 (1999), pp. 291–315.
[12] J. E. Dennis, Jr., and V. Torczon, *Direct search methods on parallel machines*, SIAM J. Optim., 1 (1991), pp. 448–474.
[13] H. Eschenauer, J. Koski, and A. Osyczka, eds., *Multicriterion Design Optimization*, Springer, Berlin, 1990.
[14] R. Fletcher and S. Leyffer, *Nonlinear programming without a penalty function*, Math. Program., 91 (2002), pp. 239–269.
[15] R. Fletcher, S. Leyffer, and Ph. L. Toint, *On the global convergence of a filter–SQP algorithm*, SIAM J. Optim., 13 (2002), pp. 44–59.
[16] R. Fletcher, N. I. M. Gould, S. Leyffer, Ph. L. Toint, and A. Wächter, *Global convergence of a trust-region SQP-filter algorithm for general nonlinear programming*, SIAM J. Optim., 13 (2002), pp. 635–659.
[17] P. D. Frank, *private communication*, 2000.
[18] P. Gilmore and C. T. Kelley, *An implicit filtering algorithm for optimization of functions with many local minima*, SIAM J. Optim., 5 (1995), pp. 269–285.
[19] J. Jahn, *Introduction to the Theory of Nonlinear Optimization*, 2nd ed., Springer, Berlin, 1996.
[20] E. B. Leach, *A note on inverse function theorems*, Proc. Amer. Math. Soc., 12 (1961), pp. 694–697.

[21] R. M. Lewis and V. Torczon, *Pattern search algorithms for bound constrained minimization*, SIAM J. Optim., 9 (1999), pp. 1082–1099.

[22] R. M. Lewis and V. Torczon, *Rank Ordering and Positive Bases in Pattern Search Algorithms*, ICASE report 96-71, NASA Langley Research Center, Hampton, VA, 1996. Available online at http://techreports.larc.nasa.gov/icase/1996/icase-1996-71.pdf.

[23] R. M. Lewis and V. Torczon, *Pattern search methods for linearly constrained minimization*, SIAM J. Optim., 10 (2000), pp. 917–941.

[24] R. M. Lewis and V. Torczon, *A globally convergent augmented Lagrangian pattern search algorithm for optimization with general constraints and simple bounds*, SIAM J. Optim., 12 (2002), pp. 1075–1089.

[25] A. L. Marsden, M. Wang, J. E. Dennis, Jr., and P. Moin, *Optimal aeroacoustic shape design using the surrogate management framework*, Optim. Eng., submitted.

[26] R. T. Rockafellar, *Convex Analysis*, Princeton Landmarks Math., Princeton University Press, Princeton, NJ, 1997.

[27] C. P. Stephens and W. Baritompa, *Global optimization requires global information*, J. Optim. Theory Appl., 96 (1998), pp. 575–588.

[28] V. Torczon, *On the convergence of pattern search algorithms*, SIAM J. Optim., 7 (1997), pp. 1–25.

SIAM J. OPTIM.
Vol. 14, No. 4, pp. 1011–1027

WEAK SHARP SOLUTIONS OF VARIATIONAL INEQUALITIES IN HILBERT SPACES*

ZILI WU† AND SOON-YI WU‡

Abstract. Under some new conditions, we present several equivalent (and sufficient) conditions for weak sharp solutions of variational inequalities in Hilbert spaces and give a finite convergence result for a class of algorithms for solving variational inequalities.

Key words. variational inequality, gap functions, weak sharpness, error bound, minimum principle sufficiency property, maximum principle sufficiency property

AMS subject classifications. 90C33, 49J52

DOI. 10.1137/S1052623403421486

1. Introduction.

Let H be a Hilbert space, C be a nonempty closed subset of H, and F be a mapping from H into H. The *variational inequality problem* (VIP(C, F)) is to find a vector $x^* \in C$ such that

$$\langle F(x^*), x - x^* \rangle \geq 0 \text{ for all } x \in C. \tag{1}$$

The *dual variational inequality problem* (DVIP(C, F)) is to find a vector $x^* \in C$ such that

$$\langle F(x), x - x^* \rangle \geq 0 \text{ for all } x \in C. \tag{2}$$

We denote the solution set of the VIP(C, F) by C^* and that of the DVIP(C, F) by C_* and assume that C^* and C_* are nonempty. From now on, we also assume C to be convex.

Variational inequality problems occur in a number of applications for which the reader can refer to Harker and Pang's survey paper [5] and references therein.

To measure the violation of the solutions to the VIP(C, F) at any point $x \in C$, one often uses the following gap functions.

The *primal gap function* $g(x)$ associated with the VIP(C, F) is defined as

$$g(x) := \max\{\langle F(x), x - c \rangle : c \in C\} \text{ for } x \in H.$$

We denote

$$\Gamma(x) := \{c \in C : \langle F(x), x - c \rangle = g(x)\}.$$

Similarly, we define the *dual gap function* $G(x)$ associated with the VIP(C, F) as

$$G(x) := \max\{\langle F(c), x - c \rangle : c \in C\} \text{ for } x \in H$$

*Received by the editors January 18, 2003; accepted for publication (in revised form) October 2, 2003; published electronically May 25, 2004. This work was supported by NSC of R.O.C.
http://www.siam.org/journals/siopt/14-4/42148.html

†Department of Mathematics and Statistics, University of Victoria, Victoria, BC, Canada V8W 3P4 (ziliwu@email.com). Current address: Department of Mathematics, National Cheng Kung University, Tainan, Taiwan 701.

‡Department of Mathematics, National Cheng Kung University, Tainan, Taiwan 701 (soonyi@mail.ncku.edu.tw).

and denote

$$\Lambda(x) := \{c \in C : \langle F(c), x - c\rangle = G(x)\}.$$

(For more general gap functions, we refer to [8] and references therein.)

It is easy to see that these two gap functions are both nonnegative on C and the dual gap function is convex. In addition, if G is bounded above on a neighborhood of a point $x \in H$, then G is Lipschitz near x; that is, there exists a constant $M > 0$ such that

$$\|G(x_1) - G(x_2)\| \le M\|x_1 - x_2\|$$

for all x_1 and x_2 in a neighborhood of x (see [2, Proposition 2.2.6]).

When G is *Gâteaux differentiable* at x, that is, there exists $\xi \in H$ such that the directional derivative $G'(x;\cdot)$ of G at x satisfies

$$G'(x;v) := \lim_{t\to 0^+} \frac{G(x+tv) - G(x)}{t} = \langle \xi, v\rangle \text{ for all } v \in H,$$

the Lipschitz continuity of G near x implies that the *Gâteaux derivative* ξ (also denoted by $\nabla G(x)$) satisfies that

$$\lim_{\substack{u\to v\\ t\to 0^+}} \frac{G(x+tu) - G(x)}{t} = \langle \xi, v\rangle \text{ for all } v \in H.$$

Following Ferris and Mangasarian [4], we say that the VIP(C, F) has the *minimum principle sufficiency* property if

$$\Gamma(x^*) = C^* \text{ for each } x^* \in C^*.$$

Similarly, the VIP(C, F) has the *maximum principle sufficiency* property if

$$\Lambda(x^*) = C^* \text{ for each } x^* \in C^*.$$

For a nonempty convex set C, the *normal cone* to C at $x \in H$ is defined by

$$N_C(x) := \begin{cases} \{\xi \in H : \langle \xi, c - x\rangle \le 0 \text{ for all } c \in C\} & \text{if } x \in C,\\ \emptyset & \text{if } x \notin C. \end{cases}$$

The *tangent cone* to C at x is given by $T_C(x) := [N_C(x)]^\circ$, where A° denotes the polar set of $A \subseteq H$ defined by

$$A^\circ := \{v \in H : \langle v, x\rangle \le 0 \text{ for all } x \in A\}.$$

It is known that

$$T_C(x) = \{v \in H : d'_C(x;v) = 0\}$$

(see [2]), where d_C stands for the *distance function* associated with C given by

$$d_C(x) := \inf\{\|c - x\| : c \in C\} \text{ for each } x \in H$$

and $d'_C(x;v)$ is the directional derivative of d_C at x in the direction $v \in H$:

$$d'_C(x;v) := \lim_{t\to 0^+} \frac{d_C(x+tv) - d_C(x)}{t}.$$

For $x^* \in C^*$, we have

$$\langle -F(x^*), x - x^* \rangle \leq 0 \text{ for all } x \in C,$$

that is, $-F(x^*) \in N_C(x^*) = [T_C(x^*)]^\circ$. If

$$-F(x^*) \in \operatorname{int} \bigcap_{x \in C^*} [T_C(x) \cap N_{C^*}(x)]^\circ \text{ for each } x^* \in C^*,$$

the set C^* is said to be *weakly sharp* (according to Patriksson [10]). This is equivalent to saying that for each $x^* \in C^*$ there exists $\alpha > 0$ such that

$$\alpha B \subseteq F(x^*) + \bigcap_{x \in C^*} [T_C(x) \cap N_{C^*}(x)]^\circ,$$

where B denotes the open unit ball in H.

A mapping $F : H \to H$ is said to be *pseudomonotone* at $x \in C$ if for each $y \in C$ there holds

$$\langle F(x), y - x \rangle \geq 0 \Rightarrow \langle F(y), y - x \rangle \geq 0.$$

We say that F is *pseudomonotone* on a subset C_1 of C if it is pseudomonotone at each $x \in C_1$. F is *pseudomonotone*$^+$ on C if it is pseudomonotone on C and, for all $x, y \in C$,

$$\langle F(x) - F(y), x - y \rangle = 0 \Rightarrow F(x) = F(y).$$

F is *pseudomonotone*$_*$ on C if it is pseudomonotone on C and, for some $k > 0$ and all $x, y \in C$,

$$\left.\begin{array}{ll} \langle F(x), x - y \rangle & = 0 \\ \langle F(y), x - y \rangle & = 0 \end{array}\right\} \Rightarrow F(x) = kF(y).$$

In particular, F is *pseudomonotone*$_*^+$ on C if it is pseudomonotone$_*$ on C with $k = 1$.

From the above definition, we see that

$$\begin{gathered} \text{pseudomonotonicity}^+ \text{ of } F \Rightarrow \text{pseudomonotonicity}_*^+ \text{ of F} \Rightarrow \\ \text{pseudomonotonicity}_* \text{ of F} \Rightarrow \text{pseudomonotonicity of } F. \end{gathered}$$

The notion of a weak sharp minimum was introduced by Burke and Ferris [1] to present sufficient conditions for the finite identification, by iterative algorithm, of local minima associated with mathematical programming in R^n. Their results have been extended by Marcotte and Zhu [9] to the variational inequality problem under the assumption that F is pseudomonotone$^+$ and continuous on a compact convex set C in R^n. Marcotte and Zhu [9] showed that C^* is weakly sharp iff G has an *error bound* on C, that is, there exists some $\mu > 0$ such that

$$d_{C^*}(x) \leq \mu G(x) \text{ for each } x \in C.$$

If C is also a compact polyhedral set in R^n, then C^* is weakly sharp iff the VIP(C, F) has the minimum principle sufficiency property.

The generalized monotonicity is used in [9] mainly to guarantee the solution set for the variational inequality coincides with the optimal solution sets of corresponding

gap functions. Such coincidence plays an important role in characterizing the weak sharpness.

Our purpose in this paper is to develop the above weak sharpness results in the Hilbert space H by presenting the following novel condition:

$$\{v \in H : \langle F(x^*), v\rangle \geq 0\} = \{v \in H : \langle F(y^*), v\rangle \geq 0\} \text{ for } x^*,\, y^* \in C; \tag{3}$$

that is, $F(x^*)$ and $F(y^*)$ have the same direction. We organize the rest of this paper as follows. In section 2, we discuss basic relations among C^*, C_*, $\Gamma(x)$, and $\Lambda(x)$ under assumption (3). We show in section 3 that C^* has the minimum principle sufficiency property if (3) holds for each $x^* \in C^* \subseteq C_*$ and each $y^* \in \Gamma(x^*)$ or if C^* is weakly sharp and (3) holds for each $x^* \in C^*$ and each $y^* \in C_*$. In parallel, we give several equivalent conditions for C^* to possess the maximum principle sufficiency property in section 4. These equivalences show that the maximum principle sufficiency property bears close relation to (3) and the value of F on $\Gamma(x^*)$ and $\Lambda(x^*)$. In section 5, under a weaker condition than in Marcotte and Zhu [9], we show that C^* is weakly sharp iff G has an error bound on C, which is in turn equivalent to saying that C^* has the minimum principle sufficiency property and there exists $\alpha > 0$ such that

$$\alpha B \cap [F(x^*) + N_C(x)] = \emptyset \text{ for each } x^* \in C^* \text{ and each } x \in C \backslash C^*.$$

We also present two sufficient conditions for C^* to be weakly sharp and prove that a class of algorithms are convergent to C^* in finitely many steps under the assumption of weak sharpness of C^*.

2. Relationship among C^*, C_*, $\Gamma(x)$, and $\Lambda(x)$. Recall that the pseudomonotonicity of F on C is sufficient for $C^* \subseteq C_*$, while the continuity of F on C guarantees the inclusion $C_* \subseteq C^*$. So $C^* = C_*$ whenever F is pseudomonotone and continuous on C. In this section, we mainly reveal the relations among C^*, C_*, $\Gamma(x)$, and $\Lambda(x)$ under assumption (3) and show how it plays a role similar to that of the pseudomonotonicity and continuity of F. However, before this, we present some basic relations first without detailed proof.

The following proposition is simple but it is very convenient for us to use it implicitly.

PROPOSITION 2.1. *For $x^* \in C$,*

(i) $x^* \in C^* \Leftrightarrow g(x^*) = 0 \Leftrightarrow x^* \in \Gamma(x^*)$;

(ii) $x^* \in C_* \Leftrightarrow G(x^*) = 0 \Leftrightarrow x^* \in \Lambda(x^*)$.

Proof. (i) has been stated in [6] (or follows directly from [8, Theorem 4.1]), while (ii) is made easily from definitions. □

The next proposition characterizes solutions to the VIP(C, F) and DVIP(C, F).

PROPOSITION 2.2. *For $x^* \in C$ and $y^* \in C$ the following are equivalent:*

(i) *$x^* \in C^*$ and $y^* \in C_*$.*

(ii) *$\langle F(x^*), x^* - y\rangle \leq \langle F(x^*), x^* - y^*\rangle \leq \langle F(x), x - y^*\rangle$ for all $x, y \in C$.*

For each $x^* \in C^*$ and each $y^* \in C_*$, by taking $y = x^*$ and $x = y^*$ in (ii) of Proposition 2.2, we have $\langle F(x^*), x^* - y^*\rangle = 0$. This implies that $y^* \in \Gamma(x^*)$ and $x^* \in \Lambda(y^*)$. So C_* is contained in $\Gamma(x^*)$, while C^* is in $\Lambda(y^*)$.

PROPOSITION 2.3. *The following hold:*

(i) *$C_* \subseteq \Gamma(x^*)$ for each $x^* \in C^*$.*

(ii) *$C^* \subseteq \Lambda(x^*)$ for each $x^* \in C_*$.*

In particular, if F is continuous on C_*, then $C_* \subseteq \Gamma(x^*)$ for each $x^* \in C_*$; if F is pseudomonotone on C^*, then $C^* \subseteq \Lambda(x^*)$ for each $x^* \in C^*$.

Note that the following hold:

$$\langle F(x^*), x^* - y^* \rangle = 0 \text{ for each } x^* \in C^* \text{ and } y^* \in \Gamma(x^*);$$
$$\langle F(y^*), x^* - y^* \rangle = 0 \text{ for each } x^* \in C_* \text{ and } y^* \in \Lambda(x^*).$$

According to these, we have the following equivalences.

PROPOSITION 2.4. *The following hold:*

(i) *For each* $x^* \in C^*$ *and* $y^* \in \Gamma(x^*)$,

$$\langle F(x^*) - F(y^*), x^* - y^* \rangle = 0 \Leftrightarrow \langle F(y^*), x^* - y^* \rangle = 0.$$

(ii) *For each* $x^* \in C_*$ *and* $y^* \in \Lambda(x^*)$,

$$\langle F(x^*) - F(y^*), x^* - y^* \rangle = 0 \Leftrightarrow \langle F(x^*), x^* - y^* \rangle = 0.$$

The next proposition states that under assumption (3) the equality $\langle F(x^*), x^* - y^* \rangle = 0$ is equivalent to $\langle F(y^*), x^* - y^* \rangle = 0$ and implies that both x^* and y^* are in C^* as long as one of them is in C^*.

PROPOSITION 2.5. *Let* $x^* \in C$ *and* $y^* \in C$ *satisfy* (3).

(i) $\langle F(x^*), x^* - y^* \rangle = 0 \Leftrightarrow \langle F(y^*), x^* - y^* \rangle = 0$.

(ii) *If either* $\langle F(x^*), x^* - y^* \rangle = 0$ *or* $\langle F(y^*), x^* - y^* \rangle = 0$, *then*

$$x^* \in C^* \Leftrightarrow y^* \in C^*.$$

From Proposition 2.5 we see that (3) is sufficient for a mapping F to possess *the cut property* as below:

$$\left.\begin{array}{l} x^* \in C^* \\ x^* \neq y^* \in C \\ \langle F(y^*), x^* - y^* \rangle = 0 \end{array}\right\} \Rightarrow y^* \in C^*,$$

which is required in an algorithm designed to solve concave-convex games (see [3]).

The following result shows that (3) plays an important role in revealing the relations among C^*, C_*, $\Gamma(x^*)$, and $\Lambda(x^*)$ like the pseudomonotonicity and continuity of F on C.

THEOREM 2.6. *Let* $x^* \in C$ *and* $y^* \in C$ *satisfy* (3).

(i) $x^* \in C^*$ *and* $y^* \in \Gamma(x^*)$ *iff* $x^* \in \Gamma(x^*)$ *and* $y^* \in C^*$.

(ii) $x^* \in C^*$ *and* $y^* \in \Lambda(x^*)$ *iff* $x^* \in \Lambda(x^*)$ *and* $y^* \in C^*$.

Proof. (i) Let $x^* \in C^*$ and $y^* \in \Gamma(x^*)$. Then $x^* \in \Gamma(x^*)$ and

$$\langle F(x^*), x^* - y^* \rangle = g(x^*) = 0,$$

which, by Proposition 2.5, implies $y^* \in C^*$.

Conversely, suppose that $x^* \in \Gamma(x^*)$ and $y^* \in C^*$. Then $x^* \in C^*$ and

$$\langle F(y^*), x^* - y^* \rangle \geq 0,$$

from which, with (3), we have $0 = g(x^*) \geq \langle F(x^*), x^* - y^* \rangle \geq 0$. So $y^* \in \Gamma(x^*)$.

(ii) To prove the necessity, let $x^* \in C^*$ and $y^* \in \Lambda(x^*)$. Then

$$\langle F(x^*), y^* - x^* \rangle \geq 0 \text{ and } \langle F(y^*), x^* - y^* \rangle = G(x^*) \geq 0,$$

which together with (3) implies $\langle F(x^*), x^* - y^* \rangle \geq 0$. Thus $\langle F(x^*), x^* - y^* \rangle = 0$. But this in turn implies $\langle F(y^*), y^* - x^* \rangle \geq 0$. Thus

$$G(x^*) = \langle F(y^*), x^* - y^* \rangle = 0,$$

that is, $x^* \in \Lambda(x^*)$ and, by Proposition 2.5, y^* must lie in C^*.

Next we prove the sufficiency. Suppose that $x^* \in \Lambda(x^*)$ and $y^* \in C^*$. Then

$$0 = G(x^*) \geq \langle F(y^*), x^* - y^* \rangle \geq 0,$$

that is, $y^* \in \Lambda(x^*)$ and $\langle F(y^*), x^* - y^* \rangle = 0$. It follows from Proposition 2.5 that $x^* \in C^*$. □

COROLLARY 2.7. *Let $x^* \in C$ and $y^* \in C$ satisfy* (3).

(i) *If $x^* \in C^*$, then $y^* \in C^* \Leftrightarrow y^* \in \Gamma(x^*)$.*
(ii) *If $x^* \in C^* \cap \Lambda(y^*)$, then $y^* \in C^* \Leftrightarrow y^* \in C_*$.*
(iii) *If $x^* \in C^* \cap C_*$, then $y^* \in C^* \Leftrightarrow y^* \in \Lambda(x^*)$.*

Note that for points x^* and y^* in C^* assumption (3) implies

$$\langle F(x^*), x^* - y^* \rangle = 0 \text{ and } \langle F(y^*), x^* - y^* \rangle = 0.$$

This turns out to be true for each pair of points x^* and y^* in a set containing C^* as below.

PROPOSITION 2.8. *Let $k > 0$. If both x^* and y^* lie in $\Gamma(x^*) \cup \Lambda(x^*) \cup C^* \cup \Gamma(y^*) \cup \Lambda(y^*)$, then the following statements satisfy*

$$\text{(i)} \Rightarrow \text{(ii)} \Rightarrow \text{(iii)} \Rightarrow \text{(iv)} \Rightarrow \text{(v)} :$$

(i) $F(y^*) = kF(x^*)$.
(ii) $\{v \in H : \langle F(x^*), v \rangle \geq 0\} = \{v \in H : \langle F(y^*), v \rangle \geq 0\}$.
(iii) $\begin{cases} \langle F(x^*), x^* - y^* \rangle \geq 0 \Leftrightarrow \langle F(y^*), x^* - y^* \rangle \geq 0, \\ \langle F(x^*), y^* - x^* \rangle \geq 0 \Leftrightarrow \langle F(y^*), y^* - x^* \rangle \geq 0. \end{cases}$
(iv) $\langle F(x^*), x^* - y^* \rangle = 0$ *and* $\langle F(y^*), x^* - y^* \rangle = 0$.
(v) $\langle F(x^*) - F(y^*), x^* - y^* \rangle = 0$.

Hence, if (iv)((v)) $\Rightarrow$ (i), *then* (i) $\Leftrightarrow$ (ii) $\Leftrightarrow$ (iii) $\Leftrightarrow$ (iv)($\Leftrightarrow$ (v)).

Proof. The implications "(i) $\Rightarrow$ (ii) $\Rightarrow$ (iii)" and "(iv) $\Rightarrow$ (v)" are immediate.

To prove (iii) $\Rightarrow$ (iv), we suppose that (iii) holds. According to Proposition 2.1, we need only to prove that (iv) holds for the following four cases:

$$x^* \in C^* \cup \Lambda(y^*) \text{ and } y^* \in \Lambda(x^*) \cup C^*,\ x^* \in C^* \cup \Lambda(y^*) \text{ and } y^* \in \Gamma(x^*) \cup \Lambda(y^*),$$

$$x^* \in \Lambda(x^*) \cup \Gamma(y^*) \text{ and } y^* \in \Lambda(x^*) \cup C^*,\ x^* \in \Lambda(x^*) \cup \Gamma(y^*) \text{ and } y^* \in \Gamma(x^*) \cup \Lambda(y^*).$$

We omit the further proof since it is trivial. □

Remark 2.1. For $x^* \in C_*$ and $y^* \in \Lambda(x^*)$, since we already have $\langle F(y^*), x^* - y^* \rangle = 0$, both (iv) and (v) in Proposition 2.8 are equivalent to $\langle F(x^*), x^* - y^* \rangle = 0$. In addition, if F is pseudomonotone$_*$ on C, then (i)–(iv) in Proposition 2.8 are equivalent to one another.

3. Minimum principle sufficiency property. Under the assumption that F be pseudomonotone and continuous on $C \subseteq R^n$, Marcotte and Zhu [9] showed that the weak sharpness of C^* implies that the VIP(C, F) has the minimum principle sufficiency property. In this section, we prove that this implication is still valid whenever (3) holds for each $x^* \in C^*$ and $y^* \in C_*$. We need the following result in which a sufficient condition is also presented for the VIP(C, F) to have the minimum principle sufficiency property.

PROPOSITION 3.1.

(i) *If* (3) *holds for* $x^* \in C^*$ *and all* $y^* \in \Gamma(x^*)$, *then* $\Gamma(x^*) \subseteq C^*$.

(ii) *If* (3) *holds for* $x^* \in C_*$ *and all* $y^* \in \Gamma(x^*)$, *then* $x^* \in \Gamma(x^*) = C^*$.

(iii) *If* (3) *holds for* $x^* \in C^* \cup C_*$ *and all* $y^* \in C^*$, *then* $x^* \in C^* \subseteq \Gamma(x^*)$.

Proof. Since (i) is direct from Theorem 2.6, it suffices to show (ii) and (iii).

Let $x^* \in C_*$. For any fixed y^* in $\Gamma(x^*)$ satisfying (3), we have

$$0 = G(x^*) \geq \langle F(y^*), x^* - y^* \rangle \text{ and } \langle F(x^*), x^* - y^* \rangle = g(x^*) \geq 0.$$

By (3), the first inequality above implies $\langle F(x^*), x^* - y^* \rangle \leq 0$, so

$$g(x^*) = \langle F(x^*), x^* - y^* \rangle = 0.$$

Thus $x^* \in \Gamma(x^*)$ (that is, $x^* \in C^*$) and, by Proposition 2.5, we obtain $y^* \in C^*$. Since y^* is arbitrary, $\Gamma(x^*) \subseteq C^*$. On the other hand, if $y^* \in C^*$, then

$$0 \leq \langle F(y^*), x^* - y^* \rangle \leq G(x^*) = 0.$$

By Proposition 2.5, $x^* \in C^*$ and $\langle F(x^*), x^* - y^* \rangle = 0 = g(x^*)$. Thus $y^* \in \Gamma(x^*)$. Hence $C^* \subseteq \Gamma(x^*)$. Therefore (ii) follows.

To show (iii), let (3) hold for $x^* \in C^* \cup C_*$ and all $y^* \in C^*$. If $x^* \in C^*$, then it follows from Theorem 2.6 that $x^* \in C^* \subseteq \Gamma(x^*)$. If $x^* \in C_*$, then

$$0 \leq \langle F(y^*), y^* - x^* \rangle \leq g(y^*) = 0,$$

that is, $\langle F(y^*), y^* - x^* \rangle = 0$. It follows from Proposition 2.5 that $x^* \in C^*$ and

$$\langle F(x^*), x^* - y^* \rangle = 0,$$

which implies that $y^* \in \Gamma(x^*)$ for all $y^* \in C^*$. Thus $C^* \subseteq \Gamma(x^*)$. Therefore (iii) follows. □

From Propositions 2.3 and 3.1 we see that the VIP(C, F) has the minimum principle sufficiency property when $C^* \subseteq C_*$ and (3) holds for each $x^* \in C^*$ and each $y^* \in \Gamma(x^*)$.

In [9, Theorem 4.2], Marcotte and Zhu presented a sufficient condition for the VIP(C, F) to have the minimum principle sufficiency property in terms of the pseudomonotonicity of F and the concepts of tangent cone and normal cone to C. Similarly, we give a sufficient condition for this as follows.

THEOREM 3.2. *Let* C_1 *be a nonempty closed convex subset of* C *and let*

$$K_1 := \operatorname{int} \bigcap_{x \in C_1} [T_C(x) \cap N_{C_1}(x)]^\circ$$

be nonempty. Then, for each $v \in K_1$, $\arg\max\{\langle v, x \rangle : x \in C\} \subseteq C_1$. *Hence, if* $C_1 = C_*$ *and* $-F(x^*) \in K_1$ *for each* $x^* \in C^*$, *then*

$$C^* \subseteq C_* = \Gamma(x^*) \text{ for each } x^* \in C^*.$$

If F is also continuous on C_* or (3) *holds for each $x^* \in C^*$ and each $y^* \in C_*$, then*

$$C^* = C_* = \Gamma(x^*) \text{ for each } x^* \in C^*.$$

Proof. Let $x \in C$ but $x \notin C_1$. Then, since C_1 is closed and convex, there exists a unique vector $\overline{x} \in C_1$ such that $\|x - \overline{x}\| = d_{C_1}(x)$. This implies that

$$x - \overline{x} \in T_C(\overline{x}) \cap N_{C_1}(\overline{x}),$$

and for each $v \in K_1$ there exists $\delta > 0$ such that

$$\langle v + u, x - \overline{x} \rangle < 0 \text{ for all } u \in \delta B,$$

from which we obtain

$$\langle v, x \rangle < \langle v, \overline{x} \rangle - \frac{\delta}{2}\|x - \overline{x}\|.$$

Thus $x \notin \arg\max\{\langle v, y \rangle : y \in C\} \subseteq C_1$.

Next, if $C_1 = C_*$ and $-F(x^*) \in K_1$ for each $x^* \in C^*$, then

$$x^* \in \Gamma(x^*) = \arg\max\{\langle -F(x^*), x \rangle : x \in C\} \subseteq C_*,$$

from which it follows that $C^* \cup \Gamma(x^*) \subseteq C_*$. In addition, by Proposition 2.3, $C_* \subseteq \Gamma(x^*)$ for each $x^* \in C^*$. Therefore

$$C^* \subseteq C_* = \Gamma(x^*) \text{ for each } x^* \in C^*.$$

Moreover, if either F is continuous on C_* or (3) holds for each $x^* \in C^*$ and each $y^* \in C_*$, then the above inclusion reduces to an equality since in either case $C_* \subseteq C^*$ holds (see (iii) in Proposition 3.1 for the second case). Hence the result desired follows. □

Remark 3.1. When F is continuous on $C \subseteq R^n$ and C^* is convex, the first part of the result in Theorem 3.2 for $C_1 = C^*$, that is,

$$\arg\max\{\langle v, x \rangle : x \in C\} \subseteq C^*,$$

was first established by Marcotte and Zhu in [9, Theorem 4.2]. Since in this case we have $C_* \subseteq C^*$, the result in Theorem 3.2 corresponding to $C_1 = C_*$ implies the above inclusion. In addition, if $-F(x^*) \in K_1$ for each $x^* \in C^*$, then C^* coincides not only with $\Gamma(x^*)$ but also with C_* without the assumption of pseudomonotonicity. Finally, if $C^* = C_*$, the first part of result in Theorem 3.2 shows that the weak sharpness of C^* implies the minimum sufficiency property of it.

4. Maximum principle sufficiency property. Like the minimum principle sufficiency property, the maximum principle sufficiency property is also very useful in characterizing the weak sharpness of C^*. To see this, we present several sufficient conditions for such a property in terms of (3) in this section.

THEOREM 4.1. *Let $C^* \neq \emptyset$.*

(i) *If* (3) *holds for $x^* \in C^* \cup C_*$ and all $y^* \in \Lambda(x^*)$, then $x^* \in \Lambda(x^*) = C^*$.*

(ii) *If for each $x^* \in C^*$ there exists $y^* \in \Lambda(x^*)$ such that* (3) *holds, then $C^* \subseteq C_*$.*

(iii) *If for each $x^* \in C_*$ there exists $y^* \in C^*$ such that* (3) *holds, then $C_* \subseteq C^*$.*

(iv) *If* (3) *holds for each $x^* \in C^* \cup C_*$ and each $y^* \in \Lambda(x^*)$, then*

$$C^* = \Lambda(x^*) = C_* \text{ for each } x^* \in C^* \cup C_*.$$

Proof. (i) If (3) holds for $x^* \in C^*$ and all $y^* \in \Lambda(x^*)$, then, by Theorem 2.6,

$$x^* \in \Lambda(x^*) \subseteq C^*.$$

On the other hand, by Propositions 2.1 and 2.3, we have $C^* \subseteq \Lambda(x^*)$ and hence, $\Lambda(x^*) = C^*$.

If (3) holds for $x^* \in C_*$ and all $y^* \in \Lambda(x^*)$, then

$$\langle F(y^*), x^* - y^* \rangle = 0 \text{ for each } y^* \in \Lambda(x^*).$$

In particular, by Proposition 2.3, this equality holds for each $y^* \in C^*$. By Proposition 2.5, $x^* \in C^*$, and hence $y^* \in C^*$ for each $y^* \in \Lambda(x^*)$. So $\Lambda(x^*) \subseteq C^*$. Thus, by Proposition 2.3 again, $\Lambda(x^*) = C^*$.

Since (ii) and (iii) are immediate from Theorem 2.6, it remains to show (iv).

Suppose that (3) holds for each $x^* \in C^* \cup C_*$ and each $y^* \in \Lambda(x^*)$. Then it follows from (i) that

$$\Lambda(x^*) = C^* \text{ for each } x^* \in C^* \cup C_*.$$

In addition, by (ii) and (iii), we have $C^* = C_*$. Therefore (iv) follows. ∎

In the case stated in next result, the sufficient condition in (iv) of Theorem 4.1 becomes equivalent for the VIP(C, F) to possess the maximum principle sufficiency property.

THEOREM 4.2. *Let*

$$\langle F(x^*), x^* - y^* \rangle = 0 \text{ and } \langle F(y^*), x^* - y^* \rangle = 0 \Rightarrow F(y^*) = kF(x^*)$$

for some $k > 0$, each $x^* \in C^* \cup C_*$, and each $y^* \in \Lambda(x^*)$. Then the following are equivalent:

(i) (3) *holds for each $x^* \in C^* \cup C_*$ and each $y^* \in \Lambda(x^*)$.*

(ii) *$C^* = \Lambda(x^*) = C_*$ for each $x^* \in C^* \cup C_*$.*

(iii) *For each $x^* \in C^* \cup C_*$ and each $y^* \in \Lambda(x^*)$,*

$$\begin{cases} \langle F(x^*), x^* - y^* \rangle \geq 0 \Leftrightarrow \langle F(y^*), x^* - y^* \rangle \geq 0, \\ \langle F(x^*), y^* - x^* \rangle \geq 0 \Leftrightarrow \langle F(y^*), y^* - x^* \rangle \geq 0. \end{cases}$$

(iv) *For each $x^* \in C^* \cup C_*$ and each $y^* \in \Lambda(x^*)$,*

$$\langle F(x^*), x^* - y^* \rangle = 0 \text{ and } \langle F(y^*), x^* - y^* \rangle = 0.$$

(v) *$F(y^*) = kF(x^*)$ for each $x^* \in C^* \cup C_*$ and each $y^* \in \Lambda(x^*)$.*

Proof. The equivalences (i) ⇔ (iii) ⇔ (iv) ⇔ (v) have been stated in Proposition 2.8. It suffices to show (i) ⇒ (ii) ⇒ (iv). But the first implication is just (iv) in Theorem 4.1, while the second one follows directly from Proposition 2.2. Hence the proof is complete. ∎

Remark 4.1. In the case $C^* \subseteq C_*$, the expression $C^* \cup C_*$ in Theorem 4.2 reduces to C_*, and the equality $\langle F(y^*), x^* - y^* \rangle = 0$ in (iv) can be omitted. When F is both continuous and pseudomonotone$_*$ on C, all statements in Theorem 4.2 are true as stated below.

THEOREM 4.3. *Let F be continuous on C_* and pseudomonotone$_*$ on C. Then, for some $k > 0$,* (i)–(v) *in Theorem* 4.2 *hold.*

Proof. If F is continuous on C_* and pseudomonotone$_*$ on C, then $C^* = C_*$. Since (iv) is easy to verify for $x^* \in C_*$ and $y^* \in \Lambda(x^*)$, by Theorem 4.2, (i)–(v) hold. ∎

Remark 4.2. Since the pseudomonotonicity$^+$ of F implies the pseudomonotonicity$_*$ of F, if F is continuous and pseudomonotone$^+$ on C, all statements in Theorem 4.3 hold and all are equivalent with $k = 1$. Hence Theorem 4.3 extends [9, Theorem 3.1], which states that $\Lambda(x^*) = C^*$ and $F(y^*) = F(x^*)$ hold for each $x^* \in C^*$ and each $y^* \in C^* \cup \Lambda(x^*)$ under the assumption that F is continuous and pseudomonotone$^+$ on $C \subseteq R^n$.

Combining Proposition 3.1 with Theorem 4.1, we obtain the following sufficient condition and necessary condition for the VIP(C, F) to have both the minimum principle sufficiency property and the maximum principle sufficiency property.

THEOREM 4.4. *For some $k > 0$, the following statements satisfy*

$$(\text{v}) \Rightarrow (\text{i}) \Rightarrow (\text{ii})((\text{iii})) \Rightarrow (\text{iv}) :$$

(i) (3) *holds for each $x^* \in C^* \cup C_*$ and each $y^* \in \Gamma(x^*) \cup \Lambda(x^*)$.*

(ii) *$C^* = \Gamma(x^*) = \Lambda(x^*) = C_*$ for each $x^* \in C^* \cup C_*$.*

(iii) *For each $x^* \in C^* \cup C_*$ and each $y^* \in \Gamma(x^*) \cup \Lambda(x^*)$,*

$$\begin{cases} \langle F(x^*), x^* - y^* \rangle \geq 0 \Leftrightarrow \langle F(y^*), x^* - y^* \rangle \geq 0, \\ \langle F(x^*), y^* - x^* \rangle \geq 0 \Leftrightarrow \langle F(y^*), y^* - x^* \rangle \geq 0. \end{cases}$$

(iv) *For each $x^* \in C^* \cup C_*$ and each $y^* \in \Gamma(x^*) \cup \Lambda(x^*)$,*

$$\langle F(x^*), x^* - y^* \rangle = 0 \text{ and } \langle F(y^*), x^* - y^* \rangle = 0.$$

(v) *$F(y^*) = kF(x^*)$ for each $x^* \in C^* \cup C_*$ and each $y^* \in \Gamma(x^*) \cup \Lambda(x^*)$.*

Hence, if also (iv) $\Rightarrow$ (v), *then* (i)–(v) *are all equivalent.*

Proof. By Proposition 2.8, we already have (v) $\Rightarrow$ (i) $\Rightarrow$ (iii) $\Rightarrow$ (iv). In addition, (i) $\Rightarrow$ (ii) is a direct result of Proposition 3.1(ii) and Theorem 4.1(iv). So it remains to show that (ii) $\Rightarrow$ (iv).

Let (ii) hold. Then, for each $x^* \in C^* \cup C_*$,

$$x^* \in C^* = C_* \text{ and } \Gamma(x^*) \cup \Lambda(x^*) = \Gamma(x^*) \cap \Lambda(x^*).$$

Thus, for each $y^* \in \Gamma(x^*) \cup \Lambda(x^*)$, the two equalities in (iv) hold, and hence (iv) follows. □

Remark 4.3. When F is pseudomonotone$_*$ on C, since the assumption in Theorem 4.4 holds, the statements (i)–(v) are equivalent. In particular, if F is continuous and pseudomonotone$^+_*$ on C, then C^* has the minimum principle sufficiency property and maximum principle sufficiency property iff (3) holds for each $x^* \in C^*$ and each $y^* \in \Gamma(x^*) \cup \Lambda(x^*)$ iff F equals $F(x^*)$ over $\Gamma(x^*) \cup \Lambda(x^*)$ for each $x^* \in C^*$.

5. Weak sharpness of C^*. In [9], Marcotte and Zhu showed that C^* is weakly sharp iff the dual gap function G has an error bound on C under the assumption that C is compact and F is continuous and pseudomonotone$^+$ over C. In this section, we present several equivalent conditions for C^* to be weakly sharp in terms of error bounds of G. The methods we use are similar to those of Marcotte and Zhu [9].

THEOREM 5.1. *Let G be Gâteaux differentiable and locally Lipschitz on C^*. Suppose that each $x^* \in C^* \cup C_*$ and each $y^* \in \Lambda(x^*)$ satisfy* (3) *and*

$$\langle F(x^*), x^* - y^* \rangle = 0 \text{ and } \langle F(y^*), x^* - y^* \rangle = 0 \Rightarrow F(y^*) = F(x^*).$$

Then C^* is weakly sharp iff there exists $\mu > 0$ such that

$$d_{C^*}(x) \leq \mu G(x) \text{ for each } x \in C. \tag{4}$$

Proof. Suppose that each $x^* \in C^* \cup C_*$ and each $y^* \in \Lambda(x^*)$ satisfy (3) and

$$\langle F(x^*), x^* - y^* \rangle = 0 \text{ and } \langle F(y^*), x^* - y^* \rangle = 0 \Rightarrow F(y^*) = F(x^*).$$

Then, by Theorem 4.2, $C^* = C_*$ and

$$F(y^*) = F(x^*) \text{ for each } y^* \in \Lambda(x^*) = C^* \text{ and each } x^* \in C^*.$$

If the set C^* is weakly sharp, then there exists $\alpha > 0$ such that

$$\alpha \overline{B} \subseteq F(x^*) + [T_C(x^*) \cap N_{C^*}(x^*)]^\circ \text{ for each } x^* \in C^*,$$

where $\overline{B}$ is the closed unit ball in H. This, as proved in [9], is equivalent to saying that

$$\langle F(x^*), v \rangle \geq \alpha \|v\| \text{ for each } v \in T_C(x^*) \cap N_{C^*}(x^*).$$

Now for each $x \in C$, since $C^* (= C_*)$ is convex, there exists $\overline{x} \in C^*$ such that

$$\|x - \overline{x}\| = d_{C^*}(x),$$

which implies that $x - \overline{x} \in T_C(\overline{x}) \cap N_{C^*}(\overline{x})$. Thus

$$G(x) \geq \langle F(\overline{x}), x - \overline{x} \rangle \geq \alpha \|x - \overline{x}\| = \alpha d_{C^*}(x).$$

Taking $\mu = \alpha^{-1}$, we obtain (4).

Conversely, suppose that (4) is satisfied for some $\mu > 0$. We claim that for $\alpha = \mu^{-1}$ there holds

$$\alpha B \subseteq F(x^*) + [T_C(x^*) \cap N_{C^*}(x^*)]^\circ \text{ for each } x^* \in C^*. \tag{5}$$

This is obvious for each point $x^* \in C^*$ satisfying $T_C(x^*) \cap N_{C^*}(x^*) = \{0\}$.

Next we show that (5) still holds for any x^* in C^* with $T_C(x^*) \cap N_{C^*}(x^*) \neq \{0\}$.

Let $x^* \in C^*$ and $0 \neq v \in T_C(x^*) \cap N_{C^*}(x^*)$. Then

$$\langle v, v \rangle > 0 \text{ and } \langle v, y^* - x^* \rangle \leq 0 \text{ for each } y^* \in C^*.$$

This implies that C^* is separated from $x^* + v$ by the hyperplane

$$H_v := \{x \in H : \langle v, x - x^* \rangle = 0\}.$$

In addition, for each sequence $\{t_k\}$ in $(0, +\infty)$ decreasing to 0, by [2, Theorem 2.4.5], there exists a sequence $\{v_k\}$ such that $v_k \to v$ and $x^* + t_k v_k \in C$ for sufficiently large k. Thus $\langle v, v_k \rangle > 0$ for sufficiently large k, and hence we can assume that $x^* + t_k v_k$ lies in the open set $\{x \in H : \langle v, x - x^* \rangle > 0\}$ which is separated from C^* by H_v. So

$$d_{C^*}(x^* + t_k v_k) \geq d_{H_v}(x^* + t_k v_k) = \frac{t_k \langle v_k, v \rangle}{\|v\|}.$$

Since G is Lipschitz near x^* and $G(x^*) = 0$,

$$\begin{aligned} \langle \nabla G(x^*), v \rangle &= \lim_{k \to +\infty} \frac{G(x^* + t_k v_k) - G(x^*)}{t_k} \\ &\geq \liminf_{k \to +\infty} \frac{\alpha d_{C^*}(x^* + t_k v_k)}{t_k} \geq \alpha \|v\|. \end{aligned}$$

Note that $\nabla G(x^*) = F(x^*)$ due to the Gâteaux differentiability of G at x^* and the inequality

$$G(x) \geq \langle F(x^*), x - x^* \rangle \text{ for all } x \in H.$$

For each $u \in B$ we have

$$\langle \alpha u - F(x^*), v \rangle = \langle \alpha u, v \rangle - \langle \nabla G(x^*), v \rangle \leq \alpha \|v\| - \alpha \|v\| = 0.$$

This implies that (5) is valid. Since F is constant over C^*, C^* is weakly sharp. □

Recall that for a lower semicontinuous convex function $f : H \to (-\infty, +\infty]$ the *subdifferential* of f at x in the sense of convex analysis is the set

$$\partial f(x) := \{\xi \in H : \langle \xi, y - x \rangle \leq f(y) - f(x) \text{ for each } y \in H\}.$$

It is easy to see that the *indicator function* associated with a convex set C given by

$$\psi_C(x) := \begin{cases} 0 & \text{if } x \in C, \\ +\infty & \text{if } x \notin C \end{cases}$$

satisfies $\partial \psi_C(x) = N_C(x)$ for $x \in C$.

Using the concept of the above subdifferential, Wu and Ye [11] showed that a proper lower semicontinuous convex function f satisfies

$$f(x) \geq \alpha d_S(x) \text{ for some } \alpha > 0 \text{ and each } x \in H$$

iff $\|\xi\| \geq \alpha$ for each $\xi \in \partial f(x)$ and each $x \in f^{-1}(0, +\infty)$, where $S := \{x \in H : f(x) \leq 0\}$. Based on this result and Theorem 5.1, we present another character of weak sharp solutions of the VIP(C, F) as follows.

THEOREM 5.2. *Let G be Gâteaux differentiable and locally Lipschitz on C^*. Suppose that each $x^* \in C^* \cup C_*$ and each $y^* \in \Lambda(x^*)$ satisfy* (3) *and*

$$\langle F(x^*), x^* - y^* \rangle = 0 \text{ and } \langle F(y^*), x^* - y^* \rangle = 0 \Rightarrow F(y^*) = F(x^*).$$

Then the following are equivalent:

(i) *C^* is weakly sharp.*

(ii) *$\Gamma(x^*) = C^*$ and there exists $\alpha > 0$ such that*

$$\alpha B \cap [F(x^*) + N_C(x)] = \emptyset \text{ for each } x^* \in C^* \text{ and each } x \in C \backslash C^*. \quad (6)$$

Proof. Based on Theorem 4.2,

$$C^* = \Lambda(x^*) = C_* \text{ and } F(y^*) = F(x^*) \text{ for each } x^* \in C^* \text{ and each } y^* \in \Lambda(x^*).$$

(i) ⇒ (ii): Let C^* be weakly sharp. Then, by Theorem 3.2, $\Gamma(x^*) = C_* = C^*$ for each $x^* \in C^*$. In addition, as in the proof of Theorem 5.1, there exists $\alpha > 0$ such that, for each $x^* \in C^*$,

$$\langle F(x^*), v \rangle \geq \alpha \|v\| \text{ for each } v \in T_C(x^*) \cap N_{C^*}(x^*).$$

Since for each $x \in C$ there exists $\overline{x} \in C^*$ such that $\|x - \overline{x}\| = d_{C^*}(x)$, which implies that $x - \overline{x} \in T_C(\overline{x}) \cap N_{C^*}(\overline{x})$,

$$\langle F(\overline{x}), x - \overline{x} \rangle \geq \alpha \|x - \overline{x}\| = \alpha d_{C^*}(x).$$

Since $F(x^*) = F(y^*)$ and $\langle F(x^*), x^* - y^* \rangle = 0$ for any $x^*, y^* \in C^* = C_*$, the lower semicontinuous convex function

$$f(x) := \langle F(x^*), x - x^* \rangle + \psi_C(x) \text{ for } x \in H$$

is well defined and satisfies

$$f(x) \geq \alpha d_{C^*}(x) \text{ for each } x \in H.$$

Denote $S := \{x \in H : f(x) \leq 0\}$. It is easy to see that

$$S = \{x \in C : \langle F(x^*), x - x^* \rangle \leq 0\} = \Gamma(x^*) = C^*.$$

So we have

$$f(x) \geq \alpha d_S(x) \text{ for each } x \in H.$$

Thus it follows from [11, Theorem 7] that

$$\|\xi\| \geq \alpha \text{ for each } \xi \in \partial f(x) \text{ and each } x \in f^{-1}(0, +\infty) = C \backslash C^*.$$

This implies that (ii) holds since $\partial f(x) = F(x^*) + \partial \psi_C(x) = F(x^*) + N_C(x)$.

(ii) $\Rightarrow$ (i): Suppose that $\Gamma(x^*) = C^*$ for each $x^* \in C^*$ and there exists $\alpha > 0$ such that (6) holds. Then, for the above f,

$$\|\xi\| \geq \alpha \text{ for each } \xi \in \partial f(x) \text{ and each } x \in f^{-1}(0, +\infty).$$

For each $x^* \in C^*$, by [11, Theorem 7] again, we have

$$f(x) \geq \alpha d_S(x) = \alpha d_{\Gamma(x^*)}(x) = \alpha d_{C^*}(x) \text{ for each } x \in H,$$

that is,

$$\langle F(x^*), x - x^* \rangle \geq \alpha d_{C^*}(x) \text{ for each } x^* \in C^* \text{ and each } x \in C.$$

Thus

$$G(x) \geq \langle F(x^*), x - x^* \rangle \geq \alpha d_{C^*}(x) \text{ for each } x \in C.$$

Hence, by Theorem 5.1, C^* is weakly sharp. □

Next we apply Theorems 5.1 and 5.2 to the VIP(C, F) in which F is pseudomonotone$^+$ to obtain the following equivalent condition for the weak sharpness of C^*.

THEOREM 5.3. *Let each $x^* \in C_*$ and each $y^* \in \Lambda(x^*)$ satisfy*

$$\langle F(x^*) - F(y^*), x^* - y^* \rangle = 0.$$

Suppose that F is pseudomonotone$^+$ over C and that G is Gâteaux differentiable and locally Lipschitz on C^*. Then the following are equivalent:

(i) *C^* is weakly sharp.*

(ii) *There exists $\mu > 0$ such that $d_{C^*}(x) \leq \mu G(x)$ for each $x \in C$.*

(iii) *$\Gamma(x^*) = C^*$ and there exists $\alpha > 0$ such that*

$$\alpha B \cap [F(x^*) + N_C(x)] = \emptyset \text{ for each } x^* \in C^* \text{ and each } x \in C \backslash C^*.$$

If C is also a polyhedral set in R^n, then (i)–(iii) *are all equivalent to*

(iv) *$\Gamma(x^*) = C^*$ for each $x^* \in C^*$.*

Proof. Based on Theorems 5.1 and 5.2, it suffices to verify the condition in Theorem 5.1 for the equivalences (i) $\Leftrightarrow$ (ii) $\Leftrightarrow$ (iii).

Since F is pseudomonotone$^+$, $C^* \subseteq C_* = C^* \cup C_*$. By Theorem 4.2, we need to show only that each $x^* \in C_*$ and each $y^* \in \Lambda(x^*)$ satisfy

$$\langle F(x^*), x^* - y^* \rangle = 0 \text{ and } \langle F(y^*), x^* - y^* \rangle = 0.$$

But this is obvious according to Proposition 2.4 and the assumption.

Now if $H = R^n$ and C is a polyhedral set, then for each $x^* \in C^*$ the linear program

$$\min\{\langle F(x^*), x - x^* \rangle : x \in C\}$$

has the solution set $\Gamma(x^*)$ and, according to [7, Lemma 1], there exists $\alpha > 0$ such that

$$\langle F(x^*), x - \overline{x} \rangle \geq \alpha \|x - \overline{x}\| \text{ for each } x \in C,$$

where $\overline{x} \in \Gamma(x^*)$ and $\|x - \overline{x}\| = d_{\Gamma(x^*)}(x)$. If $\Gamma(x^*) = C^*$, then

$$\langle F(x^*), x - x^* \rangle \geq \alpha d_{C^*}(x) \text{ for each } x \in C,$$

that is,

$$\langle F(x^*), x - x^* \rangle + \psi_C(x) \geq \alpha d_{C^*}(x) \text{ for each } x \in R^n.$$

Based on [11, Theorem 7], we obtain

$$\alpha B \cap [F(x^*) + N_C(x)] = \emptyset \text{ for each } x \in C \backslash C^*.$$

This shows that (iv) implies (iii), and hence (iii) $\Leftrightarrow$ (iv). Therefore (i)–(iv) are equivalent. ∎

Remark 5.1. (i) Under the assumption in Theorem 5.2 or 5.3,

$$C^* = \Lambda(x^*) \text{ for each } x^* \in C^*,$$

so Theorems 5.2 and 5.3 also hold with $\Gamma(x^*) = C^*$ replaced with $\Gamma(x^*) = \Lambda(x^*)$.

(ii) Unlike [9, Theorem 4.1], we do not assume the compactness of C in Theorems 5.1 and 5.3. However, if $C \subseteq R^n$ is compact and F is continuous and pseudomonotone$^+$ on C, then the conditions in Theorem 5.3 are satisfied since in this case $C^* = C_*$, G is Gâteaux differentiable (see [9, Theorem 3.1]) and locally Lipschitz on C^*, and $\langle F(y^*), x^* - y^* \rangle = 0$ for $x^* \in C_*$ and $y^* \in \Lambda(x^*)$ (which together with the pseudomonotonicity of F means $\langle F(x^*), x^* - y^* \rangle = 0$). Therefore, Theorems 5.1 and 5.3 extend [9, Theorem 4.1]. In particular, when C is a polyhedral, Theorem 5.3 covers [9, Theorem 4.3], which states that C^* is weakly sharp iff the VIP(C, F) possesses the minimum principle sufficiency property.

Note that in the case $C = H$ the equivalence in Theorem 5.1 means that G has an error bound on H iff

$$-F(x^*) \in \text{int} \bigcap_{x \in C^*} [N_{C^*}(x^*)]^\circ = \text{int} \bigcap_{x \in C^*} T_{C^*}(x) \text{ for each } x^* \in C^*,$$

which is in turn sufficient for C^* to be weakly sharp whether $C = H$ or not, since for $x^* \in C^*$ we have

$$T_{C^*}(x^*) = [N_{C^*}(x^*)]^\circ \subseteq [T_C(x^*) \cap N_{C^*}(x^*)]^\circ.$$

The following result states that the above equivalence is still valid for any nonempty convex set C in H, and hence the two equivalent statements are both sufficient for the weak sharpness of C^*.

THEOREM 5.4. *Let G be Gâteaux differentiable on C^*. Suppose that each $x^* \in C^* \cup C_*$ and each $y^* \in \Lambda(x^*)$ satisfy* (3) *and*

$$\langle F(x^*), x^* - y^* \rangle = 0 \text{ and } \langle F(y^*), x^* - y^* \rangle = 0 \Rightarrow F(y^*) = F(x^*).$$

Then the following are equivalent:

(i) $-F(x^*) \in \operatorname{int} \bigcap_{x \in C^*} T_{C^*}(x)$ *for each* $x^* \in C^*$.

(ii) *There exists* $\mu > 0$ *such that*

$$d_{C^*}(x) \le \mu G(x) \text{ for each } x \in H. \tag{7}$$

Proof. As in the proof of Theorem 5.1, we have $C^* = C_*$ and

$$F(y^*) = F(x^*) \text{ for each } y^* \in \Lambda(x^*) = C^* \text{ and each } x^* \in C^*.$$

If (i) holds, then there exists $\alpha > 0$ such that

$$\alpha \overline{B} \subseteq F(x^*) + T_{C^*}(x^*) = F(x^*) + [N_{C^*}(x^*)]^\circ \text{ for each } x^* \in C^*,$$

where $\overline{B}$ is the closed unit ball in H. Thus for each $x^* \in C^*$ and every $u \in \overline{B}$ we have

$$\alpha u - F(x^*) \in [N_{C^*}(x^*)]^\circ.$$

Therefore

$$\langle \alpha u - F(x^*), v \rangle \le 0 \text{ for each } v \in N_{C^*}(x^*).$$

Taking $u = v/\|v\|$ for $v \ne 0$ yields

$$\langle F(x^*), v \rangle \ge \alpha \|v\| \text{ for each } v \in N_{C^*}(x^*).$$

Now for each $x \in H$, since C^* is convex, there exists $\overline{x} \in C^*$ such that

$$\|x - \overline{x}\| = d_{C^*}(x),$$

which implies that $x - \overline{x} \in N_{C^*}(\overline{x})$. Thus

$$G(x) \ge \langle F(\overline{x}), x - \overline{x} \rangle \ge \alpha \|x - \overline{x}\| = \alpha d_{C^*}(x).$$

Taking $\mu = \alpha^{-1}$, we obtain (7).

Conversely, suppose that (7) is satisfied for some $\mu > 0$. To obtain (i), it suffices to show that for $\alpha = \mu^{-1}$ there holds

$$\alpha \overline{B} \subseteq F(x^*) + T_{C^*}(x^*) \text{ for each } x^* \in C^*. \tag{8}$$

Obviously, if $x^* \in C^*$ satisfies $N_{C^*}(x^*) = \{0\}$, then (8) holds.

Next we show that (8) also holds for any x^* in C^* with $N_{C^*}(x^*) \ne \{0\}$.

Let $x^* \in C^*$ and $0 \ne v \in N_{C^*}(x^*)$. Then

$$\langle v, v \rangle > 0 \text{ and } \langle v, y^* - x^* \rangle \le 0 \text{ for each } y^* \in C^*.$$

Hence C^* is separated from $x^* + v$ by the hyperplane H_v passing through x^* and orthogonal to v. In addition, for each sequence $\{t_k\}$ in $(0, +\infty)$ decreasing to 0, x^*+t_kv lies in the open set $\{x \in H : \langle v, x - x^*\rangle > 0\}$ which is separated from C^* by H_v, so we have

$$d_{C^*}(x^* + t_kv) \geq d_{H_v}(x^* + t_kv) = t_k\|v\|,$$

from which and from the inequality $G(x^* + t_kv) \geq \alpha d_{C^*}(x^* + t_kv)$ it follows that

$$\langle \nabla G(x^*), v\rangle = \lim_{k\to+\infty} \frac{G(x^* + t_kv) - G(x^*)}{t_k} \geq \alpha\|v\|.$$

Thus for each $u \in \overline{B}$ we have

$$\langle \alpha u - F(x^*), v\rangle = \langle \alpha u, v\rangle - \langle \nabla G(x^*), v\rangle \leq \alpha\|v\| - \alpha\|v\| = 0.$$

This implies that (8) holds since $0 \neq v \in N_{C^*}(x^*)$ is arbitrary and F is constant over C^*. □

To conclude this paper, we derive a finite convergence result for a class of algorithms for solving variational inequalities under the condition that C^* be weakly sharp.

THEOREM 5.5. *Let F be continuous on C_* and let C^* be weakly sharp. Suppose that each $x^* \in C^*$ and each $y^* \in \Lambda(x^*)$ satisfy* (3) *and*

$$\langle F(x^*), x^* - y^*\rangle = 0 \text{ and } \langle F(y^*), x^* - y^*\rangle = 0 \Rightarrow F(y^*) = F(x^*).$$

If $\{x_k\}$ is a sequence in H satisfying either

(i) *$d_{C^*}(x_k) \to 0$ and F is uniformly continuous on C^*, or*

(ii) *x_k converges to some $x^* \in C^*$,*

then $\Gamma(x_k) \subseteq C^*$ for sufficiently large k.

Proof. For any fixed $x^* \in C^*$, under the given conditions, it follows from Theorem 4.2 that $C^* = \Lambda(x^*) = C_*$ and $F(y^*) = F(x^*)$ for each $y^* \in C^*$. Since the set C^* is weakly sharp, that is,

$$-F(x^*) \in \operatorname{int} \bigcap_{x\in C^*} [T_C(x) \cap N_{C^*}(x)]^\circ \text{ for each } x^* \in C^*,$$

there exists $\alpha > 0$ such that

$$-F(x^*) + \alpha\overline{B} \subseteq \bigcap_{x\in C^*} [T_C(x) \cap N_{C^*}(x)]^\circ \text{ for each } x^* \in C^*,$$

where $\overline{B}$ is the closed unit ball in H.

If $\{x_k\}$ is a sequence satisfying (i), then, since C_* is convex, there exists unique $x_k^* \in C_* = C^*$ such that $\|x_k - x_k^*\| = d_{C^*}(x_k)$ and, by the uniform continuity of F on C^*,

$$\|F(x_k) - F(x^*)\| = \|F(x_k) - F(x_k^*)\| < \alpha \text{ for sufficiently large } k,$$

that is,

$$-F(x_k) \in \operatorname{int} \bigcap_{x\in C^*} [T_C(x) \cap N_{C^*}(x)]^\circ \text{ for sufficiently large } k.$$

Obviously, this still holds if $\{x_k\}$ is a sequence satisfying (ii). Therefore we obtain from Theorem 3.2 that $\Gamma(x_k) \subseteq C^*$ for sufficiently large k. □

Remark 5.2. Motivated by [9, Theorem 5.1], Theorem 5.5 has the same conclusion as in that theorem. However, in [9, Theorem 5.1], F is assumed to be not only continuous and pseudomonotone$^+$ on $C \subseteq R^n$ but also uniformly continuous on an open set containing C^* and $\{x_k\}$, so the condition in Theorem 5.5 is weaker.

Acknowledgments. We thank Professor Steve Cox and the anonymous referee for their valuable comments and suggestions.

REFERENCES

[1] J. V. Burke and M. C. Ferris, *Weak sharp minima in mathematical programming*, SIAM J. Control Optim., 31 (1993), pp. 1340–1359.

[2] F. H. Clarke, *Optimization and Nonsmooth Analysis*, Wiley-Interscience, New York, 1983; reprinted as Classics Appl. Math. 5, SIAM, Philadelphia, 1990.

[3] J. P. Crouzeix, P. Marcotte, and D. L. Zhu, *Conditions ensuring the applicability of cutting-plane methods for solving variational inequalities*, Math. Programming, 88 (2000), pp. 521–539.

[4] M. C. Ferris and O. L. Mangasarian, *Minimum principle sufficiency*, Math. Program., 57 (1992), pp. 1–14.

[5] P. T. Harker and J. S. Pang, *Finite-dimensional variational inequality and nonlinear complementarity problems: A survey of theory, algorithms and applications*, Math. Program., 48 (1990), pp. 161–220.

[6] D. W. Hearn, *The gap function of a convex program*, Oper. Res. Lett., 1 (1982), pp. 67–71.

[7] O. L. Mangasarian and R. R. Meyer, *Nonlinear perturbation of linear programs*, SIAM J. Control Optim., 17 (1979), pp. 745–752.

[8] T. Larsson and M. Patriksson, *A class of gap functions for variational inequalities*, Math. Program., 64 (1994), pp. 53–79.

[9] P. Marcotte and D. L. Zhu, *Weak sharp solutions of variational inequalities*, SIAM J. Optim., 9 (1998), pp. 179–189.

[10] M. Patriksson, *A Unified Framework of Descent Algorithms for Nonlinear Programs and Variational Inequalities*, Ph.D. thesis, Department of Mathematics, Linköping Institute of Technology, Linköping, Sweden, 1993.

[11] Z. L. Wu and J. J. Ye, *On error bounds for lower semicontinuous functions*, Math. Programming, 92 (2002), pp. 301–314.

SIAM J. OPTIM.
Vol. 14, No. 4, pp. 1028–1042

THE FINITE SAMPLE BREAKDOWN POINT OF ℓ_1-REGRESSION*

AVI GILONI† AND MANFRED PADBERG‡

Abstract. Through a new (parametric) linear programming approach, we derive a formula for the finite sample breakdown point of ℓ_1-regression with a given design matrix $\mathbf{X}$ and contamination restricted to the dependent variable. This is done using the notion of the q-strength and the s-stability of a design matrix $\mathbf{X}$, which are introduced here. We discuss the relationship between our result and existing results in the literature. Finally, we demonstrate the usefulness of our result by calculating (via the solution of mixed-integer programs) the finite sample breakdown point of ℓ_1-regression with contamination restricted to the dependent variable for nine well-known data sets from the robust regression literature.

Key words. ℓ_1-regression, breakdown point, robust designs, robustness, linear programming, mixed-integer programming

AMS subject classifications. 62J99, 62F35, 90C05, 90C11

DOI. 10.1137/S1052623403424156

1. Introduction.

In this paper we discuss the finite sample breakdown point of the ℓ_1-regression estimator, with a fixed design matrix $\mathbf{X}$ and contamination restricted to the dependent variable $\mathbf{y}$, which we denote by $\alpha(\ell_1, \mathbf{y}|\mathbf{X})$ to indicate that the design matrix $\mathbf{X}$ is given. The finite sample breakdown point, or conditional breakdown point, $\alpha(\ell_1, \mathbf{y}|\mathbf{X})$, was introduced by Donoho and Huber [2]. It is especially important in *planned experiments*, where the design matrix is under the control of the experimenter. Its study has been addressed by, among others, He et al. [7], Ellis and Morgenthaler [6], and Mizera and Müller [11, 12]. We introduce the notions of the q-strength and the s-stability of $\mathbf{X}$ based on a parametric linear programming (LP) approach to the problem, which permits us to derive a formula for $\alpha(\ell_1, \mathbf{y}|\mathbf{X})$. We show that our result is consistent with earlier results. The advantage of our framework is that it permits us to compute the breakdown value via the solution of a mixed-integer program (MIP). We present computational results for nine data sets from the robust regression literature.

1.1. ℓ_1-regression.

In linear regression we have n observations on some "dependent" variable y and some number $p \geq 1$ of "independent" variables $x_1, \ldots, x_p$, for each of which we know n values as well. We denote

$$\mathbf{y} = \begin{pmatrix} y_1 \\ \cdot \\ \cdot \\ \cdot \\ y_n \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} x_1^1 & \cdot & \cdot & \cdot & x_p^1 \\ \cdot & & & & \cdot \\ \cdot & & & & \cdot \\ \cdot & & & & \cdot \\ x_1^n & \cdot & \cdot & \cdot & x_p^n \end{pmatrix} = \begin{pmatrix} \mathbf{x}^1 \\ \cdot \\ \cdot \\ \cdot \\ \mathbf{x}^n \end{pmatrix} = (\mathbf{x}_1, \ldots, \mathbf{x}_p), \tag{1}$$

where $\mathbf{y} \in \mathbb{R}^n$ is a vector of n observations and $\mathbf{X}$ is an $n \times p$ matrix of reals referred to as the design matrix. $\mathbf{x}_1, \ldots, \mathbf{x}_p$ are column vectors with n components, and

*Received by the editors March 12, 2003; accepted for publication (in revised form) October 16, 2003; published electronically May 25, 2004. This work was supported in part by a grant from the Office of Naval Research (N00014-96-0327).
http://www.siam.org/journals/siopt/14-4/42415.html
†Sy Syms School of Business, Yeshiva University, 500 West 185th Street, BH-428, New York, NY, 10033 (agiloni@ymail.yu.edu).
‡17, Rue Vendome, 13007 Marseille, France (manfred@padberg.com).

$\mathbf{x}^1, \ldots, \mathbf{x}^n$ are row vectors with p components corresponding to the columns and rows of $\mathbf{X}$, respectively. To rule out pathologies we assume throughout that the rank $r(\mathbf{X})$ of $\mathbf{X}$ is full, i.e., that $r(\mathbf{X}) = p$.

The statistical linear regression model is $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where $\boldsymbol{\beta}^T = (\beta_1, \ldots, \beta_p)$ is the vector of *parameters* of the linear model and $\boldsymbol{\varepsilon}^T = (\varepsilon_1, \ldots, \varepsilon_n)$ a vector of n random variables corresponding to the error terms in the asserted relationship. An upper index T denotes "transposition" of a vector or matrix throughout this work. In the statistical model, the dependent variable y is a random variable for which we obtain measurements or observations that contain some "noise" or measurement errors that are captured in the error terms $\boldsymbol{\varepsilon}$. For the numerical problem that we are facing we write

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{r}, \tag{2}$$

where, given some parameter vector $\boldsymbol{\beta}$, the components r_i of the vector $\mathbf{r}^T = (r_1, \ldots, r_n)$ are the *residuals* that result, given the observations $\mathbf{y}$, a fixed design matrix $\mathbf{X}$, and the chosen vector $\boldsymbol{\beta} \in \mathbb{R}^p$. In the case of ℓ_1-regression, the (optimal) parameters $\boldsymbol{\beta} \in \mathbb{R}^p$ are those that minimize the ℓ_1-norm $\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_1 = \sum_{i=1}^n |y_i - \mathbf{x}^i\boldsymbol{\beta}|$ of the residuals.

The ℓ_1-regression problem can be formulated as the LP problem

$$\begin{aligned} &\min\ \mathbf{e}_n^T\mathbf{r}^+ + \mathbf{e}_n^T\mathbf{r}^- \\ \text{s.t.}\quad &\mathbf{X}\boldsymbol{\beta} + \mathbf{r}^+ - \mathbf{r}^- = \mathbf{y}, \\ &\boldsymbol{\beta} \text{ free}, \ \mathbf{r}^+ \geq \mathbf{0}, \ \mathbf{r}^- \geq \mathbf{0}, \end{aligned} \tag{3}$$

where $\mathbf{e}_n$ is the vector with all n components equal to 1. In (3) the residuals $\mathbf{r}$ of the general form (2) are simply replaced with a difference $\mathbf{r}^+ - \mathbf{r}^-$ of nonnegative variables; i.e., we require that $\mathbf{r}^+ \geq \mathbf{0}$ and $\mathbf{r}^- \geq \mathbf{0}$, whereas the parameters $\boldsymbol{\beta} \in \mathbb{R}^p$ are "free" to assume positive, zero, or negative values. From the properties of LP solution procedures, it follows that in any solution inspected by, e.g., the simplex algorithm, either $r_i^+ > 0$ or $r_i^- > 0$, but not both, thus giving $|r_i|$ in the objective function depending on whether $r_i > 0$ or $r_i < 0$ for any $i \in N$, where $N = \{1, \ldots, n\}$.

To characterize the optimality of coefficients $\boldsymbol{\beta} \in \mathbb{R}^p$ for ℓ_1-regression let

$$Z_{\boldsymbol{\beta}} = \left\{ i \in N : r_i^{\boldsymbol{\beta}} = 0 \right\}, \ U_{\boldsymbol{\beta}} = \left\{ i \in N : r_i^{\boldsymbol{\beta}} > 0 \right\}, \ L_{\boldsymbol{\beta}} = \left\{ i \in N : r_i^{\boldsymbol{\beta}} < 0 \right\}, \tag{4}$$

where $r_i^{\boldsymbol{\beta}} = y_i - \mathbf{x}^i\boldsymbol{\beta}$ for all $i \in N$. Let $\mathbf{X}_Z = (\mathbf{x}^i)_{i \in Z}$, $\mathbf{e}_Z = (1, \ldots, 1)^T$ with $|Z|$ components equal to 1 (i.e., $|Z|$ is the cardinality of the set Z). $\mathbf{X}_U$, $\mathbf{e}_U$, $\mathbf{r}_U$, $\mathbf{X}_L$, $\mathbf{e}_L$, and $\mathbf{r}_L$ are defined likewise.

THEOREM 1. *Let $\boldsymbol{\beta} \in \mathbb{R}^p$ and $Z_{\boldsymbol{\beta}}, U_{\boldsymbol{\beta}}, L_{\boldsymbol{\beta}}$ be as defined in* (4). *$\boldsymbol{\beta}$ is an optimal solution to $\min_{\boldsymbol{\beta}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_1$ if and only if there exists $\mathbf{v} \in \mathbb{R}^{|Z_\beta|}$ such that*

$$\mathbf{v}\mathbf{X}_{Z_{\boldsymbol{\beta}}} = -\mathbf{e}_{U_{\boldsymbol{\beta}}}^T\mathbf{X}_{U_{\boldsymbol{\beta}}} + \mathbf{e}_{L_{\boldsymbol{\beta}}}^T\mathbf{X}_{L_{\boldsymbol{\beta}}}, \qquad -\mathbf{e}_{Z_{\boldsymbol{\beta}}}^T \leq \mathbf{v} \leq \mathbf{e}_{Z_{\boldsymbol{\beta}}}^T, \tag{5}$$

i.e., if and only if (5) *is solvable.*

Proof. The dual linear program to (3) is given by

$$\max\left\{\mathbf{u}\mathbf{y} : \mathbf{u}\mathbf{X} = \mathbf{0}, \ -\mathbf{e}_n^T \leq \mathbf{u} \leq \mathbf{e}_n^T\right\} = \max\left\{\mathbf{u}\mathbf{r} : \mathbf{u}\mathbf{X} = \mathbf{0}, \ -\mathbf{e}_n^T \leq \mathbf{u} \leq \mathbf{e}_n^T\right\},$$

where the equality follows because $\mathbf{u}\mathbf{y} = \mathbf{u}(\mathbf{X}\boldsymbol{\beta} + \mathbf{r}) = \mathbf{u}\mathbf{r}$ for all $\mathbf{u} \in \mathbb{R}^n$ satisfying $\mathbf{u}\mathbf{X} = \mathbf{0}$. Suppose condition (5) is satisfied. Define $u_i = 1$ for $i \in U_{\boldsymbol{\beta}}$, $u_i = -1$ for

$i \in L_{\boldsymbol{\beta}}$, and $\mathbf{u}_{Z_{\boldsymbol{\beta}}} = \mathbf{v}$. Then $\mathbf{u}$ is a feasible solution to the dual, $\mathbf{ur} = \mathbf{e}_U^T\mathbf{r}_U - \mathbf{e}_L^T\mathbf{r}_L = \|\mathbf{r}\|_1$, and thus by the weak theorem of duality of LP, $\boldsymbol{\beta}$ is an optimal solution. Suppose $\boldsymbol{\beta}$ is an optimal solution to the ℓ_1-regression problem, but that $\mathbf{v} \in \mathbb{R}^{|Z_{\boldsymbol{\beta}}|}$ satisfying (5) does not exist. By Farkas's lemma (see, e.g., [1, Exercise 6.5] or [13]), there exist $\boldsymbol{\xi} \in \mathbb{R}^p$, $\boldsymbol{\eta}^+, \boldsymbol{\eta}^- \in \mathbb{R}^{|Z_{\boldsymbol{\beta}}|}$ such that

$$\begin{aligned}(6)\qquad &\mathbf{X}_{Z_{\boldsymbol{\beta}}}\boldsymbol{\xi} + \boldsymbol{\eta}^+ - \boldsymbol{\eta}^- = \mathbf{0},\ \left(-\mathbf{e}_{U_{\boldsymbol{\beta}}}^T\mathbf{X}_{U_{\boldsymbol{\beta}}} + \mathbf{e}_{L_{\boldsymbol{\beta}}}^T\mathbf{X}_{L_{\boldsymbol{\beta}}}\right)\boldsymbol{\xi} + \mathbf{e}_{Z_{\boldsymbol{\beta}}}^T\boldsymbol{\eta}^+ + \mathbf{e}_{Z_{\boldsymbol{\beta}}}^T\boldsymbol{\eta}^- < 0,\\ &\qquad\qquad \boldsymbol{\eta}^+ \geq \mathbf{0},\ \boldsymbol{\eta}^- \geq \mathbf{0}.\end{aligned}$$

If $Z_{\boldsymbol{\beta}} = \emptyset$, then $-\mathbf{e}_{U_{\boldsymbol{\beta}}}^T\mathbf{X}_{U_{\boldsymbol{\beta}}} + \mathbf{e}_{L_{\boldsymbol{\beta}}}^T\mathbf{X}_{L_{\boldsymbol{\beta}}} \neq \mathbf{0}$ since otherwise the dual linear program, and thus (5), has a solution. In this case we choose any $\boldsymbol{\xi} \in \mathbb{R}^p$ such that $(-\mathbf{e}_{U_{\boldsymbol{\beta}}}^T\mathbf{X}_{U_{\boldsymbol{\beta}}} + \mathbf{e}_{L_{\boldsymbol{\beta}}}^T\mathbf{X}_{L_{\boldsymbol{\beta}}})\boldsymbol{\xi} < 0$. Since $\mathbf{r}_{U_{\boldsymbol{\beta}}} > \mathbf{0}$ and $\mathbf{r}_{L_{\boldsymbol{\beta}}} < \mathbf{0}$, there exists $\lambda > 0$ such that $\mathbf{r}_{U_{\boldsymbol{\beta}}}^+(\lambda) = \mathbf{r}_{U_{\boldsymbol{\beta}}} - \lambda\mathbf{X}_{U_{\boldsymbol{\beta}}}\boldsymbol{\xi} \geq \mathbf{0}$ and $\mathbf{r}_{L_{\boldsymbol{\beta}}}^-(\lambda) = -\mathbf{r}_{L_{\boldsymbol{\beta}}} + \lambda\mathbf{X}_{L_{\boldsymbol{\beta}}}\boldsymbol{\xi} \geq \mathbf{0}$. Consequently, $\boldsymbol{\beta}(\lambda) = \boldsymbol{\beta} + \lambda\boldsymbol{\xi}$, $\mathbf{r}_{Z_{\boldsymbol{\beta}}}^{\pm}(\lambda) = \lambda\boldsymbol{\eta}^{\pm}$, $\mathbf{r}_{U_{\boldsymbol{\beta}}}^+(\lambda)$, $\mathbf{r}_{U_{\boldsymbol{\beta}}}^-(\lambda) = \mathbf{0}$, $\mathbf{r}_{L_{\boldsymbol{\beta}}}^+(\lambda) = \mathbf{0}$, and $\mathbf{r}_{L_{\boldsymbol{\beta}}}^-(\lambda)$ define a feasible solution to the linear program (3). Calculating its objective function we get

$$\begin{aligned}\mathbf{e}_N^T\mathbf{r}^+(\lambda) + \mathbf{e}_N^T\mathbf{r}^-(\lambda) &= \mathbf{e}_{U_{\boldsymbol{\beta}}}^T\mathbf{r}_{U_{\boldsymbol{\beta}}}^+(\lambda) + \mathbf{e}_{L_{\boldsymbol{\beta}}}^T\mathbf{r}_{L_{\boldsymbol{\beta}}}^-(\lambda) + \lambda\left(\mathbf{e}_{Z_{\boldsymbol{\beta}}}^T\boldsymbol{\eta}^+ + \mathbf{e}_{Z_{\boldsymbol{\beta}}}^T\boldsymbol{\eta}^-\right)\\ &= \|\mathbf{r}\|_1 + \lambda\left(-\mathbf{e}_{U_{\boldsymbol{\beta}}}^T\mathbf{X}_{U_{\boldsymbol{\beta}}}\boldsymbol{\xi} + \mathbf{e}_{L_{\boldsymbol{\beta}}}^T\mathbf{X}_{L_{\boldsymbol{\beta}}}\boldsymbol{\xi} + \mathbf{e}_{Z_{\boldsymbol{\beta}}}^T\boldsymbol{\eta}^+ + \mathbf{e}_{Z_{\boldsymbol{\beta}}}^T\boldsymbol{\eta}^-\right) < \|\mathbf{r}\|_1,\end{aligned}$$

and consequently $\boldsymbol{\beta}$ is not optimal. □

As one of the referees pointed out, a different proof of Theorem 1 can be obtained by applying Theorem 2.2.1 of [8, p. 253]. We leave the details to the interested reader.

2. The breakdown point of ℓ_1-regression.

2.1. Breakdown point. The notion of the *breakdown point* of a regression estimator due to Hampel [5] can be found, e.g., in [14], and reads as follows. Suppose we estimate the regression parameters $\boldsymbol{\beta}$ by some technique τ from some data $(\mathbf{X}, \mathbf{y})$ to be $\boldsymbol{\beta}^\tau$. If we replace any number $1 \leq m < n$ of the data with some arbitrary data $(\widetilde{\mathbf{x}}^i, \widetilde{y}_i)$, we obtain new data $(\widetilde{\mathbf{X}}, \widetilde{\mathbf{y}})$. The same technique τ applied to $(\widetilde{\mathbf{X}}, \widetilde{\mathbf{y}})$ yields estimates $\boldsymbol{\beta}^\tau(\widetilde{\mathbf{X}}, \widetilde{\mathbf{y}})$ that are different from the original ones. We can use any norm $\|\cdot\|$ on $\mathbb{R}^p$ to measure the distance $\|\boldsymbol{\beta}^\tau(\widetilde{\mathbf{X}}, \widetilde{\mathbf{y}}) - \boldsymbol{\beta}^\tau\|$ of the respective estimates. If we vary over *all* possible choices, then this distance remains either bounded or not bounded. Let

$$(7)\qquad b(m, \tau, \mathbf{X}, \mathbf{y}) = \sup_{\widetilde{\mathbf{X}}, \widetilde{\mathbf{y}}} \left\|\boldsymbol{\beta}^\tau\left(\widetilde{\mathbf{X}}, \widetilde{\mathbf{y}}\right) - \boldsymbol{\beta}^\tau\right\|$$

be the *maximum bias* that results when we replace at most m of the original data $(\mathbf{x}^i, y_i)$ with arbitrary new data. Let

$$(8)\qquad b(m, \tau, \mathbf{y}|\mathbf{X}) = \sup_{\widetilde{\mathbf{y}}} \left\|\boldsymbol{\beta}^\tau\left(\mathbf{X}, \widetilde{\mathbf{y}}\right) - \boldsymbol{\beta}^\tau\right\|$$

be the *maximum bias* that results when we replace at most m of the original values of the dependent variable y_i with arbitrary new data. The breakdown point of τ is

$$\alpha(\tau, \mathbf{X}, \mathbf{y}) = \min_{1 \leq m < n} \left\{\frac{m}{n} : b(m, \tau, \mathbf{X}, \mathbf{y}) \text{ is infinite}\right\};$$

i.e., it is the *minimum* number of rows of $(\mathbf{X}, \mathbf{y})$ that, if replaced with arbitrary new data, make the regression technique τ break down. The conditional breakdown point of τ is

$$\alpha(\tau, \mathbf{y}|\mathbf{X}) = \min_{1 \le m < n} \left\{ \frac{m}{n} : b(m, \tau, \mathbf{y}|\mathbf{X}) \text{ is infinite} \right\};$$

i.e., it is the *minimum* number of values of $\mathbf{y}$ that, if replaced with arbitrary new data, make the regression technique τ break down. We divide by n to get $\frac{1}{n} \le \alpha(\tau, \mathbf{X}, \mathbf{y}) \le 1$.

The breakdown point of ℓ_1-regression is $\frac{1}{n}$ or, asymptotically, 0; see, e.g., [14]. However, the determination of the conditional breakdown point $\alpha(\ell_1, \mathbf{y}|\mathbf{X})$ of ℓ_1-regression is not straightforward. He et al. [7] disprove a claim of Donoho and Huber [2, p. 166] that $\alpha(\ell_1, \mathbf{y}|\mathbf{X})$ is $\frac{1}{2}$ or 50%. This was observed in [3] independently of He et al.'s work.

Example 1. Let $n = 3$, $p = 2$ with data

$$(\mathbf{X}, \mathbf{y}) = \begin{pmatrix} 1 & 0 & 2 \\ 1 & 2 & 3 \\ 1 & -1 & 2 \end{pmatrix}, \tag{9}$$

and suppose that y_2 is contaminated. We replace $y_2 = 3$ with $y_2 = 3 + \vartheta$, where $\vartheta \ge 0$ is arbitrary. We calculate that the optimal ℓ_1-regression coefficients $\boldsymbol{\beta}(\vartheta)$ are $\beta_1(\vartheta) = \frac{7}{3} + \frac{\vartheta}{3}$, $\beta_2(\vartheta) = \frac{1}{3} + \frac{\vartheta}{3}$ for all $\vartheta \ge 0$. The optimal residuals are $r_1^-(\vartheta) = \frac{1}{3} + \frac{\vartheta}{3}$, and $r_i^+(\vartheta) = r_i^-(\vartheta) = 0$ otherwise. Thus $\|\boldsymbol{\beta}(\vartheta) - \boldsymbol{\beta}(0)\|_1 = \frac{2}{3}\vartheta \to +\infty$ for $\vartheta \to +\infty$, and thus a single contaminated observation in $\mathbf{y}$ may cause ℓ_1-regression to break down. □

This example can be generalized to higher dimensions. The idea that underlies the counterexample to Donoho and Huber's statement is the following: For any $\vartheta \ge 0$, the optimal basis of (3) corresponding to the data contains row $\mathbf{x}^2$ of the design matrix, but neither r_2^+ nor r_2^-. Hence the optimal $\boldsymbol{\beta}(\vartheta)$ depend on the amount ϑ of the contamination of y_2, and thus the maximum bias that results from the contamination grows beyond all bounds.

2.2. *q*-strength and *s*-stability of design matrices. To study $\alpha(\ell_1, \mathbf{y}|\mathbf{X})$, we consider the parametric linear program corresponding to (3),

$$z(\vartheta) = \min \left\{ \mathbf{e}_n^T \mathbf{r}^+ + \mathbf{e}_n^T \mathbf{r}^- : \mathbf{X}\boldsymbol{\beta} + \mathbf{r}^+ - \mathbf{r}^- = \mathbf{y} + \vartheta \mathbf{g},\ \mathbf{r}^+ \ge \mathbf{0},\ \mathbf{r}^- \ge \mathbf{0} \right\}, \tag{10}$$

where $\mathbf{g} \in \mathbb{R}^n$ is arbitrary and $\vartheta \ge 0$ is some parameter. Since $\mathbf{g} \in \mathbb{R}^n$ is arbitrary, the sign restriction on ϑ does not matter. By varying ϑ and $\mathbf{g}$, every possible contamination of the components of the observation vector $\mathbf{y} \in \mathbb{R}^n$ is obtained. It is well known (see, e.g., [13, p. 102]) that $z(\vartheta)$ is a convex, piecewise linear function of ϑ.

LEMMA 1. *Let* $\mathbf{P}^* = \{\mathbf{u} \in \mathbb{R}^n : \mathbf{uX} = \mathbf{0},\ -\mathbf{e}_n^T \le \mathbf{u} \le \mathbf{e}_n^T\}$. *Then*

$$0 \le z(\vartheta) = z(\vartheta_0) + (\vartheta - \vartheta_0)\mathbf{u}^0\mathbf{g} \le z(\vartheta_0) + (\vartheta - \vartheta_0)\|\mathbf{g}\|_1$$

for all $\vartheta \ge \vartheta_0$ *with some finite* $\vartheta_0 \ge 0$ *and* $\mathbf{u}^0$ *some extreme point of* $\mathbf{P}^*$.

Proof. From (10), $z(\vartheta) \ge 0$. Since $(\boldsymbol{\beta}, \mathbf{r}^+, \mathbf{r}^-) = (\mathbf{0}, \max\{\mathbf{0}, \mathbf{y} + \vartheta\mathbf{g}\}, -\min\{\mathbf{0}, \mathbf{y} + \vartheta\mathbf{g}\})$ is feasible for (10), it follows from the triangle inequality that $z(\vartheta) \le \|\mathbf{y} + \vartheta\mathbf{g}\|_1 \le \|\mathbf{y}\|_1 + \vartheta\|\mathbf{g}\|_1$ for all $\vartheta \ge 0$. Consequently, by the duality theorem of LP,

$$z(\vartheta) = \max\{\mathbf{u}(\mathbf{y} + \vartheta\mathbf{g}) : \mathbf{u} \in \mathbf{P}^*\}. \tag{11}$$

Since $\mathbf{0} \in \mathbf{P}^*$ and $\mathbf{P}^* \subseteq \{\mathbf{u} \in \mathbb{R}^n : -\mathbf{e}_n^T \leq \mathbf{u} \leq \mathbf{e}_n^T\}$, $\mathbf{P}^*$ is a nonempty polytope. Hence, $\mathbf{P}^* = \text{conv}\{\mathbf{u}^1, \ldots, \mathbf{u}^r\}$, where $\mathbf{u}^i$ is an extreme point of $\mathbf{P}^*$ and $r > 0$ a finite number. Thus

$$z(\vartheta) = \max\left\{\mathbf{u}^i(\mathbf{y} + \vartheta\mathbf{g}) : 1 \leq i \leq r\right\} = z(\vartheta_0) + (\vartheta - \vartheta_0)\mathbf{u}^0\mathbf{g} \tag{12}$$

for all $\vartheta \geq \vartheta_0$, where $\vartheta_0 \geq 0$ is some finite value of the parameter and $\mathbf{u}^0$ a corresponding optimal extreme point of $\mathbf{P}^*$ for $\vartheta = \vartheta_0$. Finally, $\mathbf{u}^0\mathbf{g} \leq \|\mathbf{g}\|_1$, since $-\mathbf{e}_n^T \leq \mathbf{u}^0 \leq \mathbf{e}_n^T$. ∎

To analyze $\alpha(\ell_1, \mathbf{y}|\mathbf{X})$, let

$$\mathcal{F}_q = \{(U, L, Z) : U \subseteq N,\ L \subseteq N - U,\ Z = N - (U \cup L),\ |U \cup L| = q\}.$$

We call a design matrix $\mathbf{X}$ *q-strong* if q is the largest integer such that

$$\mathbf{v}\mathbf{X}_Z = -\mathbf{e}_U^T\mathbf{X}_U + \mathbf{e}_L^T\mathbf{X}_L, \qquad -\mathbf{e}_Z^T \leq \mathbf{v} \leq \mathbf{e}_Z^T \tag{13}$$

is solvable for all $(U, L, Z) \in \mathcal{F}_q$. Note the similarity between (5) and (13). Geometrically, we require that q be the largest integer such that the *faces*

$$F(U, L) = \mathbf{P}^* \cap \{\mathbf{u} \in \mathbb{R}^n : u_j = 1 \text{ for } j \in U,\ u_j = -1 \text{ for } j \in L\}$$

of the dual polytope $\mathbf{P}^*$ are nonempty for all $(U, L, Z) \in \mathcal{F}_q$. Since $\mathbf{P}^* \neq \emptyset$, every design matrix is q-strong for some $q \geq 0$. The condition $|L \cup U| = q$ can be replaced with $|L \cup U| \leq q$ in the definition of q-strength. Thus $0 \leq q \leq n$ is well defined for every design matrix $\mathbf{X}$. In the numerical example of section 2.1, the dual polytope $\mathbf{P}^*$ has precisely two extreme points $\mathbf{u}^1 = (1, -\frac{1}{3}, -\frac{2}{3})$ and $\mathbf{u}^2 = -\mathbf{u}^1$. Hence the design matrix $\mathbf{X}$ of (9) is 0-strong, which explains its breakdown point of $\frac{1}{3}$.

PROPOSITION 1. *If* $\mathbf{X}$ *is q-strong, then* $q \leq n - p$ *and* $\alpha(\ell_1, \mathbf{y}|\mathbf{X}) \leq \frac{q+1}{n}$.

Proof. Suppose $q > n - p$ and let $(U, L, Z) \in \mathcal{F}_q$. Thus $|Z| = n - q < p$. Consequently, $\mathbf{X}_Z\boldsymbol{\xi} = \mathbf{0}$ has a solution $\boldsymbol{\xi} \neq \mathbf{0}$. Let $\boldsymbol{\eta}^+ = \boldsymbol{\eta}^- = \mathbf{0}$. If $(-\mathbf{e}_U^T\mathbf{X}_U + \mathbf{e}_L^T\mathbf{X}_L)\boldsymbol{\xi} \neq 0$, then $(\boldsymbol{\xi}, \boldsymbol{\eta}^+, \boldsymbol{\eta}^-)$ or $(-\boldsymbol{\xi}, \boldsymbol{\eta}^+, \boldsymbol{\eta}^-)$ solve (6). Thus by Farkas's lemma, (13) is not solvable, which is a contradiction. Suppose $(-\mathbf{e}_U^T\mathbf{X}_U + \mathbf{e}_L^T\mathbf{X}_L)\boldsymbol{\xi} = 0$. Since $r(\mathbf{X}) = p$, there exist $i \in U \cup L$ such that $\mathbf{x}^i\boldsymbol{\xi} \neq 0$. If $i \in U$, let $S = U - i$ and $T = L + i$. It follows that $(-\mathbf{e}_S^T\mathbf{X}_S + \mathbf{e}_T^T\mathbf{X}_T)\boldsymbol{\xi} = 2\mathbf{x}^i\boldsymbol{\xi} \neq 0$, and thus we contradict with $(S, T, Z) \in \mathcal{F}_q$ as in the previous case. If $i \in L$, we use $S = U + i$ and $T = L - i$. Thus $q \leq n - p$. To prove the second part we proceed as follows. Since $\mathbf{X}$ is q-strong, there exist $(U, L, Z) \in \mathcal{F}_{q+1}$ such that (13) is not solvable. Let

$$g_j = +1 \text{ for all } j \in U, \quad g_j = -1 \text{ for all } j \in L, \quad g_j = 0 \text{ otherwise},$$

and $\mathbf{g} = (g_j)_{j \in N}$. From Lemma 1, $z(\vartheta) = z(\vartheta_0) + (\vartheta - \vartheta_0)\mathbf{u}^0\mathbf{g}$ for all $\vartheta \geq \vartheta_0$, where $\mathbf{u}^0$ is some optimal extreme point of $\mathbf{P}^*$ for $\vartheta = \vartheta_0$. Since (13) is not solvable for U and L, it follows that $-1 < u_j^0 < 1$ for some $j \in U \cup L$. Let $(\boldsymbol{\beta}(\vartheta_0), \mathbf{r}^+(\vartheta_0), \mathbf{r}^-(\vartheta_0))$ be any optimal solution to (10) for $\vartheta = \vartheta_0$. By complementary slackness, $r_j^+(\vartheta_0) = r_j^-(\vartheta_0) = 0$, and thus $\mathbf{x}^j\boldsymbol{\beta}(\vartheta_0) = y_j + \vartheta_0 g_j$. Let $\vartheta > \vartheta_0$ be arbitrary. Since $\mathbf{u}^0$ is unchanged, it follows as before that $r_j^+(\vartheta) = r_j^-(\vartheta) = 0$ in any optimal solution to (10), and thus $\mathbf{x}^j\boldsymbol{\beta}(\vartheta) = y_j + \vartheta g_j$. Consequently, $|\mathbf{x}^j\boldsymbol{\beta}(\vartheta) - \mathbf{x}^j\boldsymbol{\beta}(\vartheta_0)| = \vartheta - \vartheta_0$. By the Cauchy–Schwarz inequality, $\vartheta - \vartheta_0 = |\mathbf{x}^j(\boldsymbol{\beta}(\vartheta) - \boldsymbol{\beta}(\vartheta_0))| \leq \|\mathbf{x}^j\|\|\boldsymbol{\beta}(\vartheta) - \boldsymbol{\beta}(\vartheta_0)\|$, and hence $\|\boldsymbol{\beta}(\vartheta) - \boldsymbol{\beta}(\vartheta_0)\| \to +\infty$ for $\vartheta \to +\infty$; i.e., the maximum bias (8) grows beyond all bounds for $\mathbf{g}$. Thus $\alpha(\ell_1, \mathbf{y}|\mathbf{X}) \leq \frac{q+1}{n}$. ∎

PROPOSITION 2.

(i) *If* $\mathbf{X}$ *is* q*-strong and* $\mathbf{g} \in \mathbb{R}^n$ *in* (10) *has* q *nonzero components, then* (10) *has an optimal solution* $(\boldsymbol{\beta}(\vartheta), \mathbf{r}^+(\vartheta), \mathbf{r}^-(\vartheta))$ *with* $\lim_{\vartheta\to\infty} \|\boldsymbol{\beta}(\vartheta)\|_1 < \infty$.

(ii) *If* $\mathbf{X}$ *is* q*-strong and the solution to* (10) *is unique for* $\vartheta \geq \vartheta_0$, *then* $\alpha(\ell_1, \mathbf{y}|\mathbf{X}) = \frac{q+1}{n}$.

Proof. (i) Let $S = \{i \in N : g_i > 0\}$, $T = \{i \in N : g_i < 0\}$. Since $\mathbf{X}$ is q-strong and $|S \cup T| = q$, it follows that there exists an extreme point $\mathbf{u}^k \in \mathbf{P}^*$ such that $u^k_\ell = 1$ for all $\ell \in S$ and $u^k_\ell = -1$ for all $\ell \in T$. Since $\mathbf{u}^i\mathbf{g} \leq \|\mathbf{g}\|_1$ for all $i = 1, \ldots, r$, and (12) holds for arbitrarily large $\vartheta \geq \vartheta_0$, it follows that any optimal dual extreme point $\mathbf{u}^0$ of $\mathbf{P}^*$ satisfies $u^0_\ell = 1$ for all $\ell \in S$ and $u^0_\ell = -1$ for all $\ell \in T$. Consequently,

$$z(\vartheta) = z(\vartheta_0) + (\vartheta - \vartheta_0)\|\mathbf{g}\|_1 \text{ for all } \vartheta \geq \vartheta_0. \tag{14}$$

Let $(\boldsymbol{\beta}(\vartheta_0), \mathbf{r}^+(\vartheta_0), \mathbf{r}^-(\vartheta_0))$ be an optimal solution to (10) for $\vartheta = \vartheta_0$ and define

$$\big(\boldsymbol{\beta}(\vartheta), \mathbf{r}^+(\vartheta), \mathbf{r}^-(\vartheta)\big) = \big(\boldsymbol{\beta}(\vartheta_0), \mathbf{r}^+(\vartheta_0) + (\vartheta - \vartheta_0)\mathbf{g}^+, \mathbf{r}^-(\vartheta_0) + (\vartheta - \vartheta_0)\mathbf{g}^-\big)\,,$$

where $\mathbf{g}^+ = \max\{\mathbf{0}, \mathbf{g}\}$ and $\mathbf{g}^- = -\min\{\mathbf{0}, \mathbf{g}\}$. $(\boldsymbol{\beta}(\vartheta), \mathbf{r}^+(\vartheta), \mathbf{r}^-(\vartheta))$ is a feasible solution to (10) for all $\vartheta \geq \vartheta_0$. Since $\mathbf{e}_n^T\mathbf{r}^+(\vartheta) + \mathbf{e}_n^T\mathbf{r}^-(\vartheta) = z(\vartheta_0) + (\vartheta - \vartheta_0)\|\mathbf{g}\|_1$, by the duality theorem, it is an optimal solution to (10). Hence $\|\boldsymbol{\beta}(\vartheta)\|_1 = \|\boldsymbol{\beta}(\vartheta_0)\|_1 < \infty$ for $\vartheta \to +\infty$ and the assertion follows.

(ii) If the solution to (10) is unique for all $\vartheta \geq \vartheta_0$, then by part (i) the maximum bias remains bounded if q components of $\mathbf{y}$ are contaminated. Since the contamination vector $\mathbf{g}$ is perfectly arbitrary, by Proposition 1, the assertion follows. □

A necessary and sufficient condition for ℓ_1-regression to have a unique solution can be found, e.g., in [4, Proposition 3].

We next give a different condition for $\mathbf{X}$ to be q-strong. It yields the basis for an algorithmic approach to find the breakdown point of ℓ_1-regression. For $(U, L, Z) \in \mathcal{F}_q$ let

$$z(U, L) = \min\{(-\mathbf{e}_U^T\mathbf{X}_U + \mathbf{e}_L^T\mathbf{X}_L)\boldsymbol{\xi} + \mathbf{e}_Z^T(\boldsymbol{\eta}^+ + \boldsymbol{\eta}^-) : \mathbf{X}_Z\boldsymbol{\xi} + \boldsymbol{\eta}^+ - \boldsymbol{\eta}^- = \mathbf{0},\ \boldsymbol{\eta}^+ \geq \mathbf{0},\ \boldsymbol{\eta}^- \geq \mathbf{0}\}$$

and note that $z(U, L) = 0$ for $U = L = \emptyset$.

THEOREM 2. *A design matrix* $\mathbf{X}$ *is* q*-strong if and only if* q *is the largest integer such that* $z(U, L) \geq 0$ *for all* $(U, L, Z) \in \mathcal{F}_q$.

Proof. We establish sufficiency first. By assumption, for every $(U, L, Z) \in \mathcal{F}_q$

$$\min\{(-\mathbf{e}_U^T\mathbf{X}_U + \mathbf{e}_L^T\mathbf{X}_L)\boldsymbol{\xi} + \mathbf{e}_Z^T\boldsymbol{\eta}^+ + \mathbf{e}_Z^T\boldsymbol{\eta}^- : \mathbf{X}_Z\boldsymbol{\xi} + \boldsymbol{\eta}^+ - \boldsymbol{\eta}^- = \mathbf{0},\ \boldsymbol{\eta}^+ \geq \mathbf{0},\ \boldsymbol{\eta}^- \geq \mathbf{0}\} = 0.$$

By the strong duality theorem of LP, the dual of this linear program,

$$\max\{\mathbf{u0} : \mathbf{u}\mathbf{X}_Z = -\mathbf{e}_U^T\mathbf{X}_U + \mathbf{e}_L^T\mathbf{X}_L,\ -\mathbf{e}_Z^T \leq \mathbf{u} \leq \mathbf{e}_Z^T\},$$

has a finite optimum. Thus (13) is solvable for all choices of $(U, L, Z) \in \mathcal{F}_q$. Suppose that $\mathbf{X}$ is m-strong. It follows that $m \geq q$. Suppose that $m > q$. Then the dual program has a finite optimum for all $(U, L, Z) \in \mathcal{F}_m$ and, by strong duality, so does the primal, i.e., $z(U, L) = 0$ for all $(U, L, Z) \in \mathcal{F}_m$. This contradicts the assumption that q is the largest such integer. Consequently, $\mathbf{X}$ is q-strong. On the other hand, let $\mathbf{X}$ be q-strong. It follows that

$$P^*(U, L) = \{\mathbf{u} \in \mathbb{R}^Z : \mathbf{u}\mathbf{X}_Z = -\mathbf{e}_U^T\mathbf{X}_U + \mathbf{e}_L^T\mathbf{X}_L,\ -\mathbf{e}_Z^T \leq \mathbf{u} \leq \mathbf{e}_Z^T\} \neq \emptyset$$

for all $(U, L, Z) \in \mathcal{F}_q$. $P^*(U, L)$ is a nonempty polytope, and thus by strong duality

$$0 = \max\{\mathbf{u0} : \mathbf{u} \in P^*(U, L)\} = z(U, L).$$

Suppose that $z(U, L) \geq 0$ for all $(U, L, Z) \in \mathcal{F}_m$ with $m > q$. Then, as in the first part of the proof, we have a contradiction by the fact that q is the largest integer such that (13) is solvable. □

It follows from Proposition 2 that for any q-strong design matrix $\mathbf{X}$ and finite $\mathbf{y}$, there exist ℓ_1-regression coefficients whose bias is bounded for any set of q contaminated elements of $\mathbf{y}$. While a q-strong $\mathbf{X}$ may permit ℓ_1-regression coefficients, for which the bias grows beyond all bounds, ℓ_1-regression with q-strong $\mathbf{X}$ still is robust since one needs only to ensure in such a case that the ℓ_1-regression coefficients used are bounded. We show by a brief example that this anomaly of ℓ_1-regression due to multiple optima to (10) can indeed occur.

Example 2. Consider the data

$$(\mathbf{X}, \mathbf{y}) = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 2 & 0 \\ 1 & 3 & 0 \\ 1 & 4 & 0 \end{pmatrix}.$$

It is easily verified that $\mathbf{X}$ is 1-strong. If the fourth observation is contaminated, we find that the parametric linear program (10)

$$\min \sum_{i=1}^{4} (r_i^+ + r_i^-) \quad \text{s.t.} \quad \begin{array}{rrrrrrrr} \beta_1 + & \beta_2 & + r_1^+ & & & - r_1^- & & = 0, \\ \beta_1 + & 2\beta_2 & & + r_2^+ & & & - r_2^- & = 0, \\ \beta_1 + & 3\beta_2 & & & + r_3^+ & & - r_3^- & = 0, \\ \beta_1 + & 4\beta_2 & & & & + r_4^+ & & - r_4^- = \vartheta, \end{array}$$

where β_1 and β_2 are free, $r_i^+ \geq 0$ and $r_i^- \geq 0$ for $i = 1, \ldots, 4$, has two basic optimal solutions given by $\beta_1^0 = \beta_2^0 = 0$, $r_4^+ = \vartheta$ and by $\beta_1^1 = -\vartheta$, $\beta_2^1 = \frac{1}{2}\vartheta$, $r_1^+ = \frac{1}{2}\vartheta$, $r_3^- = \frac{1}{2}\vartheta$, respectively, where in both cases $r_i^+ = 0$ and $r_i^- = 0$ otherwise. Thus in this example we have, in agreement with Proposition 2, the existence of optimal ℓ_1-regression coefficients $\boldsymbol{\beta}^0$ for which the bias is bounded, whereas for $\boldsymbol{\beta}^1$ the bias grows without bounds. □

We call $\mathbf{X}$ *s-stable* if $s \geq 0$ is the largest integer such that

$$\mathbf{X}_Z \boldsymbol{\xi} + \boldsymbol{\eta}^+ - \boldsymbol{\eta}^- = \mathbf{0}, \ \left(-\mathbf{e}_U^T \mathbf{X}_U + \mathbf{e}_L^T \mathbf{X}_L\right) \boldsymbol{\xi} + \mathbf{e}_Z^T \left(\boldsymbol{\eta}^+ + \boldsymbol{\eta}^-\right) \leq 0, \quad \boldsymbol{\xi} \neq \mathbf{0}, \quad \boldsymbol{\eta}^+ \geq \mathbf{0}, \quad \boldsymbol{\eta}^- \geq \mathbf{0}, \tag{15}$$

is *not solvable* for any $(U, L, Z) \in \mathcal{F}_s$. It follows that $s \geq 0$ is well defined for any $\mathbf{X}$ with $r(\mathbf{X}) = p$. If $|Z| < p$, then (15) is solvable: In this case there exists $\boldsymbol{\xi} \neq \mathbf{0}$ such that $\mathbf{X}_Z \boldsymbol{\xi} = \mathbf{0}$. If $(-\mathbf{e}_U^T \mathbf{X}_U + \mathbf{e}_L^T \mathbf{X}_L)\boldsymbol{\xi} \leq 0$, then $\boldsymbol{\xi}$ and $\boldsymbol{\eta}^+ = \boldsymbol{\eta}^- = \mathbf{0}$ solve (15); otherwise, we change the sign of $\boldsymbol{\xi}$. It follows that $s \leq n - p$. Note the subtle difference between (6) and (15). More precisely, every solution to (6) is feasible for (15), but not vice versa.

Example 2 (continued). Consider the 1-strong design matrix $\mathbf{X}$ of Example 2. $r(\mathbf{X}) = p$ and thus $s \geq 0$. Let $(U, L, Z) \in \mathcal{F}_1$, where $U = \{4\}$, $L = \emptyset$, and $Z = \{1, 2, 3\}$. Then $\xi_1 = -\theta$, $\xi_2 = \frac{\theta}{2}$, $\eta_1^+ = \eta_3^- = \frac{\theta}{2}$ for arbitrary real θ and $\eta_i^+ = \eta_i^- = 0$ otherwise solves (15), but not (6). Thus the 1-strong design matrix $\mathbf{X}$ is 0-stable. □

From the definition of s it follows that $z(U, L) \geq 0$ for all $(U, L, Z) \in \mathcal{F}_s$.

PROPOSITION 3. *If* $\mathbf{X}$ *is* s*-stable, then* $s \le n-p$ *and* $\mathbf{X}$ *is* q*-strong with* $q \ge s$.

Proof. For the proof, see the above discussion and Theorem 2. □

As the last example shows, the inequality $q \ge s$ in Proposition 3 can be sharp.

PROPOSITION 4. *If* $\mathbf{X}$ *is* s*-stable, then* $\alpha(\ell_1, \mathbf{y}|\mathbf{X}) \le \frac{s+1}{n}$.

Proof. Suppose that $\mathbf{X}$ is q-strong for some $q \ge 0$. Hence $q \ge s$. Thus if $s = q$, the assertion follows from Proposition 1. Suppose $q > s$. Then there exists $(U, L, Z) \in \mathcal{F}_{s+1}$ such that (15) is solvable. Let $\boldsymbol{\xi}, \boldsymbol{\eta}^+, \boldsymbol{\eta}^-$ be any solution to (15) with $\boldsymbol{\xi} \ne \mathbf{0}$. Define $\mathbf{g} \in \mathbb{R}^n$ by $g_j = \mathbf{x}^j \boldsymbol{\xi}$ for all $j \in U \cup L$, $g_j = 0$ otherwise, and consider the linear program (10). By Lemma 1 there exists some finite $\vartheta_0 \ge 0$ such that (12) holds for all $\vartheta \ge \vartheta_0$. Since $\mathbf{X}$ is q-strong with $q \ge s+1$, it follows, as in the proof of Proposition 2, that (14) holds. Define $\boldsymbol{\beta}(\vartheta) = \boldsymbol{\beta}(\vartheta_0) + (\vartheta - \vartheta_0)\boldsymbol{\xi}$ and

$$\mathbf{r}_Z^+(\vartheta) = \mathbf{r}_Z^+(\vartheta_0) + (\vartheta - \vartheta_0)\boldsymbol{\eta}^+, \quad \mathbf{r}_Z^-(\vartheta) = \mathbf{r}_Z^-(\vartheta_0) + (\vartheta - \vartheta_0)\boldsymbol{\eta}^-, \quad \mathbf{r}_{U\cup L}^{\pm}(\vartheta) = \mathbf{r}_{U\cup L}^{\pm}(\vartheta_0).$$

It follows that $(\boldsymbol{\beta}(\vartheta), \mathbf{r}^+(\vartheta), \mathbf{r}^-(\vartheta))$ is feasible for (10) and, since by duality

$$z(\vartheta) \le \mathbf{e}^T(\mathbf{r}^+(\vartheta) + \mathbf{r}^-(\vartheta)) = z(\vartheta_0) + (\vartheta - \vartheta_0)\mathbf{e}_Z^T(\boldsymbol{\eta}^+ + \boldsymbol{\eta}^-) \le z(\vartheta_0) + (\vartheta - \vartheta_0)\|\mathbf{g}\|_1 = z(\vartheta),$$

it is optimal for (10). But $\lim_{\vartheta \to +\infty} \|\boldsymbol{\beta}(\vartheta)\| \to +\infty$, since $\boldsymbol{\xi} \ne \mathbf{0}$, and thus $\alpha(\ell_1, \mathbf{y}|\mathbf{X}) \le \frac{s+1}{n}$. □

THEOREM 3. *A design matrix* $\mathbf{X}$ *is* s*-stable if and only if* $\alpha(\ell_1, \mathbf{y}|\mathbf{X}) = \frac{s+1}{n}$.

Proof. Let $\mathbf{X}$ be s-stable and suppose that $\alpha(\ell_1, \mathbf{y}|\mathbf{X}) \le \frac{s}{n}$. Then there exists $\mathbf{g} \in \mathbb{R}^n$ with s nonzero components such that (10) has an optimal solution $(\boldsymbol{\beta}(\vartheta), \boldsymbol{r}^+(\vartheta), \boldsymbol{r}^-(\vartheta))$ with $\|\boldsymbol{\beta}(\vartheta)\|_1 \to +\infty$ for $\vartheta \to +\infty$. By Lemma 1 there exists a finite $\vartheta_0 \ge 0$ such that $z(\vartheta) = z(\vartheta_0) + (\vartheta - \vartheta_0)\mathbf{u}^0\mathbf{g}$ for all $\vartheta \ge \vartheta_0$ and some extreme point $\mathbf{u}^0$ of $\mathbf{P}^*$. Since $\mathbf{X}$ is q-strong with $q \ge s$ it follows as before that (14) holds. By [4, Proposition 2] there exists $B \subseteq N$ with $|B| = p$ such that $\mathbf{X}_B\boldsymbol{\beta}(\vartheta) = \mathbf{y}_B + \vartheta\mathbf{g}_B$ with $\mathbf{X}_B$ nonsingular, and thus

$$\boldsymbol{\beta}(\vartheta) = \mathbf{X}_B^{-1}\mathbf{y}_B + \vartheta\mathbf{X}_B^{-1}\mathbf{g}_B = \boldsymbol{\beta}(\vartheta_0) + (\vartheta - \vartheta_0)\boldsymbol{\xi}$$

for all $\vartheta \ge \vartheta_0$, where $\boldsymbol{\xi} = \mathbf{X}_B^{-1}\mathbf{g}_B \ne \mathbf{0}$ because $\|\beta(\vartheta)\|_1 \to +\infty$. It follows that $\mathbf{r}^{\pm}(\vartheta) = \mathbf{r}^{\pm}(\vartheta_0) + (\vartheta - \vartheta_0)\boldsymbol{\eta}^{\pm}$ for all $\vartheta \ge \vartheta_0$, where $\boldsymbol{\eta}^+ = \max\{\mathbf{0}, \mathbf{g} - \mathbf{X}\boldsymbol{\xi}\}$ and $\boldsymbol{\eta}^- = -\min\{\mathbf{0}, \mathbf{g} - \mathbf{X}\boldsymbol{\xi}\}$, respectively. Since $\mathbf{X}\boldsymbol{\beta}(\vartheta) + \mathbf{r}^+(\vartheta) - \mathbf{r}^-(\vartheta) = \mathbf{y} + \vartheta\mathbf{g}$ for all $\vartheta \ge 0$, we calculate

$$\mathbf{X}\boldsymbol{\beta}(\vartheta_0) + \mathbf{r}^+(\vartheta_0) - \mathbf{r}^-(\vartheta_0) + (\vartheta - \vartheta_0)(\mathbf{X}\boldsymbol{\xi} + \boldsymbol{\eta}^+ - \boldsymbol{\eta}^-) = \mathbf{y} + \vartheta_0\mathbf{g} + (\vartheta - \vartheta_0)\mathbf{g}.$$

Consequently, $\mathbf{X}\boldsymbol{\xi} + \boldsymbol{\eta}^+ - \boldsymbol{\eta}^- = \mathbf{g}$. Define $(U, L, Z) \in \mathcal{F}_s$ by

$$Z = \{i \in N : g_i = 0\}, \quad U = \{i \in N - Z : g_i > 0\}, \quad L = \{i \in N - Z : g_i < 0\}.$$

From $\mathbf{X}_U\boldsymbol{\xi} + \boldsymbol{\eta}_U^+ - \boldsymbol{\eta}_U^- = \mathbf{g}_U$ and $\mathbf{X}_L\boldsymbol{\xi} + \boldsymbol{\eta}_L^+ - \boldsymbol{\eta}_L^- = \mathbf{g}_L$, we calculate

$$\left(\mathbf{e}_U^T\mathbf{X}_U - \mathbf{e}_L^T\mathbf{X}_L\right)\boldsymbol{\xi} + \mathbf{e}_U^T\boldsymbol{\eta}_U^+ + \mathbf{e}_L^T\boldsymbol{\eta}_L^- - \left(\mathbf{e}_U^T\boldsymbol{\eta}_U^- + \mathbf{e}_L^T\boldsymbol{\eta}_L^+\right) = \|\mathbf{g}\|_1.$$

Calculating the optimal objective function value of (10) from the primal solution, we find

$$z(\vartheta) = \mathbf{e}_N^T(\mathbf{r}^+(\vartheta) + \mathbf{r}^-(\vartheta)) = z(\vartheta_0) + (\vartheta - \vartheta_0)\mathbf{e}_N^T(\boldsymbol{\eta}^+ + \boldsymbol{\eta}^-) = z(\vartheta_0) + (\vartheta - \vartheta_0)\|\mathbf{g}\|_1.$$

Thus $\mathbf{e}_N^T(\boldsymbol{\eta}^+ + \boldsymbol{\eta}^-) = \|\mathbf{g}\|_1$ and, from the previous relation for $\|\mathbf{g}\|_1$, we get

$$\left(-\mathbf{e}_U^T\mathbf{X}_U + \mathbf{e}_L^T\mathbf{X}_L\right)\boldsymbol{\xi} + \mathbf{e}_Z^T\left(\boldsymbol{\eta}_Z^+ + \boldsymbol{\eta}_Z^-\right) = -2\left(\mathbf{e}_U^T\boldsymbol{\eta}_U^- + \mathbf{e}_L^T\boldsymbol{\eta}_L^+\right) \le 0.$$

Hence (15) is solvable for $(U, L, Z) \in \mathcal{F}_s$, which is a contradiction. Thus $\alpha(\ell_1, \mathbf{y}|\mathbf{X}) > \frac{s}{n}$, and by Proposition 4 we have equality. Suppose $\alpha(\ell_1, \mathbf{y}|\mathbf{X}) = \frac{s+1}{n}$. By definition, s is the smallest integer with this property. Suppose $\mathbf{X}$ is t-stable for some integer $t \geq 0$. By the first part of this theorem, $\alpha(\ell_1, \mathbf{y}|\mathbf{X}) = \frac{t+1}{n}$, and thus $s = t$. ∎

Proposition 2 shows that for q-strong $\mathbf{X}$ there exist optimal ℓ_1-regression coefficients such that the bias remains bounded when at most q data of $\mathbf{y}$ are contaminated. It is thus debatable whether or not we want to talk of a "breakdown" of ℓ_1-regression in this case. By Proposition 3 we know $q \geq s$, and from Example 2 we know that $q > s$ is possible. The question becomes whether or not the difference $q - s$ is "reasonably" small. Under the assumption that the data $\mathbf{X}$ are *in general position* we can answer the question affirmatively. ($\mathbf{X}$ is in general position if every $p \times p$ submatrix of $\mathbf{X}$ is nonsingular.)

PROPOSITION 5. *If* $\mathbf{X}$ *is in general position, q-strong, and s-stable, then* $0 \leq q - s \leq 1$.

Proof. Suppose that $q > s$. Then there exist $(U, L, Z) \in \mathcal{F}_{s+1}$ such that (15) has a solution $(\boldsymbol{\xi}, \boldsymbol{\eta}_Z^+, \boldsymbol{\eta}_Z^+)$. We claim that $\eta_i^+ + \eta_i^- > 0$ for some $i \in Z$. Otherwise, $\mathbf{X}_Z \boldsymbol{\xi} = \mathbf{0}$ with $\boldsymbol{\xi} \neq \mathbf{0}$ implies that $r(\mathbf{X_Z}) < p$. But, by Proposition 1, $|Z| = n-(s+1) \geq n-q \geq p$, and thus $r(\mathbf{X_Z}) = p$ since $\mathbf{X}$ is in general position. The claim follows. Assume that $\eta_i^+ > 0$ for some $i \in Z$. Let $Z^* = Z - i$, $U^* = U$, and $L^* = L + i$. We calculate

$$\begin{aligned}&\left(-\mathbf{e}_{U^*}^T \mathbf{X}_{U^*} + \mathbf{e}_{L^*}^T \mathbf{X}_{L^*}\right) \boldsymbol{\xi} + \sum_{i \in Z^*} \left(\eta_i^+ + \eta_i^-\right) \\ &\quad = \left(-\mathbf{e}_U^T \mathbf{X}_U + \mathbf{e}_L^T \mathbf{X}_L\right) \boldsymbol{\xi} + \sum_{i \in Z} \left(\eta_i^+ + \eta_i^-\right) + \mathbf{x}^i \boldsymbol{\xi} - \eta_i^+ - \eta_i^- \leq -2\eta_i^+ < 0\end{aligned}$$

because $\mathbf{x}^i \boldsymbol{\xi} + \eta_i^+ - \eta_i^- = 0$. Hence (6) has a solution for $(U^*, L^*, Z^*) \in \mathbf{F}_{s+2}$, and thus by Farkas's lemma the corresponding (13) is not solvable, i.e., $q < s + 2$. If $\eta_i^- > 0$ for some $i \in Z$, we let $Z^* = Z - i$, $U^* = U + i$, and $L^* = L$ and calculate likewise. ∎

2.3. Related work on $\alpha(\ell_1, \mathbf{y}|\mathbf{X})$. To summarize previous results on the determination of the finite sample breakdown point of ℓ_1-regression, let m_* be the largest integer such that for all $S \subseteq N$ with $|S| = m_*$

$$\inf_{\|\boldsymbol{\xi}\|=1} \frac{\sum_{i \in N-S} |\mathbf{x}^i \boldsymbol{\xi}|}{\sum_{i \in N} |\mathbf{x}^i \boldsymbol{\xi}|} > \frac{1}{2}. \tag{16}$$

He et al. [7, Theorem 5.2] show $(m_* + 1)/n \leq \alpha(\ell_1, \mathbf{y}|\mathbf{X}) \leq (m_* + 2)/n$. Mizera and Müller [11] prove that

$$\alpha(\ell_1, \mathbf{y}|\mathbf{X}) = (m_* + 1)/n. \tag{17}$$

We show that Theorem 3 is consistent with (17). From (16) it follows that

$$-\sum_{i \in S} |\mathbf{x}^i \boldsymbol{\xi}| + \sum_{i \in N-S} |\mathbf{x}^i \boldsymbol{\xi}| > 0 \quad \text{for all} \quad \boldsymbol{\xi} \in \mathbb{R}^p \quad \text{with} \quad \|\boldsymbol{\xi}\| > \mathbf{0}.$$

Hence $\sum_{i \in S} |\mathbf{x}^i \boldsymbol{\xi}| \geq \sum_{i \in U} \mathbf{x}^i \boldsymbol{\xi} - \sum_{i \in L} \mathbf{x}^i \boldsymbol{\xi}$, where $U, L \subseteq S$ with $L \cap U = \emptyset$ and $L \cup U = S$ are arbitrary. Letting $Z = N - S$ we get, from the previous inequality,

$$-\sum_{i \in U} \mathbf{x}^i \boldsymbol{\xi} + \sum_{i \in L} \mathbf{x}^i \boldsymbol{\xi} + \sum_{i \in Z} |\mathbf{x}^i \boldsymbol{\xi}| > 0 \tag{18}$$

for all $\boldsymbol{\xi} \in \mathbb{R}^p$ with $\|\boldsymbol{\xi}\| > 0$ and U and L as specified. Let $\boldsymbol{\eta}^+, \boldsymbol{\eta}^- \in \mathbb{R}^{|Z|}$ satisfy $\boldsymbol{\eta}^+, \boldsymbol{\eta}^- \geq \mathbf{0}$ and $\mathbf{X}_Z\boldsymbol{\xi} + \boldsymbol{\eta}^+ - \boldsymbol{\eta}^- = \mathbf{0}$ for some $\boldsymbol{\xi} \in \mathbb{R}^p$. It follows that

$$\eta_i^+ + \eta_i^- \geq |\mathbf{x}^i\boldsymbol{\xi}| \qquad \text{for } i \in Z \tag{19}$$

because $\mathbf{x}^i\boldsymbol{\xi} + \eta_i^+ - \eta_i^- = 0$ and $\eta_i^+ \geq 0$, $\eta_i^- \geq 0$ imply $\eta_i^+ \geq -\mathbf{x}^i\boldsymbol{\xi}$ and $\eta_i^- \geq \mathbf{x}^i\boldsymbol{\xi}$. Thus if $\mathbf{x}^i\boldsymbol{\xi} \leq 0$, then $\eta_i^+ \geq |\mathbf{x}^i\boldsymbol{\xi}|$, and if $\mathbf{x}^i\boldsymbol{\xi} > 0$, then $\eta_i^- \geq |\mathbf{x}^i\boldsymbol{\xi}|$. Equation (19) follows. From (18) and (19),

$$\left(-\mathbf{e}_U^T\mathbf{X}_U + \mathbf{e}_L^T\mathbf{X}_L\right)\boldsymbol{\xi} + \mathbf{e}_Z^T\left(\boldsymbol{\eta}^+ + \boldsymbol{\eta}^-\right) > 0$$

for all $\boldsymbol{\xi} \in \mathbb{R}^p$ with $\|\boldsymbol{\xi}\| > 0$ and U and L as specified. Consequently, since S is any subset of N with $|S| = m_*$, (15) has no solution for any $(U, L, Z) \in \mathcal{F}_{m_*}$. Hence $\mathbf{X}$ is s-stable with $s \geq m_*$. By the definition of m_*, there exists $S \subseteq N$ with $|S| = m_* + 1$ such that

$$\inf_{\|\boldsymbol{\xi}\|=1} \frac{\sum_{i\in N-S} |\mathbf{x}^i\boldsymbol{\xi}|}{\sum_{i\in N} |\mathbf{x}^i\boldsymbol{\xi}|} \leq \frac{1}{2}.$$

Hence $-\sum_{i\in S} |\mathbf{x}^i\boldsymbol{\xi}^0| + \sum_{i\in N-S} |\mathbf{x}^i\boldsymbol{\xi}^0| \leq 0$ for some $\boldsymbol{\xi}^0 \neq \mathbf{0}$. Define $U = \{i \in S : \mathbf{x}^i\boldsymbol{\xi}^0 \geq 0\}$ and $L = \{i \in S : \mathbf{x}^i\boldsymbol{\xi}^0 < 0\}$. Then $\sum_{i\in S} |\mathbf{x}^i\boldsymbol{\xi}^0| = (\mathbf{e}_U^T\mathbf{X}_U - \mathbf{e}_L^T\mathbf{X}_L)\boldsymbol{\xi}^0$. Let $\eta_i^+ = \max\{0, \mathbf{x}^i\boldsymbol{\xi}^0\}$ and $\eta_i^- = -\min\{0, \mathbf{x}^i\boldsymbol{\xi}^0\}$ for all $i \in Z = N - S$. It follows that $(\boldsymbol{\xi}^0, \boldsymbol{\eta}_Z^+, \boldsymbol{\eta}_Z^-)$ solves (15), and thus $s < m_* + 1$. Consequently, $s = m_*$ and the results agree even though the proof methodologies employed are quite different.

Mizera and Müller [12] provide an enumerative algorithm for the computation of the conditional breakdown point as follows. $m_* = |S|$ is the largest integer such that (16) holds for all $S \subseteq N$ with $|S| = m_*$ if and only if $|E| = m_* + 1$, where $m_* + 1$ is the smallest integer such that there exists an $E \subseteq N$ such that

$$\max_{\|\boldsymbol{\xi}\|=1} \frac{\sum_{i\in E} |\mathbf{x}^i\boldsymbol{\xi}|}{\sum_{i\in N} |\mathbf{x}^i\boldsymbol{\xi}|} \geq \frac{1}{2}. \tag{20}$$

Note that the restriction $\|\boldsymbol{\xi}\| = 1$ can be dropped. They then show that in order to calculate $m_* + 1$ it is sufficient to compare at most $\binom{n}{p-1}$ candidate solutions for $\boldsymbol{\xi}$ in (20).

3. Calculating the *q*-strength and *s*-stability of a design matrix. To calculate the q-strength and s-stability of a design matrix $\mathbf{X}$, there are two roads to take. The first is to find the q-strength or the s-stability of $\mathbf{X}$ by enumeration. The second is to formulate an MIP and solve the program. Mizera and Müller [12] provide a "special purpose" enumerative algorithm. Here we use the results of section 2 to formulate the problem as an MIP. Thus, in order to calculate, a user need only generate the corresponding constraint set for the data of his/her design matrix $\mathbf{X}$ (a useful exercise for a graduate student) and solve the problem using some commercially available MIP solver, such as CPLEX.

By Theorem 2, $\mathbf{X}$ is not q-strong if for some $(U, L, Z) \in \mathcal{F}_q$ there exist $\boldsymbol{\xi} \in \mathbb{R}^p$, $\boldsymbol{\eta}^+, \boldsymbol{\eta}^- \in \mathbb{R}^{|Z|}$ satisfying (6). Thus the problem of determining the q-strength of $\mathbf{X}$ consists of finding the smallest integer such that (6) is solvable for some $(U, L, Z) \in \mathcal{F}_q$.

We claim that the following MIP, called MIP1, does just that:

$$\min \sum_{i=1}^{n} u_i + \ell_i$$

$$\text{s.t.} \quad \mathbf{x}^i\boldsymbol{\xi} + \eta_i^+ - \eta_i^- + s_i - t_i = 0 \quad \text{for } i = 1, \ldots, n, \tag{21}$$

$$s_i - Mu_i \le 0, \quad s_i + Mu_i \ge 0 \quad \text{for } i = 1, \ldots, n, \tag{22}$$

$$t_i - M\ell_i \le 0, \quad t_i + M\ell_i \ge 0 \quad \text{for } i = 1, \ldots, n, \tag{23}$$

$$\eta_i^+ + \eta_i^- + Mu_i + M\ell_i \le M \quad \text{for } i = 1, \ldots, n, \tag{24}$$

$$u_i + \ell_i \le 1 \quad \text{for } i = 1, \ldots, n, \tag{25}$$

$$\sum_{i=1}^{n} s_i + t_i + \eta_i^+ + \eta_i^- \le -\varepsilon, \tag{26}$$

$$\boldsymbol{\xi}, \mathbf{s}, \mathbf{t} \text{ free}, \boldsymbol{\eta}^+ \ge \mathbf{0}, \boldsymbol{\eta}^- \ge \mathbf{0}, \quad u_i, \ell_i \in \{0, 1\} \quad \text{for } i = 1, \ldots, n. \tag{27}$$

We assume that $M > 0$ is a suitably chosen large number and $\varepsilon > 0$ a small number so that the constraints (22)–(24) corresponding to the ith observation are *nonbinding* for the solution that results if we set u_i or ℓ_i equal to 1 or $u_i = \ell_i = 0$. It can be shown by standard arguments of mixed-integer programming that such M and ε exist. Specifically, let U be the set of indices in which $u_i = 1$ in any solution to MIP1, let L be given by $\ell_i = 1$, and $Z = N - U - L$. By (25), $L \cap U = \emptyset$ and thus U, L, Z is a three-way partition of N. Moreover, to every such three-way partition of N there corresponds some setting of u_i and ℓ_i equal to 0 or 1 with $u_i + \ell_i \le 1$. Equations (22)–(24) constrain s_i, t_i, η_i^+, and η_i^- such that if $u_i = 0$, then $s_i = 0$; if $\ell_i = 0$, then $t_i = 0$; and if $u_i + \ell_i = 1$, then $\eta_i^+ = \eta_i^- = 0$. It follows that the constraints of MIP1 produce the following system of constraints:

$$\mathbf{X}_Z\boldsymbol{\xi} + \boldsymbol{\eta}_Z^+ - \boldsymbol{\eta}_Z^- = \mathbf{0}, \tag{28}$$

$$\mathbf{X}_U\boldsymbol{\xi} + \mathbf{s}_U = \mathbf{0}, \qquad \mathbf{X}_L\boldsymbol{\xi} - \mathbf{t}_L = \mathbf{0}, \tag{29}$$

$$\sum_{i \in U} s_i + \sum_{i \in L} t_i + \sum_{i \in Z} \eta_i^+ + \eta_i^- \le -\varepsilon \tag{30}$$

since (22)–(24) are redundant. By (29) and (30),

$$0 > -\varepsilon \ge \sum_{i \in U} s_i + \sum_{i \in L} t_i + \sum_{i \in Z} \eta_i^+ + \eta_i^- = \left(-\mathbf{e}_U^T\mathbf{X}_U + \mathbf{e}_L^T\mathbf{X}_L\right)\boldsymbol{\xi} + \mathbf{e}_Z^T\boldsymbol{\eta}_Z^+ + \mathbf{e}_Z^T\boldsymbol{\eta}_Z^-.$$

Thus, (6) is satisfied. Since MIP1 minimizes $\sum_{i=1}^{n} u_i + \ell_i = |L \cup U|$, it determines the smallest integer k such that $\mathbf{X}$ is not k-strong. Therefore, $\mathbf{X}$ is q-strong with $q = k - 1$.

MIP1 calculates the q-strength of a design matrix $\mathbf{X}$. However, if the size of $\mathbf{X}$ is large, e.g., if $n \ge 100$, then even with today's powerful MIP solvers the calculation may take too much time. Thus, we now provide guidelines for a heuristic to determine a good upper bound of the q-strength and the s-stability of a large design matrix.

Although solving a large MIP *exactly* may take a long time, finding a *feasible* solution is often much easier and can be accomplished quite quickly. Determining a feasible solution to MIP1 provides an upper bound Q on the q-strength of a design matrix. This upper bound can be improved upon if some subset $S \subset N$ exists where $|S| < Q$, $L, U \subset S$ with $L \cup U = S$, $L \cap U = \emptyset$, and $Z = N - S$ such that no solution

exists to (13). In such a case, the design matrix in question is at most $(|S|-1)$-strong. Furthermore, from preliminary computational results, when a matrix is not $|S|$-strong, it is often easy to find a subset $S \subset N$ such that no solution exists to (13).

HEURISTIC UPPER BOUND Q.

Step 1. Input MIP1 to a commercially available package such as CPLEX. Find some feasible (not necessarily optimal) solution to MIP1 with objective function value Q.

Step 2. Randomly select r subsets, $S \subset N$, of size $Q-1$. Set $i=1$.

Step 3. Randomly select p partitions of S_i, where $L_i \cup U_i = S_i$. Set $j=1$.

Step 4. For the jth partition of S_i with $Z = N - S_i$, solve a linear program corresponding to (13). If a solution exists, replace j with $j+1$. Otherwise, replace Q with $Q-1$ and goto Step 2.

Step 5. If $j > p$, replace i with $i+1$. Otherwise, goto Step 4.

Step 6. If $i \le r$, goto Step 3. Otherwise, Q is an upper bound for the q-strength of the design matrix in question and stop.

By Theorem 3, the finite sample breakdown point of ℓ_1-regression equals $\frac{s+1}{n}$, where the design matrix with n rows is s-stable. Although q-strength provides a robustness measure by itself, calculating the q-strength does not guarantee the exact value of the breakdown point. We now provide another mixed-integer linear program that calculates the s-stability and thus the breakdown point of ℓ_1-regression. This MIP, called MIP2, is similar to MIP1 except that here we calculate the smallest value of $|U \cup L|$ such that (15) is solvable.

$$\min \sum_{i=1}^{n} u_i + \ell_i$$

$$\text{s.t.} \quad \mathbf{x}^i \boldsymbol{\xi} + \eta_i^+ - \eta_i^- + s_i - t_i = 0 \quad \text{for } i = 1, \dots, n, \tag{31}$$

$$s_i - M u_i \le 0, \quad t_i - M \ell_i \le 0 \quad \text{for } i = 1, \dots, n, \tag{32}$$

$$\eta_i^+ + \eta_i^- + M u_i + M \ell_i \le M \quad \text{for } i = 1, \dots, n, \tag{33}$$

$$u_i + \ell_i \le 1 \quad \text{for } i = 1, \dots, n, \tag{34}$$

$$\sum_{i=1}^{n} \eta_i^+ + \eta_i^- - s_i - t_i \le 0, \qquad \sum_{i=1}^{n} s_i + t_i \ge \varepsilon, \tag{35}$$

$$\boldsymbol{\xi} \text{ free}, \ \boldsymbol{\eta}^+ \ge \mathbf{0}, \ \boldsymbol{\eta}^- \ge \mathbf{0}, \ \mathbf{s} \ge \mathbf{0}, \ \mathbf{t} \ge \mathbf{0}, \quad u_i, \ell_i \in \{0,1\} \quad \text{for } i = 1, \dots, n. \tag{36}$$

As in MIP1, we assume that M is a suitably chosen large number and ε a small number so that (32) and (33) are nonbinding for the solution that results if we set u_i or ℓ_i equal to 1 or $u_i = \ell_i = 0$. The existence of such M and ε can be shown as in the case of MIP1. As in the case of MIP1, we argue that to every three-way partition of N there corresponds a feasible solution to MIP2 and vice versa. Equations (32)–(33) constrain s_i, t_i, η_i^+, and η_i^- such that if $u_i = 0$, then $s_i = 0$; if $\ell_i = 0$, then $t_i = 0$; and if $u_i + \ell_i = 1$, then $\eta_i^+ = \eta_i^- = 0$. It follows that every feasible assignment of the zero-one variables of MIP2 reduces MIP2 to the following constraints:

$$\mathbf{X}_Z \boldsymbol{\xi} + \boldsymbol{\eta}_Z^+ - \boldsymbol{\eta}_Z^- = \mathbf{0}, \tag{37}$$

$$\mathbf{X}_U \boldsymbol{\xi} + \mathbf{s}_U = \mathbf{0}, \qquad \mathbf{X}_L \boldsymbol{\xi} - \mathbf{t}_L = \mathbf{0}, \tag{38}$$

$$-\sum_{i \in U} s_i - \sum_{i \in L} t_i + \sum_{i \in Z} \eta_i^+ + \eta_i^- \le 0, \tag{39}$$

TABLE 1
Breakdown points $\alpha(\ell_1, \mathbf{y}|\mathbf{X})$ for ℓ_1-regression.

Data set	n	p	q-strength	s-stability	Breakdown
Stackloss	21	4	3	3	$\frac{4}{21}$
Aircraft	23	5	1	1	$\frac{2}{23}$
Delivery	25	3	2	2	$\frac{3}{25}$
Engine	16	5	1	1	$\frac{2}{16}$
Gessel	21	2	2	2	$\frac{3}{21}$
Salinity	28	4	3	3	$\frac{4}{28}$
Telephone	24	2	6	5	$\frac{6}{24}$
Wood	20	6	2	2	$\frac{3}{20}$
Star	47	2	4	4	$\frac{5}{47}$

where $U = \{i \in N : u_i = 1\}$ and $L = \{i \in N : \ell_i = 1\}$. By (38) and (39),

$$0 \geq -\sum_{i \in U} s_i - \sum_{i \in L} t_i + \sum_{i \in Z} \eta_i^+ + \eta_i^- = \left(-\mathbf{e}_U^T \mathbf{X}_U + \mathbf{e}_L^T \mathbf{X}_L\right) \boldsymbol{\xi} + \mathbf{e}_Z^T \boldsymbol{\eta}_Z^+ + \mathbf{e}_Z^T \boldsymbol{\eta}_Z^-.$$

Since $\sum_{i=1}^n s_i + t_i \geq \varepsilon > 0$, it follows that either $s_i > 0$, $t_i = 0$, and $\eta^+ = \eta_i^- = 0$ or $s_i = 0$, $t_i > 0$, and $\eta^+ = \eta_i^- = 0$ for some $i \in N$. Consequently, from (31), $\mathbf{x}^i \boldsymbol{\xi} = -s_i + t_i \neq 0$, and thus $\boldsymbol{\xi} \neq \mathbf{0}$. Hence, (15) is satisfied. On the other hand, if (15) is solvable, then a feasible solution to MIP2 is readily constructed. So the formulation MIP2 does the job. Since MIP2 minimizes $\sum_{i=1}^n u_i + \ell_i = |L \cup U|$, it determines the smallest integer k such that $\mathbf{X}$ is not k-stable. Therefore, $\mathbf{X}$ is s-stable with $s = k - 1$.

To demonstrate the usefulness of our approach, we have calculated the q-strength and the s-stability of the design matrices for nine data sets from the robust regression literature. All of these data sets can be found in [14], except for the engine data set which can be found in [10, p. 529]. These results are listed in Table 1. In all but one data set, $s = q$ and the breakdown point equals $\frac{q+1}{n}$. The results were obtained by solving the corresponding MIPs by the commercially available CPLEX package on a Pentium 4 (2.26 GHz) processor. The median solution time and the median number of nodes in the branch and bound tree of the nine MIPs for the s-stability were 25.09 seconds and 22164, respectively. The median time spent and number of nodes traversed in order to find the optimal solution were 6.19 seconds and 10779, respectively. On the other hand, the median solution time and the median number of nodes in the branch and bound tree of the nine MIPs for the q-strength were 1.47 seconds and 1169, respectively. The median time spent and number of nodes traversed in order to find the optimal solution were .26 seconds and 405, respectively.

3.1. The telephone data. The telephone data set is data on the number of international phone calls from Belgium per year versus the years 1950 through 1973. The data has outliers in the values of the dependent variable (calls), in particular the data for the calls in the years 1964 through 1969. These outliers were due to the recording of the number of international phone *call minutes* from Belgium as opposed to the number of calls. Figure 1 contains four graphs of versions of this data set, each with a fitted ℓ_1-regression line. The graph in the upper left-hand corner is the original data set with the outliers as described above. Notice that the ℓ_1-regression line is hardly influenced by the outliers. To further demonstrate this, the graph in the upper right-hand corner contains an ℓ_1-fit to data that is most likely quite similar to

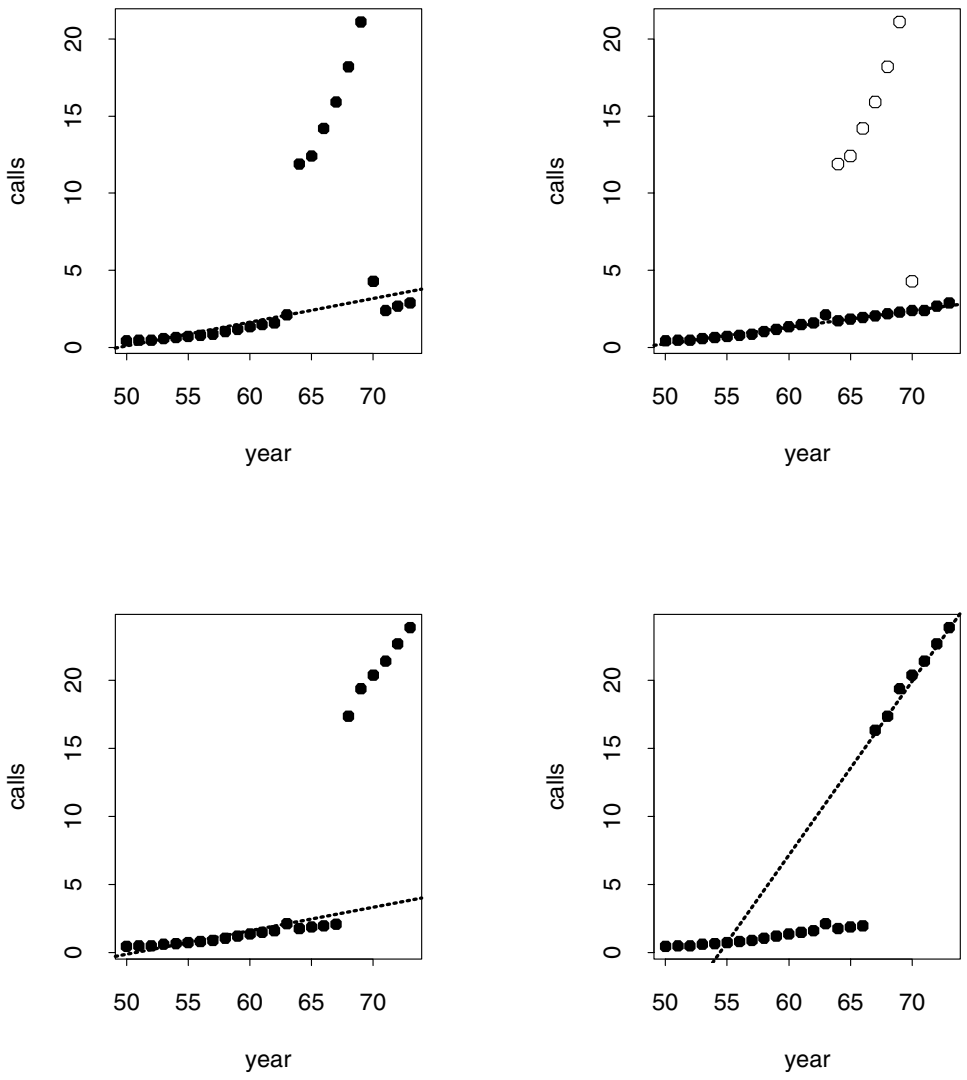


FIG. 1. *ℓ_1-regression lines for telephone data.*

the "uncontaminated" data set. The estimated observations used for the years 1964 through 1970 are determined from a least squares fit to the data excluding the years 1963 through 1970. The original outliers are plotted by empty circles but are not taken into consideration when solving for the ℓ_1-regression line. From Table 1, the q-strength for this problem is 6, its s-stability is 5, and thus $\alpha(\ell_1, \mathbf{X}|\mathbf{y}) = \frac{6}{24}$; i.e., some set of six contaminated observations may cause the ℓ_1-regression estimator to break down in this example. However, by Proposition 2 and the fact that the q-strength is 6, there exist bounded ℓ_1-regression coefficients for every set of six contaminated observations, but not for every set of seven contaminated observations. The graph in the bottom left-hand corner shows the uncontaminated data set with six new outliers for the years 1968 through 1973. The ℓ_1-regression estimator performs quite well with these six outliers in the sense that ℓ_1-regression coefficients exist that are hardly

influenced by the outliers. However, for just one more contaminated observation, there are no longer ℓ_1-regression coefficients that are not influenced by the outliers; see the graph in the bottom right corner. In our minds this points to the suitability of the q-strength of a design matrix as a measure of the breakdown of ℓ_1-regression. In any case, by Proposition 5, q-strength and s-stability are, in general, reasonably close to each other.

4. Conclusion. We have provided an exact formula for the breakdown point $\alpha(\ell_1, \mathbf{y}|\mathbf{X})$ of ℓ_1-regression with contamination restricted to the dependent variable. This is done using the notion of the s-stability of the design matrix $\mathbf{X}$, which is introduced here. We have shown that our results agree with results known in the literature. We have also introduced the notion of q-strength, a new robustness measure for ℓ_1-regression. Most important, we have shown that one can indeed calculate the conditional breakdown point of ℓ_1-regression by solving an appropriate MIP. We give computational experiments using the proposed approach for nine data sets from the robust regression literature. For large data sets, we provide a heuristic that provides an upper bound on the q-strength of a design matrix. This is important in the design of robustly planned experiments, as it provides a computable assessment of the vulnerability of the experiment's design to errors in the measurement on the dependent variable. Finally, we provide an illustrative example to demonstrate the difference between q-strength and s-stability of design matrices.

REFERENCES

[1] D. Alevras and M. Padberg, *Linear Optimization and Extensions: Problems and Solutions*, Springer-Verlag, Berlin, 2001.

[2] D. L. Donoho and P. J. Huber, *The notion of breakdown point*, in A Festschrift for Erich Lehmann, P. Bickel, K. Doksum, and J. L. Hodges, eds., Wadsworth, Belmont, CA, 1983, pp. 157–184.

[3] A. Giloni, *Essays on Optimization in Data Analysis and Operations Management*, Ph.D. Thesis, Stern School of Business, New York University, New York, NY, 2000.

[4] A. Giloni and M. Padberg, *Alternative methods of linear regression*, Math. Comput. Modelling, 35 (2002), pp. 361–374.

[5] F. R. Hampel, *Contributions to the Theory of Robust Estimation*, Ph.D. Thesis, University of California, Berkeley, CA, 1968.

[6] S. P. Ellis and S. Morgenthaler, *Leverage and breakdown in l_1-regression*, J. Amer. Statist. Assoc., 87 (1993), pp. 143–148.

[7] X. He, J. Jureckova, R. Koenker, and S. Portnoy, *Tail behavior of regression estimators and their breakdown points*, Econometrica, 58 (1990), pp. 1195–1214.

[8] J. Hiriart-Urruty and C. Lemaréchal, *Convex Analysis and Minimization Algorithms* I, Springer-Verlag, Berlin, 1993.

[9] R. Koenker and G. Bassett, *Regression quantiles*, Econometrica, 46 (1978), pp. 33–50.

[10] R. L. Mason, R. F. Gunst, and J. L. Hess, *Statistical Design and Analysis of Experiments*, John Wiley, New York, 1989.

[11] I. Mizera and C. H. Müller, *Breakdown points and variation exponents of robust M-estimators in linear models*, Ann. Statist., 27 (1999), pp. 1164–1177.

[12] I. Mizera and C. H. Müller, *The influence of the design on the breakdown point of ℓ_1-type M-estimators*, in MODA6—Advances in Model-Oriented Design and Analysis, A. Atkinson, P. Hackl, and W. Müller, eds., Physica-Verlag, Heidelberg, 2001, pp. 193–200.

[13] M. Padberg, *Linear Optimization and Extensions*, Springer-Verlag, Berlin, 1995.

[14] P. J. Rousseeuw and A. M. Leroy, *Robust Regression and Outlier Detection*, Wiley, New York, 1987.

SIAM J. OPTIM.
Vol. 14, No. 4, pp. 1043–1056

A NONMONOTONE LINE SEARCH TECHNIQUE AND ITS APPLICATION TO UNCONSTRAINED OPTIMIZATION*

HONGCHAO ZHANG† AND WILLIAM W. HAGER†

Abstract. A new nonmonotone line search algorithm is proposed and analyzed. In our scheme, we require that an average of the successive function values decreases, while the traditional nonmonotone approach of Grippo, Lampariello, and Lucidi [*SIAM J. Numer. Anal.*, 23 (1986), pp. 707–716] requires that a maximum of recent function values decreases. We prove global convergence for nonconvex, smooth functions, and R-linear convergence for strongly convex functions. For the L-BFGS method and the unconstrained optimization problems in the CUTE library, the new nonmonotone line search algorithm used fewer function and gradient evaluations, on average, than either the monotone or the traditional nonmonotone scheme.

Key words. nonmonotone line search, R-linear convergence, unconstrained optimization, L-BFGS method

AMS subject classifications. 90C06, 90C26, 65Y20

DOI. 10.1137/S1052623403428208

1. Introduction.

We consider the unconstrained optimization problem

$$\min_{\mathbf{x}\in\Re^n} f(\mathbf{x}), \tag{1.1}$$

where $f : \Re^n \mapsto \Re$ is continuously differentiable. Many iterative methods for (1.1) produce a sequence $\mathbf{x}_0$, $\mathbf{x}_1$, $\mathbf{x}_2$, ..., where $\mathbf{x}_{k+1}$ is generated from $\mathbf{x}_k$, the current direction $\mathbf{d}_k$, and the stepsize $\alpha_k > 0$ by the rule

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k \mathbf{d}_k.$$

In monotone line search methods, α_k is chosen so that $f(\mathbf{x}_{k+1}) < f(\mathbf{x}_k)$. In nonmonotone line search methods, some growth in the function value is permitted. As pointed out by many researchers (for example, see [4, 16]), nonmonotone schemes can improve the likelihood of finding a global optimum; also, they can improve convergence speed in cases where a monotone scheme is forced to creep along the bottom of a narrow curved valley. Encouraging numerical results have been reported [6, 8, 11, 14, 15, 16] when nonmonotone schemes were applied to difficult nonlinear problems.

The earliest nonmonotone line search framework was developed by Grippo, Lampariello, and Lucidi in [7] for Newton's methods. Their approach was roughly the following: Parameters λ_1, λ_2, σ, and δ are introduced where $0 < \lambda_1 < \lambda_2$ and $\sigma, \delta \in (0, 1)$, and they set $\alpha_k = \bar{\alpha}_k \sigma^{h_k}$ where $\bar{\alpha}_k \in (\lambda_1, \lambda_2)$ is the "trial step" and h_k is the smallest nonnegative integer such that

$$f(\mathbf{x}_k + \alpha_k \mathbf{d}_k) \le \max_{0\le j\le m_k} f(\mathbf{x}_{k-j}) + \delta \alpha_k \nabla f(\mathbf{x}_k)\mathbf{d}_k. \tag{1.2}$$

*Received by the editors May 20, 2003; accepted for publication (in revised form) October 2, 2003; published electronically May 25, 2004. This material is based upon work supported by National Science Foundation grant 0203270. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

http://www.siam.org/journals/siopt/14-4/42820.html

†PO Box 118105, Department of Mathematics, University of Florida, Gainesville, FL 32611-8105 (hzhang@math.ufl.edu, http://www.math.ufl.edu/∼hzhang; hager@math.ufl.edu, http://www.math.ufl.edu/∼hager).

Here the gradient of f at $\mathbf{x}_k$, $\nabla f(\mathbf{x}_k)$, is a row vector. The memory m_k at step k is a nondecreasing integer, bounded by some fixed integer M. More precisely,

$$m_0 = 0 \text{ and for } k > 0,\ 0 \le m_k \le \min\{m_{k-1} + 1, M\}.$$

Many subsequent papers, such as [2, 6, 8, 11, 15, 18], have exploited nonmonotone line search techniques of this nature.

Although these nonmonotone techniques based on (1.2) work well in many cases, there are some drawbacks. First, a good function value generated in any iteration is essentially discarded due to the max in (1.2). Second, in some cases, the numerical performance is very dependent on the choice of M (see [7, 15, 16]). Furthermore, it has been pointed out by Dai [4] that although an iterative method is generating R-linearly convergent iterations for a strongly convex function, the iterates may not satisfy the condition (1.2) for k sufficiently large, for any fixed bound M on the memory. Dai's example is

$$f(x) = \frac{1}{2}x^2, \quad x \in \Re, \quad x_0 \neq 0, \quad d_k = -x_k, \quad \text{and} \tag{1.3}$$

$$\alpha_k = \begin{cases} 1 - 2^{-k} & \text{if } k = i^2 \text{ for some integer } i, \\ 2 & \text{otherwise.} \end{cases}$$

The iterates converge R-superlinearly to the minimizer $x^* = 0$; however, condition (1.2) is not satisfied for k sufficiently large and any fixed M.

Our nonmonotone line search algorithm, which was partly studied in the first author's masters thesis [17], has the same general form as the scheme of Grippo, Lampariello, and Lucidi, except that their "max" is replaced by an average of function values. More precisely, our nonmonotone line search algorithm is the following:

Nonmonotone Line Search Algorithm (NLSA).

- **Initialization:** Choose starting guess $\mathbf{x}_0$, and parameters $0 \le \eta_{\min} \le \eta_{\max} \le 1$, $0 < \delta < \sigma < 1 < \rho$, and $\mu > 0$. Set $C_0 = f(\mathbf{x}_0)$, $Q_0 = 1$, and $k = 0$.
- **Convergence test:** If $\|\nabla f(\mathbf{x}_k)\|$ sufficiently small, then stop.
- **Line search update:** Set $\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k \mathbf{d}_k$ where α_k satisfies either the (nonmonotone) Wolfe conditions:

$$f(\mathbf{x}_k + \alpha_k \mathbf{d}_k) \le C_k + \delta \alpha_k \nabla f(\mathbf{x}_k)\mathbf{d}_k, \tag{1.4}$$

$$\nabla f(\mathbf{x}_k + \alpha_k \mathbf{d}_k)\mathbf{d}_k \ge \sigma \nabla f(\mathbf{x}_k)\mathbf{d}_k, \tag{1.5}$$

  or the (nonmonotone) Armijo conditions: $\alpha_k = \bar{\alpha}_k \rho^{h_k}$, where $\bar{\alpha}_k > 0$ is the trial step, and h_k is the largest integer such that (1.4) holds and $\alpha_k \le \mu$.
- **Cost update:** Choose $\eta_k \in [\eta_{\min}, \eta_{\max}]$, and set

$$Q_{k+1} = \eta_k Q_k + 1, \qquad C_{k+1} = (\eta_k Q_k C_k + f(\mathbf{x}_{k+1}))/Q_{k+1}. \tag{1.6}$$

  Replace k by $k + 1$ and return to the convergence test.

Observe that C_{k+1} is a convex combination of C_k and $f(\mathbf{x}_{k+1})$. Since $C_0 = f(\mathbf{x}_0)$, it follows that C_k is a convex combination of the function values $f(\mathbf{x}_0), f(\mathbf{x}_1), \ldots, f(\mathbf{x}_k)$. The choice of η_k controls the degree of nonmonotonicity. If $\eta_k = 0$ for each k, then the line search is the usual monotone Wolfe or Armijo line search. If $\eta_k = 1$ for each k, then $C_k = A_k$, where

$$A_k = \frac{1}{k+1} \sum_{i=0}^{k} f_i, \quad f_i = f(\mathbf{x}_i),$$

is the average function value. The scheme with $C_k = A_k$ was suggested to us by Yu-hong Dai. In [9], the possibility of comparing the current function value with an average of M previous function values was also analyzed; however, since M is fixed, not all previous function values are averaged together as in (1.6). As we show in Lemma 1.1, for any choice of $\eta_k \in [0, 1]$, C_k lies between f_k and A_k, which implies that the line search update is well-defined. As η_k approaches 0, the line search closely approximates the usual monotone line search, and as η_k approaches 1, the scheme becomes more nonmonotone, treating all the previous function values with equal weight when we compute the average cost value C_k.

LEMMA 1.1. *If $\nabla f(\mathbf{x}_k)\mathbf{d}_k \le 0$ for each k, then for the iterates generated by the nonmonotone line search algorithm, we have $f_k \le C_k \le A_k$ for each k. Moreover, if $\nabla f(\mathbf{x}_k)\mathbf{d}_k < 0$ and $f(\mathbf{x})$ is bounded from below, then there exists α_k satisfying either the Wolfe or Armijo conditions of the line search update.*

Proof. Defining $D_k : \Re \to \Re$ by

$$D_k(t) = \frac{tC_{k-1} + f_k}{t+1},$$

we have

$$D_k'(t) = \frac{C_{k-1} - f_k}{(t+1)^2}.$$

Since $\nabla f(\mathbf{x}_k)\mathbf{d}_k \le 0$, it follows from (1.4) that $f_k \le C_{k-1}$, which implies that $D_k'(t) \ge 0$ for all $t \ge 0$. Hence, D_k is nondecreasing, and $f_k = D_k(0) \le D_k(t)$ for all $t \ge 0$. In particular, taking $t = \eta_{k-1}Q_{k-1}$ gives

$$f_k = D_k(0) \le D_k(\eta_{k-1}Q_{k-1}) = C_k. \tag{1.7}$$

This establishes the lower bound for C_k in Lemma 1.1.

The upper bound $C_k \le A_k$ is proved by induction. For $k = 0$, this holds by the initialization $C_0 = f(\mathbf{x}_0)$. Now assume that $C_j \le A_j$ for all $0 \le j < k$. By (1.6), the initialization $Q_0 = 1$, and the fact that $\eta_k \in [0, 1]$, we have

$$Q_{j+1} = 1 + \sum_{i=0}^{j} \prod_{m=0}^{i} \eta_{j-m} \le j + 2. \tag{1.8}$$

Since D_k is monotone nondecreasing, (1.8) implies that

$$C_k = D_k(\eta_{k-1}Q_{k-1}) = D_k(Q_k - 1) \le D_k(k). \tag{1.9}$$

By the induction step,

$$D_k(k) = \frac{kC_{k-1} + f_k}{k+1} \le \frac{kA_{k-1} + f_k}{k+1} = A_k. \tag{1.10}$$

Relations (1.9) and (1.10) imply the upper bound of C_k in Lemma 1.1.

Since both the standard Wolfe and Armijo conditions can be satisfied when $\nabla f(\mathbf{x}_k)\mathbf{d}_k < 0$ and $f(\mathbf{x})$ is bounded from below, and since $f_k \le C_k$, it follows that for each k, α_k can be chosen to satisfy either the Wolfe or the Armijo line search conditions in the nonmonotone line search algorithm. ∎

Our paper is organized as follows: In section 2 we prove global convergence under appropriate conditions on the search directions. In section 3 necessary and sufficient conditions for R-linear convergence are established. In section 4 we implement our scheme in the context of Nocedal's L-BFGS quasi-Newton method [10, 13], and we give numerical comparisons using the unconstrained problems in the CUTE test problem library [3].

2. Global convergence. To begin, we give a lower bound for the step generated by the nonmonotone line search algorithm. Here and elsewhere, $\|\cdot\|$ denotes the Euclidean norm, and $\mathbf{g}_k = \nabla f(\mathbf{x}_k)^\mathsf{T}$, a column vector.

LEMMA 2.1. *Suppose the nonmonotone line search algorithm is employed in a case where* $\mathbf{g}_k^\mathsf{T}\mathbf{d}_k \le 0$ *and* ∇f *satisfies the following Lipschitz conditions with Lipschitz constant* L*:*

1. $\|\nabla f(\mathbf{x}_{k+1}) - \nabla f(\mathbf{x}_k)\| \le L\|\mathbf{x}_{k+1} - \mathbf{x}_k\|$ *if the Wolfe conditions are used, or*
2. $\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{x}_k)\| \le L\|\mathbf{x} - \mathbf{x}_k\|$ *for all* $\mathbf{x}$ *on the line segment connecting* $\mathbf{x}_k$ *and* $\mathbf{x}_k + \alpha_k\rho\mathbf{d}_k$ *if the Armijo condition is used and* $\rho\alpha_k \le \mu$.

If the Wolfe conditions are satisfied, then

$$\alpha_k \ge \left(\frac{1-\sigma}{L}\right)\frac{|\mathbf{g}_k^\mathsf{T}\mathbf{d}_k|}{\|\mathbf{d}_k\|^2}. \tag{2.1}$$

If the Armijo conditions are satisfied, then

$$\alpha_k \ge \min\left\{\frac{\mu}{\rho},\ \left(\frac{2(1-\delta)}{L\rho}\right)\frac{|\mathbf{g}_k^\mathsf{T}\mathbf{d}_k|}{\|\mathbf{d}_k\|^2}\right\}. \tag{2.2}$$

Proof. We consider the lower bounds (2.1) and (2.2) in the following two cases.

Case 1. Suppose that α_k satisfies the Wolfe conditions. By (1.5), we have

$$(\nabla f(\mathbf{x}_k + \alpha_k\mathbf{d}_k) - \nabla f(\mathbf{x}_k))\mathbf{d}_k \ge (\sigma - 1)\nabla f(\mathbf{x}_k)\mathbf{d}_k.$$

Since $\mathbf{g}_k^\mathsf{T}\mathbf{d}_k \le 0$ and $\sigma < 1$, $(\sigma-1)\mathbf{g}_k^\mathsf{T}\mathbf{d}_k \ge 0$, and by the Lipschitz continuity of f,

$$\alpha_k L\|\mathbf{d}_k\|^2 \ge (\sigma - 1)\mathbf{g}_k^\mathsf{T}\mathbf{d}_k,$$

which implies (2.1).

Case 2. Suppose that α_k satisfies the Armijo conditions. If $\rho\alpha_k \ge \mu$, then $\alpha_k \ge \mu/\rho$, which gives (2.2). Conversely, if $\rho\alpha_k < \mu$, then since h_k is the largest integer such that $\alpha_k = \bar{\alpha}_k\rho^{h_k}$ satisfies (1.4) and since $f_k \le C_k$, we have

$$f(\mathbf{x}_k + \rho\alpha_k\mathbf{d}_k) > C_k + \delta\rho\alpha_k\mathbf{g}_k^\mathsf{T}\mathbf{d}_k \ge f(\mathbf{x}_k) + \delta\rho\alpha_k\mathbf{g}_k^\mathsf{T}\mathbf{d}_k. \tag{2.3}$$

When ∇f is Lipschitz continuous,

$$\begin{aligned} f(\mathbf{x}_k + \alpha\mathbf{d}_k) - f(\mathbf{x}_k) &= \alpha\mathbf{g}_k^\mathsf{T}\mathbf{d}_k + \int_0^\alpha [\nabla f(\mathbf{x}_k + t\mathbf{d}_k) - \nabla f(\mathbf{x}_k)]\mathbf{d}_k\, dt \\ &\le \alpha\mathbf{g}_k^\mathsf{T}\mathbf{d}_k + \int_0^\alpha tL\|\mathbf{d}_k\|^2\, dt \\ &= \alpha\mathbf{g}_k^\mathsf{T}\mathbf{d}_k + \frac{1}{2}L\alpha^2\|\mathbf{d}_k\|^2. \end{aligned}$$

Combining this with (2.3) gives (2.2). ∎

Our global convergence result utilizes the following assumption (see, for example, [4, 7]) concerning the search directions.

Direction Assumption. There exist positive constants c_1 and c_2 such that

$$\mathbf{g}_k^\mathsf{T}\mathbf{d}_k \le -c_1\|\mathbf{g}_k\|^2, \tag{2.4}$$

and

$$\|\mathbf{d}_k\| \le c_2\|\mathbf{g}_k\| \tag{2.5}$$

for all sufficiently large k.

THEOREM 2.2. *Suppose $f(\mathbf{x})$ is bounded from below and the direction assumption holds. Moreover, if the Wolfe conditions are used, we assume that ∇f is Lipschitz continuous, with Lipschitz constant L, on the level set*

$$\mathcal{L} = \{\mathbf{x} \in \Re^n : f(\mathbf{x}) \le f(\mathbf{x}_0)\}.$$

Let $\bar{\mathcal{L}}$ denote the collection of $\mathbf{x} \in \Re^n$ whose distance to $\mathcal{L}$ is at most $\mu d_{\max}$, where $d_{\max} = \sup_k \|\mathbf{d}_k\|$. If the Armijo conditions are used, we assume that ∇f is Lipschitz continuous, with Lipschitz constant L, on $\bar{\mathcal{L}}$. Then the iterates $\mathbf{x}_k$ generated by the nonmonotone line search algorithm have the property that

$$\liminf_{k\to\infty} \|\nabla f(\mathbf{x}_k)\| = 0. \tag{2.6}$$

Moreover, if $\eta_{\max} < 1$, then

$$\lim_{k\to\infty} \nabla f(\mathbf{x}_k) = \mathbf{0}. \tag{2.7}$$

Hence, every convergent subsequence of the iterates approaches a point $\mathbf{x}^*$, where $\nabla f(\mathbf{x}^*) = \mathbf{0}$.

Proof. We first show that

$$f_{k+1} \le C_k - \beta\|\mathbf{g}_k\|^2, \tag{2.8}$$

where

$$\beta = \min\left\{\frac{\delta\mu c_1}{\rho},\ \frac{2\delta(1-\delta)c_1^2}{L\rho c_2^2},\ \frac{\delta(1-\sigma)c_1^2}{Lc_2^2}\right\}. \tag{2.9}$$

Case 1. If the Armijo conditions are used and $\rho\alpha_k \ge \mu$, then $\alpha_k \ge \mu/\rho$. By (1.4) and (2.4), it follows that

$$f_{k+1} \le C_k + \delta\alpha_k\mathbf{g}_k^{\mathsf{T}}\mathbf{d}_k \le C_k - \delta\alpha_k c_1\|\mathbf{g}_k^{\mathsf{T}}\|^2 \le C_k - \frac{\delta\mu c_1}{\rho}\|\mathbf{g}_k\|^2,$$

which implies (2.8).

Case 2. If the Armijo conditions are used and $\rho\alpha_k \le \mu$, then by (2.2),

$$\alpha_k \ge \left(\frac{2(1-\delta)}{L\rho}\right)\frac{|\mathbf{g}_k^{\mathsf{T}}\mathbf{d}_k|}{\|\mathbf{d}_k\|^2}, \tag{2.10}$$

and by (1.4), we have

$$f_{k+1} \le C_k - \left(\frac{2\delta(1-\delta)}{L\rho}\right)\left(\frac{\mathbf{g}_k^{\mathsf{T}}\mathbf{d}_k}{\|\mathbf{d}_k\|}\right)^2. \tag{2.11}$$

Finally, by (2.4) and (2.5),

$$f_{k+1} \le C_k - \left(\frac{2\delta(1-\delta)c_1^2}{L\rho c_2^2}\right)\|\mathbf{g}_k\|^2, \tag{2.12}$$

which implies (2.8).

Case 3. If the Wolfe conditions are used, then the analysis is the same as in Case 2, except that the lower bound (2.10) is replaced by the corresponding lower bound (2.1).

Combining the cost update relation (1.6) and the upper bound (2.8),

$$\begin{aligned} C_{k+1} &= \frac{\eta_k Q_k C_k + f_{k+1}}{Q_{k+1}} \\ &\le \frac{\eta_k Q_k C_k + C_k - \beta\|\mathbf{g}_k\|^2}{Q_{k+1}} = C_k - \frac{\beta\|\mathbf{g}_k\|^2}{Q_{k+1}}. \end{aligned} \tag{2.13}$$

Since f is bounded from below and $f_k \le C_k$ for all k, we conclude that C_k is bounded from below. It follows from (2.13) that

$$\sum_{k=0}^{\infty} \frac{\|\mathbf{g}_k\|^2}{Q_{k+1}} < \infty. \tag{2.14}$$

If $\|\mathbf{g}_k\|$ were bounded away from 0, (2.14) would be violated since $Q_{k+1} \le k+2$ by (1.8). Hence, (2.6) holds. If $\eta_{\max} < 1$, then by (1.8),

$$Q_{k+1} = 1 + \sum_{j=0}^{k} \prod_{i=0}^{j} \eta_{k-i} \le 1 + \sum_{j=0}^{k} \eta_{\max}^{j+1} \le \sum_{j=0}^{\infty} \eta_{\max}^{j} = \frac{1}{1-\eta_{\max}}. \tag{2.15}$$

Consequently, (2.14) implies (2.7). ∎

REMARK. *The bound condition $\alpha_k \le \mu$ in the Armijo conditions of the line search update can be removed if ∇f satisfies the Lipschitz condition slightly outside of $\mathcal{L}$. In the proof of Theorem* 2.2*, this bound ensures that when $\rho\alpha_k < \mu$, the point $\mathbf{x}_k + \rho\alpha_k\mathbf{d}_k$ lies in the region $\bar{\mathcal{L}}$, where ∇f is Lipschitz continuous, which is required for establishing Lemma* 2.1*.*

Similar to [4], a slightly different global convergence result is obtained when (2.5) is replaced by the following growth condition on $\mathbf{d}_k$: There exist positive constants τ_1 and τ_2 such that

$$\|\mathbf{d}_k\|^2 \le \tau_1 + \tau_2 k \tag{2.16}$$

for each k.

COROLLARY 2.3. *Suppose $\eta_{\max} < 1$ and all the assumptions of Theorem* 2.2 *are in effect except the direction assumption which is replaced by* (2.4) *and* (2.16)*. If $\tau_2 \ne 0$, then*

$$\liminf_{k\to\infty} \|\nabla f(\mathbf{x}_k)\| = 0. \tag{2.17}$$

If $\tau_2 = 0$, then

$$\lim_{k\to\infty} \|\nabla f(\mathbf{x}_k)\| = 0. \tag{2.18}$$

Proof. We assume, without loss of generality, that $\tau_1 \ge 1$. The analysis is identical to that given in the proof of Theorem 2.2 except that the bound $\|\mathbf{d}_k\| \le c_2\|\mathbf{g}_k\|$ used in the transition from (2.11) to (2.12) is replaced by the bound (2.16). As a result, the inequality (2.8) is replaced by

$$f_{k+1} \le C_k - \left(\frac{\beta_1}{\tau_1 + \tau_2 k}\right)\|\mathbf{g}_k\|^{l_k}, \tag{2.19}$$

where $l_k = 2$ in Case 1, $l_k = 4$ in Cases 2 and 3, and

$$\beta_1 = \min\left\{\frac{\delta\mu c_1}{\rho},\ \frac{2\delta(1-\delta)c_1^2}{L\rho},\ \frac{\delta(1-\sigma)c_1^2}{L}\right\}.$$

Using the upper bound (2.19) for $f(\mathbf{x}_{k+1})$ in the series of inequalities (2.13) gives

$$C_{k+1} \le C_k - \left(\frac{\beta_1}{Q_k(\tau_1+\tau_2 k)}\right)\|\mathbf{g}_k\|^{l_k}.$$

By (2.15),

$$C_{k+1} \le C_k - \left(\frac{\beta_1(1-\eta_{\max})}{\tau_1+\tau_2 k}\right)\|\mathbf{g}_k\|^{l_k}. \tag{2.20}$$

Since f is bounded from below and $C_k \ge f_k$, we obtain (2.17) when $\tau_2 \ne 0$ and (2.18) when $\tau_2 = 0$. This completes the proof. □

3. Linear convergence. In [4] Dai proves R-linear convergence for the nonmonotone max-based line search scheme (1.2), when the cost function is strongly convex. Similar to [4], we now establish R-linear convergence for our nonmonotone line search algorithm when f is strongly convex. Recall that f is strongly convex if there exists a scalar $\gamma > 0$ such that

$$f(\mathbf{x}) \ge f(\mathbf{y}) + \nabla f(\mathbf{y})(\mathbf{x}-\mathbf{y}) + \frac{1}{2\gamma}\|\mathbf{x}-\mathbf{y}\|^2 \tag{3.1}$$

for all $\mathbf{x}$ and $\mathbf{y} \in \Re^n$. After interchanging $\mathbf{x}$ and $\mathbf{y}$ and adding,

$$(\nabla f(\mathbf{x}) - \nabla f(\mathbf{y}))(\mathbf{x}-\mathbf{y}) \ge \frac{1}{\gamma}\|\mathbf{x}-\mathbf{y}\|^2. \tag{3.2}$$

If $\mathbf{x}^*$ denotes the unique minimizer of f, it follows from (3.2), with $\mathbf{y} = \mathbf{x}^*$, that

$$\|\mathbf{x}-\mathbf{x}^*\| \le \gamma\|\nabla f(\mathbf{x})\|. \tag{3.3}$$

For $t \in [0,1]$, define $\mathbf{x}(t) = \mathbf{x}^* + t(\mathbf{x}-\mathbf{x}^*)$. Since f is convex, $f(\mathbf{x}(t))$ is a convex function of t, and the derivative $f'(\mathbf{x}(t))$ is an increasing function of $t \in [0,1]$ with $f'(\mathbf{x}(0)) = 0$. Hence, for $t \in [0,1]$, $f'(\mathbf{x}(t))$ attains its maximum value at $t = 1$. This observation combined with (3.3) gives

$$\begin{aligned} f(\mathbf{x}) - f(\mathbf{x}^*) &= \int_0^1 f'(\mathbf{x}(t))dt \le f'(\mathbf{x}(1)) = \nabla f(\mathbf{x})(\mathbf{x}-\mathbf{x}^*) \\ &\le \|\nabla f(\mathbf{x})\|\|\mathbf{x}-\mathbf{x}^*\| \le \gamma\|\nabla f(\mathbf{x})\|^2. \end{aligned} \tag{3.4}$$

THEOREM 3.1. *Suppose that f is strongly convex with unique minimizer $\mathbf{x}^*$, the search directions $\mathbf{d}_k$ in the nonmonotone line search algorithm satisfy the direction assumption, there exist $\mu > 0$ such that $\alpha_k \le \mu$ for all k, $\eta_{\max} < 1$, and ∇f is Lipschitz continuous on bounded sets. Then there exists $\theta \in (0,1)$ such that*

$$f(\mathbf{x}_k) - f(\mathbf{x}^*) \le \theta^k(f(\mathbf{x}_0) - f(\mathbf{x}^*)) \tag{3.5}$$

for each k.

Proof. Since $f(\mathbf{x}_{k+1}) \le C_k$ and C_{k+1} is a convex combination of C_k and $f(\mathbf{x}_{k+1})$, we have $C_{k+1} \le C_k$ for each k. Hence,

$$f(\mathbf{x}_{k+1}) \le C_k \le C_{k-1} \le \cdots \le C_0 = f(\mathbf{x}_0),$$

which implies that all the iterates $\mathbf{x}_k$ are contained in the level set

$$\mathcal{L} = \{\mathbf{x} \in \Re^n : f(\mathbf{x}) \le f(\mathbf{x}_0)\}.$$

Since f is strongly convex, it follows that $\mathcal{L}$ is bounded and ∇f is Lipschitz continuous on $\mathcal{L}$. By the direction assumption and the fact that $\|\nabla f(\mathbf{x})\|$ is bounded on $\mathcal{L}$, $d_{\max} = \sup_k \|\mathbf{d}_k\| < \infty$. Let $\bar{\mathcal{L}}$ denote the collection of $\mathbf{x} \in \Re^n$ whose distance to $\mathcal{L}$ is at most $\mu d_{\max}$ and let L be a Lipschitz constant for ∇f on the $\bar{\mathcal{L}}$.

As shown in the proof of Theorem 2.2,

$$f(\mathbf{x}_{k+1}) \le C_k - \beta \|\mathbf{g}_k\|^2, \tag{3.6}$$

where β is given in (2.9). Also, by the direction assumption and the upper bound μ on α_k, $\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k \mathbf{d}_k$ satisfies

$$\|\mathbf{x}_{k+1} - \mathbf{x}_k\| = \alpha_k \|\mathbf{d}_k\| \le \mu c_2 \|\mathbf{g}_k\|.$$

Combining this with the Lipschitz continuity of ∇f gives

$$\|\nabla f(\mathbf{x}_{k+1}) - \nabla f(\mathbf{x}_k)\| = \|\mathbf{g}_{k+1} - \mathbf{g}_k\| \le L\|\mathbf{x}_{k+1} - \mathbf{x}_k\| \le \mu c_2 L \|\mathbf{g}_k\|,$$

from which it follows that

$$\|\mathbf{g}_{k+1}\| \le \|\mathbf{g}_{k+1} - \mathbf{g}_k\| + \|\mathbf{g}_k\| \le b\|\mathbf{g}_k\|, \quad b = 1 + \mu c_2 L. \tag{3.7}$$

We now show that for each k,

$$C_{k+1} - f(\mathbf{x}^*) \le \theta (C_k - f(\mathbf{x}^*)), \tag{3.8}$$

where

$$\theta = 1 - \beta b_2 (1 - \eta_{\max}) \quad \text{and} \quad b_2 = \frac{1}{\beta + \gamma b^2}.$$

This immediately yields (3.5) since $f(\mathbf{x}_k) \le C_k$ and $C_0 = f(\mathbf{x}_0)$.

Case 1. $\|\mathbf{g}_k\|^2 \ge b_2 (C_k - f(\mathbf{x}^*))$. By the cost update formula (1.6), we have

$$C_{k+1} - f(\mathbf{x}^*) = \frac{\eta_k Q_k (C_k - f(\mathbf{x}^*)) + (f_{k+1} - f(\mathbf{x}^*))}{1 + \eta_k Q_k}. \tag{3.9}$$

Utilizing (3.6) gives

$$\begin{aligned} C_{k+1} - f(\mathbf{x}^*) &\le \frac{\eta_k Q_k (C_k - f(\mathbf{x}^*)) + (C_k - f(\mathbf{x}^*)) - \beta \|\mathbf{g}_k\|^2}{1 + \eta_k Q_k} \\ &= C_k - f(\mathbf{x}^*) - \frac{\beta \|\mathbf{g}_k\|^2}{Q_{k+1}}. \end{aligned}$$

Since $Q_{k+1} \le 1/(1 - \eta_{\max})$ by (2.15), it follows that

$$C_{k+1} - f(\mathbf{x}^*) \le C_k - f(\mathbf{x}^*) - \beta(1 - \eta_{\max}) \|\mathbf{g}_k\|^2.$$

Since $\|\mathbf{g}_k\|^2 \ge b_2 (C_k - f(\mathbf{x}^*))$, (3.8) has been established in Case 1.

Case 2. $\|\mathbf{g}_k\|^2 < b_2(C_k - f(\mathbf{x}^*))$. By (3.4) and (3.7), we have

$$f(\mathbf{x}_{k+1}) - f(\mathbf{x}^*) \le \gamma\|\mathbf{g}_{k+1}\|^2 \le \gamma b^2\|\mathbf{g}_k\|^2.$$

And by the Case 2 bound for $\|\mathbf{g}_k\|$, this gives

$$f(\mathbf{x}_{k+1}) - f(\mathbf{x}^*) \le \gamma b^2 b_2(C_k - f(\mathbf{x}^*)).$$

Inserting this bound for $f(\mathbf{x}_{k+1}) - f(\mathbf{x}^*)$ in (3.9) yields

$$\begin{aligned} C_{k+1} - f(\mathbf{x}^*) &\le \frac{(\eta_k Q_k + \gamma b^2 b_2)(C_k - f(\mathbf{x}^*))}{1 + \eta_k Q_k} \\ &= \left(1 - \frac{1 - \gamma b^2 b_2}{Q_{k+1}}\right)(C_k - f(\mathbf{x}^*)). \end{aligned} \tag{3.10}$$

Rearranging the expression for b_2, we have $\gamma b^2 b_2 = 1 - \beta b_2$. Inserting this relation in (3.10) and again utilizing the bound (2.15), we obtain (3.8).

This completes the proof of (3.8), and as indicated above, the linear convergence estimate (3.5) follows directly. ∎

In the introduction, example (1.3) revealed that linearly convergent iterates may not satisfy (1.2) for any fixed choice of the memory M. We now show that with our choice for C_k, we can always satisfy (1.4), when k is sufficiently large, provided η_k is close enough to 1. We begin with a lower bound for $f(\mathbf{x}) - f(\mathbf{x}^*)$, analogous to the upper bound (3.4). By (3.1) with $\mathbf{y} = \mathbf{x}^*$, we have

$$f(\mathbf{x}) - f(\mathbf{x}^*) \ge \frac{1}{2\gamma}\|\mathbf{x} - \mathbf{x}^*\|^2. \tag{3.11}$$

If ∇f satisfies the Lipschitz condition

$$\|\nabla f(\mathbf{x})\| = \|\nabla f(\mathbf{x}) - \nabla f(\mathbf{x}^*)\| \le L\|\mathbf{x} - \mathbf{x}^*\|,$$

then (3.11) gives

$$f(\mathbf{x}) - f(\mathbf{x}^*) \ge \frac{1}{2\gamma L^2}\|\nabla f(\mathbf{x})\|^2. \tag{3.12}$$

THEOREM 3.2. *Let $\mathbf{x}^*$ denote a minimizer of f and suppose that the sequence $f(\mathbf{x}_k)$, $k = 0, 1, \ldots$, converges R-linearly to $f(\mathbf{x}^*)$; that is, there exist constants $\theta \in (0,1)$ and c such that $f(\mathbf{x}_k) - f(\mathbf{x}^*) \le c\theta^k$. Assume that the $\mathbf{x}_k$ are contained in a closed, bounded convex set K, f is strongly convex on K, satisfying* (3.1), *∇f is Lipschitz continuous on K, with Lipschitz constant L, the direction assumption holds, and the stepsize α_k is bounded by a constant μ. If $\eta_{\min} > \theta$, then* (1.4) *is satisfied for k sufficiently large, where C_k is given by the recursion* (1.6).

Proof. By (3.9) and the bound $Q_k \le k + 1$ (see (1.8)), we have

$$\begin{aligned} C_k - f(\mathbf{x}^*) &= \frac{\sum_{i=0}^{k}\left[(\prod_{j=i}^{k-1}\eta_j)(f(\mathbf{x}_i) - f(\mathbf{x}^*))\right]}{Q_k} \\ &\ge \frac{\prod_{j=0}^{k-1}\eta_j}{k+1}\sum_{i=0}^{k}\left[\frac{f(\mathbf{x}_i) - f(\mathbf{x}^*)}{\prod_{j=0}^{i-1}\eta_j}\right] \\ &\ge \frac{(\eta_{\min})^k}{k+1}\phi_k, \text{ where } \phi_k = \sum_{i=0}^{k}\frac{f(\mathbf{x}_i) - f(\mathbf{x}^*)}{\prod_{j=0}^{i-1}\eta_j}. \end{aligned} \tag{3.13}$$

Here we define a product $\prod_{j=i}^{k-1}\eta_j$ to be 1 whenever the range of indices is vacuous; in particular, $\prod_{j=k}^{k-1}\eta_j = 1$. Let Φ denote the limit (possibly $+\infty$) of the positive, monotone increasing sequence $\phi_0, \phi_1, \ldots$.

By the direction assumption and (3.12), we have

$$\alpha_k \mathbf{g}_k^{\mathsf{T}}\mathbf{d}_k \geq -\mu c_2 \|\mathbf{g}_k\|^2 \geq -2\gamma\mu c_2 L^2 (f(\mathbf{x}_k) - f(\mathbf{x}^*)). \tag{3.14}$$

Combining the R-linear convergence of $f(\mathbf{x}_k)$ to $f(\mathbf{x}^*)$ with (3.14) gives

$$\begin{aligned} f(\mathbf{x}_{k+1}) - f(\mathbf{x}^*) - \delta\alpha_k \mathbf{g}_k^{\mathsf{T}}\mathbf{d}_k &\leq c\theta^{k+1} - \delta\alpha_k \mathbf{g}_k^{\mathsf{T}}\mathbf{d}_k \\ &\leq c\theta^k(\theta + 2\gamma\mu c_2 L^2). \end{aligned} \tag{3.15}$$

Comparing (3.13) with (3.15), it follows that when

$$\frac{\Phi}{k+1} \geq c\left(\frac{\theta}{\eta_{\min}}\right)^k (\theta + 2\gamma\mu c_2 L^2), \tag{3.16}$$

(1.4) is satisfied. Since $\eta_{\min} > \theta$, the inequality (3.16) holds for k sufficiently large, and the proof is complete. □

As a consequence of Theorem 3.2, the iterates of example (1.3) satisfy the Wolfe condition (1.4) for k sufficiently large, when $\eta_k = 1$ for all k.

4. Numerical comparisons. In this section we compare three methods:

(i) the monotone line search, corresponding to $\eta_k = 0$ in the nonmonotone line search algorithm;
(ii) the nonmonotone scheme [7] based on a maximum of recent function values;
(iii) the new nonmonotone line search algorithm based on an average function value.

In our implementation, we chose the stepsize α_k to satisfy the Wolfe conditions with $\delta = 10^{-4}$ and $\sigma = .9$. For the monotone line search scheme (i), C_k in (1.4) is replaced by $f(\mathbf{x}_k)$; in the nonmonotone scheme (ii) based on the maximum of recent function values, C_k in (1.4) is replaced by

$$\max_{0\leq j\leq m_k} f(\mathbf{x}_{k-j}).$$

As recommended in [7], we set $m_0 = 0$ and $m_k = \min\{m_{k-1} + 1, 10\}$ for $k > 0$. Although our best convergence results were obtained by dynamically varying η_k, using values closer to 1 when the iterates were far from the optimum, and using values closer to 0 when the iterates were near an optimum, the numerical experiments reported here employ a fixed value $\eta_k = .85$, which seemed to work reasonably well for a broad class of problems.

The search directions were generated by the L-BFGS method developed by Nocedal in [13] and Liu and Nocedal in [10]; their software is available from the web page http://www.ece.northwestern.edu/∼nocedal/software.html.

We now briefly summarize how the search directions are generated: $\mathbf{d}_k = -\mathbf{B}_k^{-1}\mathbf{g}_k$, where the matrices $\mathbf{B}_k$ are given by the update

$$\begin{aligned} \mathbf{B}_{k-1}^{(0)} &= \gamma_k \mathbf{I}, \\ \mathbf{B}_{k-1}^{(l+1)} &= \mathbf{B}_{k-1}^{(l)} - \frac{\mathbf{B}_{k-1}^{(l)}\mathbf{s}_l\mathbf{s}_l^{\mathsf{T}}\mathbf{B}_{k-1}^{(l)}}{\mathbf{s}_l^{\mathsf{T}}\mathbf{B}_{k-1}^{(l)}\mathbf{s}_l} + \frac{\mathbf{y}_l^{\mathsf{T}}\mathbf{y}_l}{\mathbf{y}_l\mathbf{s}_l}, \quad l = 0, 1, \ldots, M_k - 1, \\ \mathbf{B}_k &= \mathbf{B}_{k-1}^{M_k}. \end{aligned}$$

We took $M_k = \min\{k, 5\}$,

$$\mathbf{y}_l = \mathbf{g}_{j_l+1} - \mathbf{g}_{j_l}, \quad \mathbf{s}_l = \mathbf{x}_{j_l+1} - \mathbf{x}_{j_l}, \quad j_l = k - M_k + l,$$

and

$$\gamma_k = \begin{cases} \frac{\|\mathbf{y}_{k-1}\|^2}{\mathbf{y}_{k-1}\mathbf{s}_{k-1}} & \text{if} \quad k > 0, \\ 1 & \text{if} \quad k = 0. \end{cases}$$

The analysis in [10] reveals that when f is twice continuously differentiable and strongly convex, with the norm of the Hessian uniformly bounded, $\mathbf{B}_k^{-1}$ is uniformly bounded, which implies that the direction assumption is satisfied.

Our numerical experiments use double precision versions of the unconstrained optimization problems in the CUTE library [3]. Altogether, there were 80 problems. Our stopping criterion was

$$\|\nabla f(\mathbf{x}_k)\|_\infty \le 10^{-6}(1 + |f(\mathbf{x}_k)|), \quad \|\mathbf{y}\|_\infty = \max_{1 \le i \le n} |y_i|,$$

except for problems PENALTY1, PENALTY2, and QUARTC, which would stop at $k = 0$ with this criterion. For these three problems, the stopping criterion was

$$\|\nabla f(\mathbf{x}_k)\|_\infty \le 10^{-8}\|\nabla f(\mathbf{x}_0)\|_\infty.$$

In Tables 4.1 and 4.2, we give the dimension (Dim) of each test problem, the number n_i of iterations, and the number n_f of function or gradient evaluations. An "F" in the table means that the line search could not be satisfied. The line search routine in the L-BFGS code, according to the documentation, is a slight modification of the code CSRCH of Moré and Thuente. In the cases where the line search failed, it reported that "Rounding errors prevent further progress. There may not be a step which satisfies the sufficient decrease and curvature conditions. Tolerances may be too small." Basically, it was not possible to satisfy the first Wolfe condition (1.4) due to rounding errors. With our nonmonotone line search algorithm, on the other hand, the value of C_k was a bit larger than either the function value $f(\mathbf{x}_k)$ used in the monotone scheme (i) or the local maximum used in (ii). As a result, we were able to satisfy (1.4) using the Moré and Thuente code, despite rounding errors, in cases where the other schemes were not successful.

We now give an overview of the numerical results reported in Tables 4.1 and 4.2. First, in many cases, the numbers of function and gradient evaluations of the three line search algorithms are identical. When comparing the monotone scheme (i) to the nonmonotone schemes (ii) and (iii), we see that either of the nonmonotone schemes was superior to the monotone scheme. In particular, there were

- 20 problems where monotone (i) was superior to nonmonotone (ii),
- 35 problems where nonmonotone (ii) was superior to monotone (i),
- 15 problems where monotone (i) was superior to nonmonotone (iii),
- 43 problems where nonmonotone (iii) was superior to monotone (i).

When comparing the nonmonotone schemes, we see that the new nonmonotone line search algorithm (iii) was superior to the previous, max-based scheme (ii). In particular, there were

- 10 problems where (ii) was superior to (iii),
- 20 problems where (iii) was superior to (ii).

As the test problems were solved, we tabulated the number of iterations where the function increased in value. We found that for either of the nonmonotone schemes (ii) or (iii), in roughly 7% of the iterations, the function value increased.

TABLE 4.1
Numerical comparisons.

Problem name	Dim	Monotone (i)		Maximum (ii)		Average (iii)	
		n_i	n_f	n_i	n_f	n_i	n_f
ARGLINA	500	2	4	2	4	2	4
ARGLINB	500	F	F	F	F	35	44
ARGLINC	500	F	F	F	F	74	111
ARWHEAD	10000	12	15	12	14	12	14
BDQRTIC	5000	129	156	180	200	162	175
BROWNAL	400	6	14	6	14	6	14
BROYDN7D	2000	662	668	660	662	660	662
BRYBND	5000	29	32	38	41	38	41
CHAINWOO	800	3578	3811	3503	3530	3223	3258
CHNROSNB	50	295	308	313	315	298	300
COSINE	1000	11	16	12	16	12	16
CRAGGLVY	5000	61	68	59	63	59	63
CURLY10	1000	990	1024	1302	1310	1482	1488
CURLY20	1000	2392	2462	2019	2025	2322	2325
CURLY30	1000	3034	3123	3052	3060	2677	2683
DECONVU	61	605	634	324	326	324	326
DIXMAANA	3000	11	13	11	13	11	13
DIXMAANB	3000	11	13	11	13	11	13
DIXMAANC	6000	12	14	12	14	12	14
DIXMAAND	6000	14	16	14	16	14	16
DIXMAANE	6000	355	368	341	343	341	343
DIXMAANF	6000	284	295	258	260	258	260
DIXMAANG	6000	300	307	297	299	297	299
DIXMAANH	6000	294	305	303	305	303	305
DIXMAANI	6000	2355	2426	2616	2618	2576	2579
DIXMAANJ	6000	251	259	272	274	272	274
DIXMAANK	6000	258	266	220	222	220	222
DIXMAANL	6000	215	220	190	192	190	192
DIXON3DQ	800	4733	4874	4515	4516	4353	4356
DQDRTIC	10000	14	23	11	17	11	17
EDENSCH	5000	22	27	28	31	28	31
EG2	1000	4	5	4	5	4	5
EIGENALS	420	4377	4549	4016	4031	4381	4396
EIGENBLS	420	4572	4698	4214	4226	4288	4301
EIGENCLS	462	3327	3416	3615	3623	3615	3623
ENGVAL1	10000	14	17	14	17	14	17
ERRINROS	50	160	176	184	191	154	162
EXTROSNB	50	13789	17217	10128	10658	10606	11427
FLETCBV2	1000	1223	1265	1419	1420	1284	1286
FLETCBV3	1000	3	11	3	11	3	11

TABLE 4.2
Numerical comparisons (continued).

Problem name	Dim	Monotone (i)		Maximum (ii)		Average (iii)	
		n_i	n_f	n_i	n_f	n_i	n_f
FLETCHBV	500	2	10	2	10	2	10
FLETCHCR	5000	25245	27605	26449	26553	26257	26515
FMINSRF2	10000	385	395	387	389	387	389
FMINSURF	10000	601	611	686	688	686	688
FREUROTH	5000	16	23	16	22	16	22
GENHUMPS	1000	1892	2418	1978	2168	1944	2187
GENROSE	2000	4169	4510	4387	4444	4309	4380
HILBERTA	200	356	388	237	243	365	371
HILBERTB	200	7	9	7	9	7	9
INDEF	500	2	10	2	10	2	10
JIMACK	82	4423	4644	5531	5552	3892	3912
LIARWHD	10000	26	30	28	32	31	34
MANCINO	100	11	15	11	15	11	15
MOREBV	10000	74	77	77	79	77	79
NCB20	3010	429	474	337	347	316	323
NONCVXU2	1000	1227	1262	1583	1591	1583	1591
NONCVXUN	1000	1936	1987	1657	1664	1657	1664
NONDIA	10000	21	27	21	26	21	26
NONDQUAR	10000	3331	3685	3625	3751	3315	3444
PENALTY1	10000	23	31	23	31	23	31
PENALTY2	200	F	F	131	136	130	133
PENALTY3	200	F	F	F	F	73	107
POWELLSG	10000	55	63	59	62	68	71
POWER	5000	297	305	302	304	302	304
QUARTC	10000	23	31	23	31	23	31
SCHMVETT	10000	20	25	21	23	21	23
SENSORS	200	25	29	26	29	26	29
SINQUAD	5000	267	329	319	371	366	431
SPARSINE	1000	6692	6989	7173	7176	6220	6227
SPARSQUR	10000	34	39	35	37	35	37
SPMSRTLS	10000	245	260	243	250	243	250
SROSENBR	10000	17	20	17	20	18	20
TESTQUAD	2000	6431	6628	4549	4551	4456	4462
TOINTGOR	50	88	94	92	93	92	93
TOINTGSS	10000	17	22	17	22	17	22
TQUARTIC	10000	24	29	25	29	25	29
TRIDIA	10000	2781	2860	2977	2980	2637	2641
VARDIM	10000	1	2	1	2	1	2
VAREIGVL	5000	18	21	18	20	18	20
WOODS	10000	15	20	21	24	21	24

Acknowledgments. The first author thanks Yu-hong Dai for interesting discussions concerning the relations between monotone and nonmonotone optimization methods. The authors gratefully acknowledge the comments and suggestions of the referees.

REFERENCES

[1] J. Barzilai and J. M. Borwein, *Two-point step size gradient methods*, IMA J. Numer. Anal., 8 (1988), pp. 141–148.
[2] E. G. Birgin, J. M. Martinez, and M. Raydan, *Nonmonotone spectral projected gradient methods on convex sets*, SIAM J. Optim., 10 (2000), pp. 1196–1211.
[3] I. Bongartz, A. R. Conn, N. I. M. Gould, and P. L. Toint, *CUTE: Constrained and unconstrained testing environments*, ACM Trans. Math. Software, 21 (1995), pp. 123–160.
[4] Y. H. Dai, *On the nonmonotone line search*, J. Optim. Theory Appl., 112 (2002), pp. 315–330.
[5] Y. H. Dai, *R-linear convergence of the Barzilai and Borwein gradient method*, IMA J. Numer. Anal., 22 (2002), pp. 1–10.
[6] Y. H. Dai, *A nonmonotone conjugate gradient algorithm for unconstrained optimization*, J. Syst. Sci. Complex., 15 (2002), pp. 139–145.
[7] L. Grippo, F. Lampariello, and S. Lucidi, *A nonmonotone line search technique for Newton's method*, SIAM J. Numer. Anal., 23 (1986), pp. 707–716.
[8] L. Grippo, F. Lampariello, and S. Lucidi, *A truncated Newton method with nonmonotone line search for unconstrained optimization*, J. Optim. Theory Appl., 60 (1989), pp. 401–419.
[9] L. Grippo, F. Lampariello, and S. Lucidi, *A class of nonmonotone stabilization methods in unconstrained optimization*, Numer. Math., 59 (1991), pp. 779–805.
[10] D. C. Liu and J. Nocedal, *On the limited memory BFGS method for large scale optimization*, Math. Prog., 45 (1989), pp. 503–528.
[11] S. Lucidi, F. Rochetich, and M. Roma, *Curvilinear stabilization techniques for truncated Newton methods in large-scale unconstrained optimization*, SIAM J. Optim., 8 (1998), pp. 916–939.
[12] J. J. Moré, B. S. Garbow, and K. E. Hillstrom, *Testing unconstrained optimization software*, ACM Trans. Math. Software, 7 (1981), pp. 17–41.
[13] J. Nocedal, *Updating quasi-Newton matrices with limited storage*, Math. Comp., 35 (1980), pp. 773–782.
[14] E. R. Panier and A. L. Tits, *Avoiding the Maratos effect by means of a nonmonotone linesearch*, SIAM J. Numer. Anal., 28 (1991), pp. 1183–1195.
[15] M. Raydan, *The Barzilai and Borwein gradient method for the large scale unconstrained minimization problem*, SIAM J. Optim., 7 (1997), pp. 26–33.
[16] P. L. Toint, *An assessment of non-monotone line search techniques for unconstrained optimization*, SIAM J. Sci. Comput., 17 (1996), pp. 725–739.
[17] H. Zhang, *A Few Nonmonotone Methods in Nonlinear Optimization*, Master's Thesis, ICMSEC, Chinese Academy of Sciences, Beijing, China, 2001 (in Chinese).
[18] J. L. Zhou and A. L. Tits, *Nonmonotone line search for minimax problem*, J. Optim. Theory Appl., 76 (1993), pp. 455–476.

SIAM J. OPTIM.
Vol. 14, No. 4, pp. 1057–1073

ON FIRST- AND SECOND-ORDER CONDITIONS FOR ERROR BOUNDS*

L. R. HUANG† AND K. F. NG‡

Abstract. We establish first-order and second-order sufficient conditions ensuring that a proper lower semicontinuous function f on a Banach space X has an error bound. We also consider similar problems with constraint, namely, that f is replaced by its restriction to a subset of X. These results are employed to identify exactly when a quadratic function on X has an error bound.

Key words. error bound, Hadamard directional derivative, second-order directional derivative, complete metric space, lower semicontinuous function

AMS subject classifications. 90C31, 90C25, 49J52

DOI. 10.1137/S1052623401390549

1. Introduction.

Let X be a Banach space and $f : X \to R \cup \{+\infty\}$ a proper lower semicontinuous function. Let $\lambda \in R$ with $\lambda \geqslant \inf f$ ($:= \inf_{x\in X} f(x)$). Let $L_f(x) := \{x \in X; f(x) \leqslant \lambda\}$. Assuming $L_f(\lambda) \neq \emptyset$, we say that f has a (Lipschitz) error bound $\delta > 0$ for $L_f(\lambda)$ if the distance of x to $L_f(\lambda)$ satisfies the inequality

$$\delta d(x, L_f(\lambda)) \leqslant f(x) - \lambda \quad \forall x \notin L_f(\lambda). \tag{1.1}$$

Many authors have studied this problem (see [4], [8], [10], [11], [12], [15], [16], [13], [14], and the references therein). In this paper, we study the error bound problem by using the Hadamard directional derivative $d_-f(x;u)$ and the second-order directional derivatives. Let $\mathfrak{D}$ denote the set of all x satisfying $d_-f(x;0) = 0$. We show in Theorem 2.5 that $\delta > 0$ is an error bound for $L_f(\lambda)$ if

$$\sup_{x\in\mathfrak{D}\setminus L_f(\lambda)} \inf_{\|u\|=1} d_-f(x;u) \leqslant -\delta. \tag{1.2}$$

The consideration of the Hadamard derivative instead of the Dini derivative provides not only a better sufficient condition than the one provided in [13] but also a better tool in dealing with the constraint problems (see Theorems 2.6 and 2.7). If the above first-order sufficient condition (1.2) fails to apply, then one may look at second-order conditions. In section 3, we establish a second-order Taylor expansion (in an inequality form) for lower semicontinuous functions, and this expansion is then applied to provide a second-order sufficient condition for error bounds. This result appears to be new even for C^2-functions (see Corollary 3.3). In section 4, we study quadratic functions f on X and identify exactly when f has a Lipschitz error bound, and our classification is more concise than the corresponding results in [13] given for $X = R^n$.

*Received by the editors June 5, 2001; accepted for publication (in revised form) October 17, 2003; published electronically May 25, 2004. This research was supported by the United College of the Chinese University of Hong Kong, a direct grant (CUHK), and an earmarked grant from the Research Grant Council of Hong Kong.

http://www.siam.org/journals/siopt/14-4/39054.html

†Department of Mathematics, South China Normal University, Guang Zhou, The People's Republic of China (huanglr@scnu.edu.cn). The research of this author was partially supported by NNSF (10171034), China.

‡Department of Mathematics, The Chinese University of Hong Kong, Shatin, N.T., Hong Kong (kfng@math.cuhk.edu.hk).

2. First-order sufficient conditions for the existence of an error bound. Let Z be a complete metric space and S a closed subset of Z. Let $f : Z \to R \cup \{+\infty\}$ be a lower semicontinuous function. Let $\delta > 0$. Define $M_f(x, \delta)$ by

$$M_f(x, \delta) := \{y \in Z; f(x) \geqslant f(y) + \delta d(x, y)\}.$$

Let

$$O_f(\delta) := \{x \in Domf; M_f(x, \delta) = \{x\}\}.$$

When $\delta = 1$, we denote $M_f(x, 1)$ by $M_f(x)$ and $O_f(1)$ by O_f for short. The following result is similar to [8] and [13] but we do not assume that S is a level set of f.

PROPOSITION 2.1. *Let Z be a complete metric space and S a nonempty closed subset of X. Let $f : Z \to R \cup \{+\infty\}$ be a lower semicontinuous function bounded from below. Suppose that $O_f(\delta) \subseteq S$. Then for any $x \in Domf \backslash S$, there exists $z \in S$ such that*

$$\delta d(x, z) \leqslant f(x) - f(z).$$

Proof. We suppose without loss of generality that $\delta = 1$. Let $x \in Z \backslash S$, and assume that $f(x) < +\infty$. It is sufficient to show that

$$M_f(x) \cap S \neq \emptyset.$$

Note that $M_f(x)$ is closed as f is lower semicontinuous. By Ekeland's variational principle (cf. [3, Theorem 7.5.1]), there exists $z \in M_f(x)$ such that

$$f(y) + d(y, z) > f(z) \quad \forall y \in M_f(x) \backslash \{z\}. \tag{2.1}$$

We will show that $z \in S$ (and hence complete the proof). By way of contradiction we assume that $z \notin S$. Then, by assumption, $z \notin O_f$ and hence there exists $w \in M_f(z) \backslash \{z\}$ such that

$$f(z) - f(w) \geqslant d(z, w). \tag{2.2}$$

Since $w \in M_f(z)$ and $z \in M_f(x)$, we have $w \in M_f(x)$ and so the inequality (2.1) holds for $y = w$. But this is impossible since the inequality (2.2) contradicts (2.1) for $y = w$. □

In terms of level set $L_f(\lambda)$ of f, we have the following result [13, Lemma 2.3].

COROLLARY 2.2. *Let Z be a complete metric space and $f : Z \to R \cup \{+\infty\}$ a proper lower semicontinuous function. Let $\delta > 0, \lambda \in R$ be such that $L_f(\lambda) \neq \emptyset$ and for any $x \in Domf \backslash L_f(\lambda)$, there exists $y \in f^{-1}[\lambda, +\infty)$ with*

$$f(x) - f(y) \geqslant \delta d(x, y) > 0. \tag{2.3}$$

Then one has

$$f(x) - \lambda \geqslant \delta d(x, L_f(\lambda)) \quad \forall x \in Z \backslash L_f(\lambda).$$

Proof. Since f is lower semicontinuous, $L_f(\lambda)$ is closed. Let g be defined by $g(x) = \max\{\lambda, f(x)\}$. Then $g : Z \to [\lambda, +\infty]$ is lower semicontinuous and $L_g(\lambda) = L_f(\lambda)$. By the assumption (2.3) one has $O_g(\delta) \subseteq L_g(\lambda)$. It follows from Proposition 2.1 that for any $x \in Z \backslash L_g(\lambda)$ there exists $y \in L_g(\lambda)$ such that

$$g(x) - g(y) = g(x) - \lambda \geqslant \delta d(x, L_g(\lambda)).$$

Hence, we have

$$f(x) - \lambda \geqslant \delta d(x, L_f(\lambda)). \qquad \square$$

Throughout the entire paper, X denotes a Banach space and $f : X \to R \cup \{+\infty\}$ is a lower semicontinuous function. Suppose that f is finite at a point $x \in X$. Let $u \in X$. We use $d_-f(x;u)$ and $D_-f(x;u)$ to denote, respectively, Hadamard's and Dini's lower directional derivatives of f at x in the direction u. They are defined by

$$d_-f(x;u) := \liminf_{u' \to u, t \downarrow 0} \frac{1}{t}(f(x+tu') - f(x))$$

and

$$D_-f(x;u) := \liminf_{t \downarrow 0} \frac{1}{t}(f(x+tu) - f(x)).$$

Thus, $d_-f \leqslant D_-f$. Let $\mathfrak{D}$ or $\mathfrak{D}_f$ denote the set of all x such that $d_-f(x;0) > -\infty$. By the homogeneity of $d_-f(x;\cdot)$ we have

$$\mathfrak{D} = \{x \in Dom f; d_-f(x;0) = 0\}. \tag{2.4}$$

LEMMA 2.3. *Let $x \in X$ and $f(x)$ be finite. If $d_-f(x;0) = -\infty$, then for any $\delta > 0$, there exists $y \in X$ such that*

$$f(x) - f(y) \geqslant \delta d(x,y) > 0.$$

Proof. Suppose that $d_-f(x;0) = -\infty$. Then for any $\delta > 0$, it follows from the definition of $d_-f(x;\cdot)$ that there exist a sequence $(t_n) \downarrow 0$ and a sequence $(u_n) \to 0$ such that

$$f(x) - f(x + t_n u_n) > t_n \delta = \frac{\delta}{\|u_n\|} d(x, x + t_n u_n).$$

Thus, noting $\frac{1}{\|u_n\|} > 1$ for some large n and letting $y = x + t_n u_n$, we complete the proof. $\square$

In terms of Hadamard's directional derivative, we have the following sufficient condition for f to admit an error bound.

THEOREM 2.4. *Let S be a closed subset of X. Let $\theta := \inf_{z \in X \setminus S} f(z)$ and $\delta > 0$. Suppose that $\theta > -\infty$ and*

$$\sup_{x \in \mathfrak{D} \setminus S} \inf_{\|u\|=1} d_-f(x;u) \leqslant -\delta. \tag{2.5}$$

Then

$$\delta d(x,S) \leqslant f(x) - \theta \quad \forall x \in X \setminus S.$$

Proof. Define a function $g : X \to R \cup \{+\infty\}$ by

$$g(x) := \begin{cases} f(x), & x \in X \setminus S, \\ \theta, & x \in S. \end{cases}$$

Then $g \geqslant \theta$ is a lower semicontinuous function with $d_-f(x;u) = d_-g(x;u)$ for any $x \in X\backslash S$ and $u \in X$. Thus, one has $\mathfrak{D}\backslash S = \mathfrak{D}_g\backslash S$, where $\mathfrak{D}_g := \{x \in X; d_-g(x;0) > -\infty\}$. Therefore, (2.5) can be rewritten as

$$\sup_{x\in\mathfrak{D}_g\backslash S} \inf_{\|u\|=1} d_-g(x;u) \leqslant -\delta. \tag{2.6}$$

Let $\delta' \in (0,\delta)$ and $x \in Domg\backslash S$. We claim that there exists $y \in X$ such that

$$g(x) - g(y) \geqslant \delta' d(x,y) \neq 0. \tag{2.7}$$

If $x \notin \mathfrak{D}_g$, then the claim is certainly true by Lemma 2.3. Thus we may suppose that $x \in \mathfrak{D}_g$. By (2.6) there exists a unit vector u such that $d_-g(x;u) < -\delta'$. Hence, there exist sequences $(t_n) \downarrow 0$ and $(u_n) \to u$ such that

$$\lim_{n\to+\infty} \frac{1}{t_n}(g(x+t_nu_n) - g(x)) < -\delta';$$

that is,

$$\lim_{n\to+\infty} \frac{1}{t_n\|u_n\|}(g(x+t_nu_n) - g(x)) < -\delta'.$$

Taking $y = x + t_nu_n$ with a sufficiently large n, we have $y \neq x$ and

$$g(y) - g(x) < -\delta' d(x,y).$$

Therefore the claim made in (2.7) holds. Consequently, $Domg\backslash S \subset X\backslash O_{\delta'}(g)$ which ensures (since $O_{\delta'}(g) \subset Domg$) that $O_{\delta'}(g) \subset S$. Then, we deduce from Proposition 2.1 that

$$g(x) - \theta \geqslant \delta' d(x,S) \quad \forall x \in X\backslash S.$$

Since δ' lying in $(0,\delta)$ is arbitrary, we have

$$g(x) - \theta \geqslant \delta d(x,S) \quad \forall x \in X\backslash S. \qquad \square$$

By considering level sets $L_f(\lambda)$ in place of S, we have the following theorem.

THEOREM 2.5. *Let $\delta > 0$ and let $\lambda \in R$ be such that $L_f(\lambda) \neq \emptyset$. Suppose that*

$$\sup_{x\in\mathfrak{D}\backslash L_f(\lambda)} \inf_{\|u\|=1} d_-f(x;u) \leqslant -\delta. \tag{2.8}$$

Then

$$\delta d(x, L_f(\lambda)) \leqslant f(x) - \lambda \quad \forall x \in X\backslash L_f(\lambda).$$

Proof. Let $\theta := \inf_{x\in X\backslash L_f(\lambda)} f(x)$. Then $\theta \geqslant \lambda$. Thus, the result follows from Theorem 2.4. $\square$

Remark 1. In place of (2.8), the sufficient condition established by Ng and Zheng in [13] is that

$$\sup_{x\in X\backslash L_f(\lambda)} \inf_{\|u\|=1} D_-f(x;u) \leqslant -\delta. \tag{2.9}$$

Since $d_-f \leqslant D_-f$ and $\mathfrak{D}\backslash L_f(\lambda) \subseteq X\backslash L_f(\lambda)$, it is clear that Theorem 2.5 improves the earlier result. In particular, the consideration of d_-f is of importance to us as it provides information about f along "curvilinear tangents" (rather than only linear ones) when we consider constrained problems. This in turn will provide a key argument for our error bound result of quadratic functions studied in section 4. In the appendix of this paper, we give an example which is covered by Theorem 2.5 but not by [13, Theorem 2.5].

We now turn to a result for a constrained problem; this result will provide a key step for us to establish a second sufficient condition in Theorem 2.7. Let C be a closed subset of X and let $x \in C$. Recall (see [3]) that the contingent cone of C at x is defined by

$$T_C(x) := \left\{ u \in X; \liminf_{t\downarrow 0} \frac{1}{t}(d_C(x+tu)) = 0 \right\}.$$

Thus, $u \in T_C(x)$ if and only if there exist sequences $(t_n) \downarrow 0$ and $(u_n) \to u$ such that $x + t_n u_n \in C$ for each n.

THEOREM 2.6. *Let S and C be closed subsets of X with $\emptyset \neq S \subseteq C$. Let $f : C \to R \cup \{+\infty\}$ be a proper lower semicontinuous function such that $\lambda := \inf_{x\in C\backslash S} f(x) > -\infty$, and let $\delta > 0$. Suppose that for any $x \in \mathfrak{D}_f\backslash S$ there exists a unit vector $u \in T_C(x)$ such that $d_-f(x;u) \leqslant -\delta$. Then one has $\delta d(x,S) \leqslant (f(x) - \lambda)$ for each $x \in C\backslash S$.*

Proof. f can be regarded as being defined on X by putting $f(x) = +\infty$ for all $x \in X\backslash C$. Let g be defined by

$$g(x) := \begin{cases} f(x), & x \in X\backslash S, \\ \lambda, & x \in S. \end{cases}$$

Thus, g is a lower semicontinuous function on X bounded from below. Since $f(x) = g(x) = +\infty$ for $x \in X\backslash C$,

$$\mathfrak{D}_g\backslash S = \mathfrak{D}_f\backslash S. \tag{2.10}$$

Let $x \in \mathfrak{D}_g\backslash S$. By (2.10) and assumption, take a unit vector $u \in T_C(x)$ such that $d_-f(x;u) \leqslant -\delta$. Take sequences $(t_n) \downarrow 0$, $(u_n) \to u$ such that

$$d_-f(x;u) = \lim_{n\to+\infty} \frac{1}{t_n}(f(x+t_nu_n) - f(x)) \leqslant -\delta.$$

By considering subsequences if necessary, we can further assume that

$$x + t_n u_n \in C \quad \forall n.$$

Indeed, if $x + t_n u_n \notin C$ for infinitely many n, then $f(x + t_n u_n) = +\infty$ for such n; this contradicts

$$\lim_{n\to\infty} \frac{f(x+t_nu_n) - f(x)}{t_n} < +\infty.$$

Noting that $x + t_n u_n \notin S$ so that $f(x + t_n u_n) = g(x + t_n u_n)$ for all large enough n, it follows that

$$\begin{aligned} d_-g(x;u) &\leqslant \liminf_{n\to+\infty} \tfrac{1}{t_n}(g(x+t_nu_n) - g(x)) \\ &= \liminf_{n\to+\infty} \tfrac{1}{t_n}(f(x+t_nu_n) - f(x)) \\ &= d_-f(x;u) \leqslant -\delta. \end{aligned}$$

Thus, one has

$$\sup_{x\in\mathfrak{D}_g\setminus S}\inf_{\|u\|=1} d_-g(x;u)\leqslant -\delta.$$

It follows from Theorem 2.4 that

$$\delta d(x,S)\leqslant g(x)-\lambda=f(x)-\lambda\quad\forall x\in C\setminus S.\qquad\square$$

Let $\varepsilon>0$ and

$$\mathfrak{D}(\varepsilon):=\left\{x\in\mathfrak{D}\setminus L_f(\lambda);\ \inf_{\|u\|=1} d_-f(x;u)\geqslant-\varepsilon\right\}. \tag{2.11}$$

In terms of $\mathfrak{D}(\varepsilon)$, the sufficient condition in Theorem 2.5 (ensuring that f has an error bound for $L_f(\lambda)$) is clearly equivalent to the following: for some $\varepsilon>0$,

$$\mathfrak{D}(\varepsilon)=\emptyset. \tag{2.12}$$

The following result improves Theorem 2.5 by relaxing condition (2.12).

THEOREM 2.7. *Let $f:X\to R\cup\{+\infty\}$ be a lower semicontinuous function, and let $\lambda\in R$ be such that $X\neq L_f(\lambda)\neq\emptyset$. Suppose that there exist $\varepsilon>0,\upsilon>\lambda$ and an open set Q with*

$$\mathfrak{D}(\varepsilon)\subseteq Q\subseteq X\setminus L_f(\upsilon) \tag{2.13}$$

and

$$d_0:=\sup_{x\in Q} d(x,L_f(\lambda))<+\infty. \tag{2.14}$$

Then f admits an error bound for $L_f(\lambda)$.

Proof. We may assume that Q is nonempty (see Remark 2 below). Then, if $x\in Q$, $d_0\geqslant d(x,L_f(\lambda))>0$. Let $Q(s):=\{x\in X\setminus L_f(\upsilon);d(x,L_f(\lambda))<d_0+s\},\forall s>0$, and denote Q by $Q(0)$. Then for any $s\geqslant 0, Q(s)$ is an open subset of X with $Q(s)\subseteq X\setminus L_f(\upsilon)$ and

$$\sup_{x\in Q(s)} d(x,L_f(\lambda))\leqslant d_0+s<+\infty.$$

Note that

$$f(x)-\lambda\geqslant\frac{\upsilon-\lambda}{d_0+s}d(x,L_f(\lambda))\quad\forall x\in Q(s) \tag{2.15}$$

because $f(x)>\upsilon$. Let $\delta=\min\{\frac{\varepsilon}{2},\frac{\upsilon-\lambda}{2d_0}\}$. We will show that

$$f(x)-\lambda\geqslant\delta d(x,L_f(\lambda))\quad\forall x\in X\setminus L_f(\lambda). \tag{2.16}$$

Since

$$X\setminus L_f(\lambda)=[L_f(\upsilon)\setminus L_f(\lambda)]\cup[\overline{Q}\setminus L_f(\lambda)]\cup[X\setminus(\overline{Q}\cup L_f(\upsilon))], \tag{2.17}$$

we need only verify (2.16) for x belonging to each of the three sets on the right-hand side of (2.17). Accordingly we divide our proof into three steps.

(I) We first show that

$$f(x) - \lambda \geqslant \delta d(x, L_f(\lambda)) \quad \forall x \in L_f(\upsilon)\backslash L_f(\lambda). \tag{2.18}$$

To do this, let $\bar{f}$ denote the "restriction" of f to $L_f(\upsilon)$; more precisely, let $\bar{f}$ be defined by $\bar{f}(x) = f(x)$ if $x \in L_f(\upsilon)$ or $\bar{f}(x) = +\infty$ otherwise. Since

$$f(q) \geqslant f(w) \quad \forall q \in X\backslash L_f(\upsilon) \text{ and } w \in L_f(\upsilon),$$

$d_-f(x;u) = d_-\bar{f}(x;u)$ for each $x \in L_f(\upsilon)$ and $u \in T_{L_f(\upsilon)}(x)$. Therefore $\mathfrak{D}_{\bar{f}} = \mathfrak{D} \cap L_f(\upsilon)$. Let $z \in \mathfrak{D} \cap L_f(\upsilon)\backslash L_f(\lambda)$. By (2.13), $z \notin \mathfrak{D}(\varepsilon)$ and so $\inf_{\|u\|=1} d_-f(z;u) < -\varepsilon$. Then there exist a unit vector u and sequences $(t_n) \downarrow 0$, $(u_n) \to u$ such that

$$\lim_{n\to+\infty} \frac{f(z + t_n u_n) - f(z)}{t_n} = d_-f(z;u) < -\varepsilon \leqslant -\delta.$$

Since $f(z) \leqslant \upsilon$, it follows that for all large n, $z + t_n u_n \in L_f(\upsilon)$ and so $u \in T_{L_f(\upsilon)}(z)$. By Theorem 2.6 (applied to $\overline{f}$, $L_f(\upsilon), L_f(\lambda)$ in place of f, C, S), we have (2.18).

(II) Next we show that

$$f(x) - \lambda \geqslant \delta d(x, L_f(\lambda)) \quad \forall x \in \overline{Q}\backslash L_f(\lambda). \tag{2.19}$$

In fact, let $x \in \overline{Q}\backslash L_f(\lambda)$; we may assume that $x \notin L_f(\upsilon)$ (because of (2.18)). Take a sequence $(x_n) \subseteq Q$ such that $x_n \to x$. Then $d(x_n, L_f(\lambda)) \to d(x, L_f(\lambda))$ and so $d(x, L_f(\lambda)) \leqslant d_0$. This implies that

$$f(x) - \lambda \geqslant \frac{\upsilon - \lambda}{d_0} d(x, L_f(\lambda)) \geqslant \delta d(x, L_f(\lambda)),$$

verifying (2.19).

(III) Let S denote the set $\overline{Q} \cup L_f(\upsilon)$, and define $\theta := \inf_{X\setminus S} f(x)$. Clearly, $\theta \geqslant \upsilon$. By (2.13), one has that $\mathfrak{D}(\varepsilon) \cap (\mathfrak{D}\backslash S) = \emptyset$; that is,

$$\inf_{\|u\|=1} d_-f(x;u) < -\varepsilon \leqslant -2\delta \quad \forall x \in \mathfrak{D}\backslash S.$$

By Theorem 2.4 one has

$$f(x) - \upsilon \geqslant f(x) - \theta \geqslant 2\delta d(x, S) \quad \forall x \in X\backslash S. \tag{2.20}$$

By definition of S, $X\backslash S$ can be expressed as $S_1 \cup S_2$, where S_1, S_2 are defined by

$$S_1 := \{x \in X\backslash S; d(x, S) = d(x, \overline{Q})\},$$

$$S_2 := \{x \in X\backslash S; d(x, S) = d(x, L_f(\upsilon))\}.$$

Let $x \in S_1$. If $x \in Q(d_0)$, then by (2.15) one has

$$f(x) - \lambda \geqslant \frac{\upsilon - \lambda}{2d_0} d(x, L_f(\lambda)) \geqslant \delta d(x, L_f(\lambda)). \tag{2.21}$$

If $x \notin Q(d_0)$, then one has $d(x, L_f(\lambda)) \geqslant 2d_0$ and

$$d(x, S) = d(x, \overline{Q}) \geqslant d(x, L_f(\lambda)) - d_0 \geqslant \frac{1}{2} d(x, L_f(\lambda)), \tag{2.22}$$

where the first (nonstrict) inequality holds as

$$d(x,q)+d_0 \geqslant d(x,q)+d(q,L_f(\lambda)) \geqslant d(x,L_f(\lambda)) \quad \forall q \in Q.$$

Combining (2.20) and (2.22), one has

$$f(x)-\lambda \geqslant f(x)-\upsilon \geqslant 2\delta d(x,S) \geqslant \delta d(x,L_f(\lambda)).$$

Combining this with (2.21), we arrive at

$$f(x)-\lambda \geqslant \delta d(x,L_f(\lambda)) \quad \forall x \in S_1. \tag{2.23}$$

Finally, we consider the case when $x \in S_2$. Take $y_n \in L_f(\upsilon)$ such that $d(x,y_n) < d(x,L_f(\upsilon))+\frac{1}{n}$; thus,

$$d(x,L_f(\lambda)) \leqslant d(x,L_f(\upsilon))+d(y_n,L_f(\lambda))+\frac{1}{n}. \tag{2.24}$$

We suppose without loss of generality that each $y_n \notin L_f(\lambda)$. Indeed, if $y_n \in L_f(\lambda)$ for infinitely many n, then passing to the limits in

$$d(x,L_f(\lambda)) \leq d(x,y_n) < d(x,L_f(\upsilon))+\frac{1}{n} \leq d(x,L_f(\lambda))+\frac{1}{n}$$

gives that $d(x,L_f(\lambda)) = d(x,L_f(\upsilon))$, and it follows from (2.20) and the fact that $x \in S_2$ that

$$f(x)-\lambda > f(x)-\upsilon \geqslant \delta d(x,S) = \delta d(x,L_f(\upsilon)) = \delta d(x,L_f(\lambda)).$$

It remains to consider the case when each $y_n \in L_f(\upsilon)\backslash L_f(\lambda)$. Noting that $d(x,S) = d(x,L_f(\upsilon))$ as $x \in S_2$, it follows from (2.18) and (2.20) that

$$\begin{aligned} f(x)-\lambda &\geqslant f(x)-\upsilon+f(y_n)-\lambda \\ &\geqslant \delta[d(x,L_f(\upsilon))+d(y_n,L_f(\lambda))]. \end{aligned}$$

Combining this with (2.24), we have

$$f(x)-\lambda \geqslant \delta d(x,L_f(\lambda)) \quad \forall x \in S_2. \tag{2.25}$$

Therefore (2.16) is seen to hold by (2.18), (2.19), (2.23), and (2.25). □

Remark 2. If there exists $\varepsilon > 0$ such that $\mathfrak{D}(\varepsilon) = \emptyset$, then there exist $\upsilon > \lambda$ and a nonempty open set Q satisfying (2.13) and (2.14). Indeed, pick $x_0 \in X\backslash L_f(\lambda)$ and take υ such that $f(\cdot) > \upsilon \geqslant \lambda$ at x_0 and hence on a bounded open neighborhood Q of x_0. This is possible because f is lower semicontinuous. Therefore Theorem 2.7 extends the result of Theorem 2.5.

Let $C_{L_f(\lambda)}(x)$ denote the cone defined by

$$C_{L_f(\lambda)}(x) := \{u \in X; \text{there exists } T > 0 \text{ such that } x+tu \in L_f(\lambda)\ \forall t \in (0,T]\}.$$

This may be referred to as the linear cone of $L_f(\lambda)$ at x, in general a rather smaller set than the contingent cone $T_{L_f(\lambda)}(x)$.

PROPOSITION 2.8. *Let X be a Banach space, let $\lambda \in R$, and let $f : X \to R$ be a C^1-function. Suppose there exists $x \in L_f(\lambda)$ such that $\nabla f(x) = 0$ and $C_{L_f(\lambda)}(x) \neq X$. Then f has no error bound for $L_f(\lambda)$.*

Proof. Let $x \in L_f(\lambda), \nabla f(x) = 0$, and $u \notin C_{L_f(\lambda)}(x)$. By definition there exists a sequence $(t_n) \downarrow 0$ such that

$$x + t_n u \notin L_f(\lambda) \ \forall n. \tag{2.26}$$

Take $z_n \in L_f(\lambda)$ such that

$$\|x + t_n u - z_n\| \leqslant \left(1 + \frac{1}{n}\right) d(x + t_n u, L_f(\lambda)) \ \forall n. \tag{2.27}$$

By virtue of the intermediate value theorem and by replacing z_n with a point in the line segment $(x+t_n u, z_n)$ if necessary, we can assume that $f(z_n) = \lambda$. Note that, since $x \in L_f(\lambda)$, the right-hand side of (2.27) converges to 0 as $n \to +\infty$. Thus, letting $t'_n = \|x + t_n u - z_n\|$, one has $t'_n \to 0$. Let $v_n = \frac{x+t_n u - z_n}{t'_n}$. Then $x + t_n u = z_n + t'_n v_n$. For each t'_n, by the mean value theorem there exists ξ_n lying in the line segment $(z_n + t'_n v_n, z_n)$ such that

$$\frac{f(z_n + t'_n v_n) - f(z_n)}{t'_n} = \nabla f(\xi_n) v_n \to 0 \text{ as } n \to +\infty \tag{2.28}$$

because $\xi_n \to x$ and $\nabla f(x) = 0$. It follows from (2.27) and (2.28) that

$$\begin{aligned} \lim_{n\to+\infty} \frac{f(x + t_n u) - \lambda}{d(x + t_n u, L_f(\lambda))} &= \lim_{n\to+\infty} \frac{t'_n}{d(x + t_n u, L_f(\lambda))} \cdot \frac{f(z_n + t'_n v_n) - f(z_n)}{t'_n} \\ &= \lim_{n\to+\infty} \nabla f(\xi_n) v_n = 0. \end{aligned}$$

This implies that f has no error bound for $L_f(\lambda)$. □

3. Second-order directional derivatives and second-order sufficient conditions. In this section we suppose that $f : X \to R$ is a Gateaux differentiable function with derivative denoted by ∇f. Recall that the second-order lower directional derivative of f at x in the directions u, v is defined by

$$d^2_- f(x; u, v) := \liminf_{t \downarrow 0} \frac{f(x + tu + t^2 v) - f(x) - t\nabla f(x) u}{t^2}. \tag{3.1}$$

This second-order directional derivative has been studied by many authors; see [1], [6], [7], [9], and [17]. Note in particular that when f is a C^2-function, then

$$d^2_- f(x; u, 0) = \frac{1}{2} \nabla^2 f(x)(u, u). \tag{3.2}$$

The following result may be regarded as a generalization of the Taylor expansion. For related results see [2] and [5].

THEOREM 3.1 (inequality form of the Taylor expansion). *Let $f : X \to R$ be lower semicontinuous and Gateaux differentiable. Let $x, u \in X$ and $T > 0$. Then there exists $\alpha \in [0, T)$ such that*

$$\frac{f(x + Tu) - f(x) - T\nabla f(x) u}{T^2} \leqslant d^2_- f(x + \alpha u; u, 0). \tag{3.3}$$

Proof. Define $\rho := f(x+Tu) - f(x) - T\nabla f(x)u$ and $h : [0,T] \to R$ by

$$h(t) := f(x+tu) - t\nabla f(x)(u) - \frac{t^2}{T^2}\rho. \tag{3.4}$$

Then h is lower semicontinuous on $[0,T]$ and $h(T) = h(0)$. Thus, h attains a minimum at some point $\alpha \in [0,T)$ and hence

$$\nabla f(x+\alpha u)u - \nabla f(x) - \frac{2\alpha}{T^2}\rho = 0 \tag{3.5}$$

((3.5) holds trivially if $\alpha = 0$). Therefore,

$$\begin{aligned}
&\frac{h(\alpha+\lambda) - h(\alpha)}{\lambda^2} \\
&\quad = \frac{f(x+\alpha u+\lambda u) - f(x+\alpha u) - \lambda\nabla f(x+\alpha u)u}{\lambda^2} - \frac{\rho}{T^2} + \frac{\nabla f(x+\alpha u)u - \nabla f(x)u - \frac{2\alpha}{T^2}\rho}{\lambda} \\
&\quad = \frac{f(x+\alpha u+\lambda u) - f(x+\alpha u) - \lambda\nabla f(x+\alpha u)u}{\lambda^2} - \frac{\rho}{T^2}.
\end{aligned}$$

By the minimality of α, it follows that

$$0 \leqslant \liminf_{\lambda \downarrow 0} \frac{f(x+\alpha u+\lambda u) - f(x+\alpha u) - \lambda\nabla f(x+\alpha u)u}{\lambda^2} - \frac{\rho}{T^2};$$

that is,

$$\frac{\rho}{T^2} \leqslant d_-^2 f(x+\alpha u; u, 0),$$

thus proving (3.3). □

We are now ready to present our second-order sufficient condition for f to have an error bound. Recall that $\mathfrak{D}(\varepsilon)$ is defined by

$$\mathfrak{D}(\varepsilon) := \{x \in X \backslash L_f(\lambda); \|\nabla f(x)\| \leqslant \varepsilon\} \qquad (\varepsilon > 0). \tag{3.6}$$

THEOREM 3.2. *Let $f : X \to R$ be continuous and Gateaux differentiable. Suppose that $L_f(\lambda) \neq \emptyset$ and that there exist $\upsilon > \lambda$ and $T, \delta, \rho > 0$ such that the following conditions hold:*

(a) $\mathfrak{D}(\rho) \subseteq X \backslash L_f(\upsilon)$.

(b) *For each $x \in \mathfrak{D}(\rho)$, there exists a unit vector u such that*

$$d_-^2 f(x+\alpha u; u, 0) < -\delta \quad \forall \alpha \in [0,T).$$

Then f has an error bound for $L_f(\lambda)$.

Proof. Let $\varepsilon = \min\{\rho, \frac{\delta}{4}, \frac{T\delta}{2}\}$ and $\beta = \min\{\upsilon - \lambda, \frac{\varepsilon}{2}, 1, \frac{T\delta}{2}\}$. We claim that for each $x \in Dom f \backslash L_f(\lambda)$ there exists $y \in f^{-1}[\lambda, +\infty)$ such that

$$f(x) - f(y) \geqslant \beta d(x,y) > 0.$$

Granting this, one can complete the proof by virtue of Corollary 2.2. To establish our claim, we let $x \in Dom f \backslash L_f(\lambda)$ and we consider the following three cases.

Case 1. $x \in \mathfrak{D}(\varepsilon)$ and $d(x, L_f(\lambda)) < 1$. Since f is continuous, we can take $y \in L_f(\lambda)$ such that $f(y) = \lambda$ and

$$\|x - y\| < 1.$$

Since $x \in \mathfrak{D}(\varepsilon) \subseteq X \backslash L_f(\upsilon)$, one has $f(x) > \upsilon$ and hence

$$f(x) - f(y) \geqslant \upsilon - \lambda \geqslant (\upsilon - \lambda) d(x, y) \geqslant \beta d(x, y) > 0. \tag{3.7}$$

Case 2. $x \in \mathfrak{D}(\varepsilon)$ and $d(x, L_f(\lambda)) \geqslant 1$. By assumption let u be a unit vector such that

$$d_-^2 f(x + \alpha u; u, 0) < -\delta \quad \forall \alpha \in [0, T).$$

Let $T' = \min\{\frac{1}{2}, T\}$; then $\varepsilon \leqslant \frac{T'\delta}{2}$. Note that $x + T'u \notin L_f(\lambda)$ as $d(x, L_f(\lambda)) \geqslant 1$ and that by Theorem 3.1 there exists $\alpha \in [0, T')$ such that

$$f(x + T'u) - f(x) - T' \nabla f(x) u \leqslant T'^2 d_-^2 f(x + \alpha u; u, 0) < -T'^2 \delta.$$

Note that $\nabla f(x) u \leqslant \varepsilon$ since $x \in \mathfrak{D}(\varepsilon)$; hence

$$f(x + T'u) - f(x) - T'\varepsilon \leqslant -T'^2 \delta.$$

It follows that

$$\begin{aligned} f(x) - f(x + T'u) &\geqslant T'(T'\delta - \varepsilon) \\ &\geqslant \tfrac{T'^2\delta}{2} = \tfrac{T'\delta}{2} d(x, x + T'u) \\ &\geqslant \beta d(x, x + T'u) > 0. \end{aligned} \tag{3.8}$$

Case 3. $x \in (X \backslash \mathfrak{D}(\varepsilon)) \backslash L_f(\lambda)$. Then $\|\nabla f(x)\| > \varepsilon$ and hence there exists a unit vector u such that $\nabla f(x) u < -\varepsilon$. Thus, there exists $(t_n) \downarrow 0$ such that

$$f(x) - f(x + t_n u) > t_n \varepsilon = \varepsilon d(x, x + t_n u) \geqslant \beta d(x, x + t_n u) \ \forall n. \tag{3.9}$$

Since $f(x) > \lambda$, letting $y = x + t_n u$ with n sufficiently large one has $y \notin L_f(\lambda)$. ∎

COROLLARY 3.3. *Let $f : X \to R$ be a C^2-function. Let $\lambda \in R$ be such that $L_f(\lambda) \neq \emptyset$.*

Let $\rho, \delta > 0$. We suppose that $\nabla^2 f$ is uniformly continuous on the δ-neighborhood U of $\mathfrak{D}(\rho)$. Then f has an error bound for $L_f(\lambda)$ if either of the following conditions is satisfied:

(a) $\mathfrak{D}(\rho) = \emptyset$.

(b) *There exists $\upsilon > \lambda$ with $\mathfrak{D}(\rho) \subseteq X \backslash L_f(\upsilon)$ such that*

$$\sup_{x \in \mathfrak{D}(\rho)} \inf_{\|u\|=1} \nabla^2 f(x)(u, u) < 0.$$

Proof. We need to deal only with the case (b). Let $\varepsilon > 0$ such that

$$\sup_{x \in \mathfrak{D}(\rho)} \inf_{\|u\|=1} \nabla^2 f(x)(u, u) < -3\varepsilon. \tag{3.10}$$

Take $T \in (0, \delta)$ such that

$$\|\nabla^2 f(x) - \nabla^2 f(Y)\| < \delta \tag{3.11}$$

for all $x, y \in U$ with $\|x - y\| < T$. Let $x \in \mathfrak{D}(\rho)$; by (3.10) there exists a unit vector u such that

$$\nabla^2 f(x)(u, u) < -3\varepsilon,$$

and so by (3.11)

$$2d_-^2 f(y, u, 0) = \nabla^2 f(y)(u, u) < -2\varepsilon$$

whenever $\|y - x\| < T$. Therefore the result follows from Theorem 3.2. ∎

In the next section, we shall apply the results obtained in sections 2 and 3 to describe exactly when a quadratic function on X has a Lipschitz error bound.

4. Error bound for a quadratic function on Banach space. Throughout this section, f will denote a quadratic function

$$f(x) := \frac{1}{2}\langle x, Ax\rangle + \langle x, z^*\rangle + \gamma \quad \forall x \in X, \tag{4.1}$$

where X is a Banach space with the dual space X^*, $\gamma \in R, z^* \in X^*$, and $A : X \to X^*$ is a bounded linear operator from X into X^*. Let A' denote the dual operator of A defined by

$$\langle z, A'x\rangle = \langle x, Az\rangle$$

for all $x, z \in X$. Let $\frac{1}{2}(A + A')$ be denoted by $\bar{A}$. Then

$$\begin{aligned} \nabla f(x) &= \frac{1}{2}(\langle \cdot, Ax\rangle + \langle x, A\cdot\rangle) + \langle \cdot, z^*\rangle \\ &= \frac{1}{2}(A + A')x + z^* = \bar{A}x + z^* \end{aligned} \tag{4.2}$$

and

$$\nabla^2 f(x)(u, u) = \langle u, \overline{A}u\rangle \quad \forall x, u \in X.$$

We note that $\overline{A}$ is symmetric: $\langle z, \overline{A}x\rangle = \langle x, \overline{A}z\rangle$ for all $x, z \in X$. Since $\langle x, Ax\rangle = \langle x, \overline{A}x\rangle = \langle x, A'x\rangle$, we can suppose henceforth that A is symmetric (replace A by $\overline{A}$ if necessary). Further, we can show that

$$f(x + u) - f(x) = \nabla f(x)u + \frac{1}{2}\nabla^2 f(x)(u, u) \quad \forall x, u \in X. \tag{4.3}$$

Denote the set of all critical (stationary) points of f by $N_{\nabla f}$, that is,

$$N_{\nabla f} := \{x \in X; \nabla f(x) = 0\}.$$

By (4.2) one has $N_{\nabla f} = \{x \in X; Ax = -z^*\}$. We recall that (x_n) is a critical (or stationary) sequence of f if $\lim_{n\to+\infty} \nabla f(x_n) = 0$. η is called a critical value of f if $\eta = f(x)$ for some $x \in N_{\nabla f}$. We will show in the following lemma that there is only one critical value (provided that $N_{\nabla f}$ is nonempty).

LEMMA 4.1. *Suppose that $N_{\nabla f}$ is nonempty. Then f and z^* are constant on $N_{\nabla f}$. Moreover, if $z_0 \in N_{\nabla f}$, then it holds $\forall x \in N_{\nabla f}$ that*

$$z^*(x) = -\langle z_0, Az_0\rangle \tag{4.4}$$

and that

$$f(x) = \gamma - \frac{1}{2}\langle z_0, Az_0\rangle \quad \forall x \in N_{\nabla f}. \tag{4.5}$$

If we assume in addition that A is of closed range, then there exists $m > 0$ such that, whenever (x_n) is a critical sequence of f, one has

$$\limsup_{n\to+\infty} d(x_n, N_{\nabla f} - z_0) \leqslant \frac{\|z^*\|}{m} \tag{4.6}$$

and

$$\lim_{n\to+\infty} f(x_n) = \gamma - \frac{1}{2}\langle z_0, Az_0\rangle. \tag{4.7}$$

($\gamma - \frac{1}{2}\langle z_0, Az_0\rangle$ will be referred to as the critical value of f.)

Proof. Since A is assumed symmetric, (4.2) reads

$$\nabla f(x) = Ax + z^*. \tag{4.8}$$

Then $Az_0 + z^* = 0$ and $Ax + z^* = 0$; hence $Ax = Az_0$ for each $x \in N_{\nabla f}$. Therefore, by symmetry of A,

$$z^*(x) = \langle x, -Az_0\rangle = \langle z_0, -Ax\rangle = \langle z_0, -Az_0\rangle \quad \forall x \in N_{\nabla f},$$

verifying (4.4). This also implies that

$$\langle x, Ax\rangle = \langle z_0, Az_0\rangle \quad \forall x \in N_{\nabla f}. \tag{4.9}$$

Consequently,

$$\begin{aligned} f(x) &= \frac{1}{2}\langle x, Ax\rangle + z^*(x) + \gamma \\ &= \gamma - \frac{1}{2}\langle z_0, Az_0\rangle \quad \forall x \in N_{\nabla f}, \end{aligned}$$

verifying (4.5).

Suppose in addition that AX is closed. Then there exists $m > 0$ such that

$$\|Ax\| \geqslant m d(x, \ker A) \quad \forall x \in X$$

(cf. [18, IV. 5, Theorem 5.9]). Since (4.8) reads $\nabla f(x) = A(x - z_0)$, it follows that

$$\|\nabla f(x_n) - z^*\| \geqslant m \cdot d(x_n, N_{\nabla f} - z_0).$$

Hence, if $\lim_{n\to+\infty} \nabla f(x_n) = 0$, one has

$$\|z^*\| \geqslant m \limsup_{n\to+\infty} d(x_n, N_{\nabla f} - z_0),$$

thus proving (4.6).

Now, for each large n, one applies (4.6) to obtain $y_n \in N_{\nabla f}$ such that

$$\|x_n - (y_n - z_0)\| \leqslant \frac{\|z^*\|}{m} + 1.$$

Note that by (4.5), each $f(y_n) = \eta$, where $\eta := \gamma - \langle z_0, Az_0\rangle$. On the other hand, by the mean value theorem, there exists ξ_n in the line segment (x_n, y_n) such that

$$\begin{aligned} |f(x_n) - \eta| &= |f(x_n) - f(y_n)| \\ &= |\nabla f(\xi_n)(x_n - y_n)| \\ &\leqslant \|\nabla f(\xi_n)\|\,\|x_n - y_n\| \\ &\leqslant \|\nabla f(\xi_n)\| \left(\frac{\|z^*\|}{m} + \|z_0\| + 1\right) \to 0 \end{aligned}$$

because $\nabla f(\xi_n) \to 0$ as ∇f is affine, $\nabla f(x_n) \to 0$, and $\nabla f(y_n) = 0$. Therefore (4.7) holds. □

Let $\lambda \in R$. In the following, we consider the problem of whether f has an error bound for $L_f(\lambda)$ or not. To avoid trivialities we suppose that $X \neq L_f(\lambda) \neq \emptyset$.

PROPOSITION 4.2. *Suppose that $z^* \notin \overline{AX}$, the closure of the range AX of A. Then f has an error bound for $L_f(\lambda)$.*

Proof. By the Hahn–Banach theorem, there exists $u^{**} \in X^{**}$ of norm 1 such that

$$u^{**}(z^*) = \|z^*\| \text{ and } u^{**}(Ax) = 0 \quad \forall x \in X.$$

Since $\nabla f(x) = Ax + z^*$, it follows that

$$u^{**}(\nabla f(x)) = \|z^*\| \quad \forall x \in X.$$

By the bipolar theorem, there exists a net (u_κ) in the unit ball of X which converges to u^{**} in the $\sigma(X^{**}, X^*)$-topology. Then, for any $x \in X$ there exists u_κ such that

$$\nabla f(x)\left(\frac{u_\kappa}{\|u_\kappa\|}\right) > \frac{\|z^*\|}{2\,\|u_\kappa\|} \geqslant \frac{\|z^*\|}{2}.$$

Therefore (2.8) is satisfied (with $\delta = \frac{\|z^*\|}{2}$) and so Theorem 2.5 implies that f admits an error bound for $L_f(\lambda)$ whenever $L_f(\lambda) \neq \emptyset$. □

PROPOSITION 4.3. *Let $\lambda \in R$. Suppose that A is not positive semidefinite and is of closed range. If $N_{\nabla f} \cap L_f(\lambda) = \emptyset$, then f has an error bound for $L_f(\lambda)$.*

Proof. Let us first consider the case that $N_{\nabla f} = \emptyset$. Since $\nabla f(x) = Ax + z^*$, it follows that $z^* \notin AX$. Since A is assumed to be of closed range, it follows from Proposition 4.2 that f has an error bound for $L_f(\lambda)$. In the following we assume that $N_{\nabla f} \neq \emptyset$. We claim that there exist $\upsilon > \lambda$ and $\varepsilon > 0$ such that $\mathfrak{D}(\varepsilon) \subseteq X \backslash L_f(\upsilon)$. Indeed, if not, then there exists a sequence (x_n) such that $\|\nabla f(x_n)\| \leqslant \frac{1}{n}$ and $f(x_n) \leqslant \lambda + \frac{1}{n}$ for each n. Thus, letting η denote the critical value of f, it follows from Lemma 4.1 that $\eta \leqslant \lambda$, contradicting the assumption that $N_{\nabla f} \cap L_f(\lambda) = \emptyset$. Therefore our claim stands. Moreover, since A is not positive semidefinite, there exists a unit vector u_0 such that $\langle u_0, Au_0\rangle < 0$. Since $\nabla^2 f(z) = A$ for each $z \in X$, Corollary (3.3) implies that f has an error bound for $L_f(\lambda)$. □

For the case when A is of closed range, the next theorem together with Proposition 4.2 provides a complete answer for the question, When does f defined by (4.1) have an error bound?

THEOREM 4.4. *Let $\lambda \in R$. Suppose further that $z^* \in AX$ (namely, $N_{\nabla f} \neq \emptyset$). Then the following assertions hold.*

(i) *If λ is the critical value of f, then f has no error bound for $L_f(\lambda)$.*
(ii) *If λ is not the critical value of f, then f has an error bound for $L_f(\lambda)$ provided that A is of closed range.*

Proof. (i) We assume that $z^* \in AX$ and that $\lambda = f(x)$ for some (and hence for all by Lemma 4.1) x in $N_{\nabla f}$. By virtue of Proposition 2.8 it is sufficient to show $C_{L_f(\lambda)}(x) \neq X$. By the way of contradiction we assume that $C_{L_f(\lambda)}(x) = X$. Then for any $u \in X$, there exists $T_u > 0$ such that $x + tu \in L_f(\lambda)\ \forall t \in [0, T_u]$. Thus, for each $t \in [0, T_u]$, it follows from (4.3) that

$$\begin{aligned} \lambda \geqslant f(x+tu) &= f(x) + \nabla f(x)tu + \frac{t^2}{2}\nabla^2 f(x)(u,u) \\ &= \lambda + \frac{t^2}{2}\langle u, Au\rangle. \end{aligned}$$

This implies that

$$\langle u, Au\rangle \leqslant 0 \quad \forall u \in X.$$

Thus, f is a concave function and hence, any $x \in N_{\nabla f}$ is a maximum point of f. Consequently, $L_f(\lambda) = X$, the case that we have rejected at the outset.

(ii) By Lemma 4.1, there exists $\eta \in R$ such that $f(x) = \eta$ for each $x \in N_{\nabla f}$. By assumption, $\lambda \neq \eta$. Thus, we have two cases to consider: (a) $\lambda < \eta$ and (b) $\lambda > \eta$. For case (a), we can assume that A is not positive semidefinite (if A is positive semidefinite, then f is a convex function and so $\eta = \inf f$. Since $\eta > \lambda$, this is not possible because we have assumed that $L_f(\lambda) \neq \emptyset$). Thus, since S is of closed range it follows from Proposition 4.3 that f has an error bound for $L_f(\lambda)$.

Next consider case (b): $\lambda > \eta$. Then

$$\inf_{x \notin L_f(\lambda)} \|\nabla f(x)\| > 0 \tag{4.10}$$

(and so (2.8) is satisfied for some $\delta > 0$ and hence it follows from Theorem 2.5 that f has an error bound for $L_f(\lambda)$). To verify (4.10), we suppose on the contrary that there exists a sequence $(x_n) \subseteq X \backslash L_f(\lambda)$ such that $\nabla f(x_n) \to 0$. Then, by Lemma 4.1 it follows that $\lim_{n \to +\infty} f(x_n) = \eta$. But this is not possible because η is strictly smaller than λ and $\lambda < f(x_n)$ for all n. □

In the special case when $X = R^n$, the consideration of an error bound for a general quadratic function f is equivalent to that for ϕ of the form

$$\phi(x) := \frac{1}{2}\sum_{i=1}^{k} x_i^2 - \frac{1}{2}\sum_{i=k+1}^{m} x_i^2 + \sum_{i=m+1}^{n} c_i x_i + \gamma,$$

where $0 \leqslant k \leqslant m \leqslant n$ (cf. [13]). Note that

$$\phi(x) = \frac{1}{2}\langle x, Hx\rangle + \langle c, x\rangle + \gamma, \tag{4.11}$$

where $H = \begin{bmatrix} I_k & 0 & 0 \\ 0 & -I_{m-k} & 0 \\ 0 & 0 & 0 \end{bmatrix}$, I_k and I_m are, respectively, the $k \times k$ unit matrix and the $(m-k) \times (m-k)$ unit matrix, $c = (0, \ldots, 0, c_{m+1}, \ldots, c_n)$, and $x \in R^n$. The (operator associated with) matrix H is certainly of closed range. To compare our results with those in [13], let us fix $\lambda = 0$, and write S_ϕ for $L_\phi(0)$. We assume that $\emptyset \neq S_\phi \neq R^n$; that is, we reject the following trivial cases:

(a) $k = m, c = 0$, and $\gamma > 0$.

(b) $0 = k < m, c = 0$, and $\gamma \leqslant 0$.

Note also that $c = 0$ if and only if c belongs to the range of H.

COROLLARY 4.5. *Let ϕ be of the form* (4.11).

(I) *Suppose that $c = 0$. Then ϕ has an error bound for S_ϕ if and only if $\gamma \neq 0$.*

(II) *If $c \neq 0$, then ϕ has an error bound for S_ϕ.*

Proof. (I) Since $\nabla\phi(x) = Hx + c = Hx$, we see that $0 \in N_{\nabla\phi}$ and $\gamma = \phi(0)$ is the critical value of ϕ. It follows from Theorem 4.4 that $\gamma \neq \lambda$ if and only if ϕ has an error bound.

(II) This follows from Proposition 4.2. □

5. Appendix. We end the paper with an example showing that Theorem 2.5 improves a result in [13].

Example. Let $X = \ell^2$. Let $e_n \in \ell^2$ denote the nth unit vector $e_n := (0, \ldots, 0, 1, 0, \ldots 0)$, where 1 appears at the nth coordinate. Let L_n be the line segment defined by

$$L_n := \{x \in \ell^2; x = te_n, n^{-1} \leqslant t \leqslant 1\},$$

and let $A := \cup_{n=2}^{+\infty} L_n$. Then $A \cup \{0\}$ is closed and for any unit vector v, there exists $\varepsilon > 0$ such that

$$tB(v,\varepsilon) \cap A = \emptyset \quad \forall\, 0 \leqslant t < \varepsilon. \tag{5.1}$$

In fact, the claim is clearly true if $v = e_n$ for some $n \geqslant 2$. Moreover, if the claim is not true for some unit vector $v \neq e_n, n = 2, 3, \dots$, then there exist sequences $(v_n) \subseteq \ell^2$, $(\varepsilon_n) \downarrow 0$, and $(t_n) \downarrow 0$ such that each

$$v_n \in t_n B(v,\varepsilon_n) \cap A. \tag{5.2}$$

Since $v_n \in L_{m_n}$ for some $m_n \geqslant 2$, one can write $\tau_n e_{m_n}$ for v_n with $m_n^{-1} \leqslant \tau_n \leqslant 1$. It follows from (5.2) that

$$\frac{\tau_n}{t_n} e_{m_n} = \frac{v_n}{t_n} \to v,$$

which implies that $\frac{\tau_n}{t_n} \to 1$ and that $(e_{m_n}) \to v$. By definition of e_n's, this is impossible. Define a function $f : X \to R \cup \{+\infty\}$ by

$$f(x) := \begin{cases} -(1-n^{-\frac{1}{3}})\left[\dfrac{t-n^{-1}}{1-n^{-1}} - 1\right], & x \in L_n, x = te_n, n = 2, 3, \dots, \\ 1 + \|x\|, & x \notin L_n, \|x\| < 1, n = 2, 3, \dots, \\ 0, & \|x\| \geqslant 1. \end{cases}$$

Then f is lower semicontinuous and

$$0 \notin L_f(0) = \{x \in \ell^2; \|x\| \geqslant 1\} \cup \{e_n; n = 2, 3, \dots\}.$$

Let $u_n = n^{-\frac{1}{3}} e_n$. Then $n^{-\frac{2}{3}} u_n = \frac{1}{n} e_n \in L_n$ and so one has

$$\begin{aligned} d_- f(0;0) &\leqslant \lim_{n\to+\infty} \frac{1}{n^{-\frac{2}{3}}} (f(n^{-\frac{2}{3}} u_n) - f(0)) \\ &= \lim_{n\to+\infty} \frac{-n^{-\frac{1}{3}}}{n^{-\frac{2}{3}}} = -\infty, \end{aligned}$$

thus showing that $0 \notin \mathfrak{D}$. Moreover, by (5.1), it is easy to see that

$$d_- f(0;v) \geqslant 0 \quad \forall v \neq 0.$$

Thus, one has

$$\inf_{\|v\|=1} d_- f(0;v) \geqslant 0.$$

Since $0 \notin L_f(0)$, this implies that the sufficient condition of [13, Theorem 2.5] is not satisfied. On the other hand, note that for any x with $x \notin A \cup \{0\}$ and $\|x\| < 1$,

$$\nabla f(x) = \frac{x}{\|x\|}.$$

Then for $u = -\frac{x}{\|x\|}$,

$$d_- f(x;u) = \nabla f(x) u = -1. \tag{5.3}$$

If $x \in A : x = te_n, \frac{1}{n} \leqslant t < 1$ for some $n \geqslant 2$, then

$$d_- f(x; e_n) \leqslant -\frac{1 - n^{-\frac{1}{3}}}{1 - n^{-1}} \leqslant -(1 - 2^{-\frac{1}{3}}). \tag{5.4}$$

Combining (5.3) and (5.4) and noting that $0 \notin \mathfrak{D}$, we have

$$\sup_{x \in \mathfrak{D} \setminus L_f(0)} \inf_{\|v\|=1} d_- f(x; v) \leqslant -(1 - 2^{-\frac{1}{3}}) < 0.$$

Thus, (2.8) is satisfied and Theorem 2.5 is applicable.

REFERENCES

[1] A. Ben-Tal, *Second-order and related extremality conditions in nonlinear programming*, J. Optim. Theory Appl., 3 (1980), pp. 143–165.

[2] W. L. Chan, L. R. Huang, and K. F. Ng, *On generalized second-order derivatives and Taylor expansions in nonsmooth optimization*, SIAM J. Control Optim., 32 (1994), pp. 591–611.

[3] F. H. Clarke, *Optimization and Nonsmooth Analysis*, Wiley, New York, 1983.

[4] R. Cominetti, *Metric Regularity, tangent sets, and second-order optimality conditions*, Appl. Math. Optim., 21 (1990), pp. 265–287.

[5] R. Cominetti and R. Correa, *A generalized second-order derivative in nonsmooth optimization*, SIAM J. Control Optim., 28 (1990), pp. 789–809.

[6] R. Cominetti and J.-P. Penot, *Tangent sets of order one and two to the positive cones of some functional space*, Appl. Math. Optim., 36 (1997), pp. 291–312.

[7] R. Cominetti, *On pseudo-differentiability*, Trans. Amer. Math. Soc., 324 (1991), pp. 843–865.

[8] A. Hamel, *Remarks to equivalent formulation of Ekeland's variational principle*, Optimization, 31 (1994), pp. 233–238.

[9] L. R. Huang and K. F. Ng, *On some relations between Chaney's generalized second-order directional derivative and that of Ben-Tal and Zowe*, SIAM J. Control Optim., 34 (1996), pp. 1220–1234.

[10] L. R. Huang, K. F. Ng, and J.-P. Penot, *On minimizing and critical sequences in nonsmooth optimization*, SIAM J. Optim., 10 (2000), pp. 999–1019.

[11] A. Lewis and J. S. Pang, *Error Bounds for Convex Inequality Systems*, in Generalized Convexity, Generalized Monotonicity: Recent Results, Proceedings of the Fifth Symposium on Generalized Convexity (Luminy, France, 1996), J.-P. Crouzeix, J.-E. Martinez-Legaz, and M. Volle, eds., Kluwer Academic Publishers, Dordrecht, The Netherlands, 1998, pp. 75–110.

[12] Z. Q. Luo and J. F. Sturm, *Error bounds for quadratic systems*, in High Performance Optimization, Appl. Optim. 33, H. Frenk, K. Roos, T. Terlaky, and S. Zhang, eds., Kluwer Academic Publishers, Dordrecht, The Netherlands, 2000, pp. 383–404.

[13] K. F. Ng and X. Y. Zheng, *Error bounds for lower semicontinuous functions in normed spaces*, SIAM J. Optim., 12 (2001), pp. 1–17.

[14] K. F. Ng and X. Y. Zheng, *Global error bounds with fractional exponents*, Math. Program., 88 (2000), pp. 357–370.

[15] J.-P. Penot, *Conditioning convex and nonconvex problems*, J. Optim. Theory Appl., 90 (1997), pp. 209–221.

[16] J. S. Pang, *Error bounds in mathematical programming*, Math. Programming, 79 (1997), pp. 299–332.

[17] R. T. Rockafellar, *First and second-order epi-differentiability in nonlinear programming*, Trans. Amer. Math. Soc., 307 (1988), pp. 75–107.

[18] A. E. Taylor and D. C. Lay, *Introduction to Function Analysis*, John Wiley, New York, 1980.

SIAM J. OPTIM.
Vol. 14, No. 4, pp. 1074–1105

A FEASIBLE TRUST-REGION SEQUENTIAL QUADRATIC PROGRAMMING ALGORITHM*

STEPHEN J. WRIGHT† AND MATTHEW J. TENNY‡

Abstract. An algorithm for smooth nonlinear constrained optimization problems is described, in which a sequence of feasible iterates is generated by solving a trust-region sequential quadratic programming (SQP) subproblem at each iteration and by perturbing the resulting step to retain feasibility of each iterate. By retaining feasibility, the algorithm avoids several complications of other trust-region SQP approaches: the objective function can be used as a merit function, and the SQP subproblems are feasible for all choices of the trust-region radius. Global convergence properties are analyzed under various assumptions on the approximate Hessian. Under additional assumptions, superlinear convergence to points satisfying second-order sufficient conditions is proved.

Key words. nonlinear constrained optimization, feasible algorithm, sequential quadratic programming, trust-region algorithms

AMS subject classifications. 90C30, 65K05

DOI. 10.1137/S1052623402413227

1. Introduction.

We consider the general smooth constrained optimization problem,

$$\min f(z) \text{ subject to } c(z) = 0, \;\; d(z) \le 0, \tag{1.1}$$

where $z \in \mathsf{R}^n$, $f : \mathsf{R}^n \to \mathsf{R}$, $c : \mathsf{R}^n \to \mathsf{R}^m$, and $d : \mathsf{R}^n \to \mathsf{R}^r$ are smooth (twice continuously differentiable) functions. We denote the set of feasible points for (1.1) by $\mathcal{F}$.

At a feasible point z, let H be an $n \times n$ symmetric matrix. The basic sequential quadratic programming (SQP) approach obtains a step Δz by solving the subproblem

$$\min_{\Delta z} m(\Delta z) \stackrel{\text{def}}{=} \nabla f(z)^T \Delta z + \tfrac{1}{2} \Delta z^T H \Delta z \text{ subject to} \tag{1.2a}$$

$$c(z) + \nabla c(z)^T \Delta z = 0, \;\; d(z) + \nabla d(z)^T \Delta z \le 0. \tag{1.2b}$$

The matrix H is chosen as some approximation to the Hessian of the Lagrangian, possibly obtained by a quasi-Newton technique, or possibly a "partial Hessian" computed in some application-dependent way from some of the objective and constraint functions and Lagrange multiplier estimates. The function $m(\cdot)$ is the quadratic model for the change in f around the current point z.

Although the basic approach of (1.2a,b) often works well in the vicinity of a solution to (1.1), trust-region or line-search devices must be added to improve its robustness and global convergence behavior. In this paper, we consider a trust region of the form

$$\|D \Delta z\|_p \le \Delta, \tag{1.3}$$

*Received by the editors August 19, 2002; accepted for publication (in revised form) September 29, 2003; published electronically May 25, 2004.
http://www.siam.org/journals/siopt/14-4/41322.html

†Computer Sciences Department, University of Wisconsin, 1210 W. Dayton Street, Madison, WI 53706 (swright@cs.wisc.edu). The research of this author was supported in part by NSF grants MTS-0086559, ACI-0196485, and EIA-0127857.

‡Chemical Engineering Department, University of Wisconsin, Madison, WI 53706 (tenny@bevo.che.wisc.edu). The research of this author was supported in part by NSF grant CTS-0105360.

where the scaling matrix D is uniformly bounded above and $p \in [1, \infty]$. The choice $p = \infty$ makes (1.2a,b), (1.3) a quadratic program since we can then restate the trust-region constraint as $-\Delta e \le D\Delta z \le \Delta e$, where $e = (1, 1, \ldots, 1)^T$. The choice $p = 2$ produces the quadratic constraint $\Delta z^T D^T D \Delta z \le \Delta^2$, and since z is feasible for (1.1), we can show that the solution Δz of (1.2a,b), (1.3) is identical to the solution of (1.2a,b) alone, with H replaced by $H + \gamma D^T D$ for some $\gamma \ge 0$. For generality, we develop most of the convergence theory to apply to any choice of $p \in [1, \infty]$, making frequent use of the equivalence between $\|\cdot\|_p$ and $\|\cdot\|_2$.

By allowing D to have zero eigenvalues, the constraint (1.3) generally allows Δz to be unrestricted by the trust region in certain directions. We assume, however, that the combination of (1.3) and (1.2b) ensures that the all components of the step are controlled by the trust region; see Assumption 1 below.

When the iterate z is not feasible for the original problem (1.1) we cannot, in general, simply add the restriction (1.3) to the constraints in the subproblem (1.2a,b), since the resulting subproblem will be infeasible for small Δ. Practical trust-region methods, such as those due to Celis, Dennis, and Tapia [3] and Omojokun [12] do not insist on satisfaction of the constraints (1.2b) by the step Δz but, rather, achieve some reduction in the infeasibility while staying within the trust region (1.3) and reducing the objective in subproblem (1.2a).

Another issue that arises in the practical SQP methods is the use of a merit or penalty function to measure the worth of each point z. Typically, this function is some combination of the objective $f(z)$ and the violations of the constraints, that is, $|c_i(z)|$, $i = 1, 2, \ldots, m$ and $d_i^+(z)$, $i = 1, 2, \ldots, r$. The merit function may also depend on estimates of the Lagrange multipliers for the constraints in (1.1). It is sometimes difficult to appropriately choose weighting parameters in these merit functions in a way that drives the iterates to a solution (or at least a point satisfying Karush–Kuhn–Tucker (KKT) conditions) of (1.1).

In this paper, we propose an algorithm called Algorithm FP-SQP (feasibility perturbed SQP), in which all iterates z^k are feasible; that is, $z^k \in \mathcal{F}$ for all k. We obtain a step by solving a problem of the form (1.2a,b)at a feasible point $z \in \mathcal{F}$ with a trust-region constraint of the form (1.3). We then find a perturbation $\widetilde{\Delta z}$ of the step Δz that satisfies the following two crucial properties. First, *feasibility*:

$$z + \widetilde{\Delta z} \in \mathcal{F}; \tag{1.4}$$

second, *asymptotic exactness*: There is a continuous monotonically increasing function $\phi : \mathsf{R}^+ \to [0, 1/2]$ with $\phi(0) = 0$ such that

$$\|\Delta z - \widetilde{\Delta z}\|_2 \le \phi(\|\Delta z\|_2)\, \|\Delta z\|_2. \tag{1.5}$$

Note that because $\phi(t) \le 1/2$ for all $t \ge 0$, we have that

$$(1/2)\|\Delta z\|_2 \le \|\widetilde{\Delta z}\|_2 \le (3/2)\|\Delta z\|_2. \tag{1.6}$$

These conditions on $\widetilde{\Delta z}$ suffice to prove good global convergence properties for the algorithm. Additional assumptions on the feasibility perturbation technique can be made to obtain fast local convergence; see section 4.

The effectiveness of our method depends on its ability to calculate efficiently a perturbed step $\widetilde{\Delta z}$ with properties (1.4) and (1.5). This task is not difficult for certain structured problems, including some problems in optimal control. Additionally, in the

special case in which the constraints c and d are linear, we can simply set $\widetilde{\Delta z} = \Delta z$. When some constraints are nonlinear, $\widetilde{\Delta z}$ can be obtained from the projection of $z + \Delta z$ onto the feasible set $\mathcal{F}$. For general problems, this projection is nontrivial to compute, but for problems with structured constraints, it may be inexpensive.

By maintaining feasible iterates, our method gains several advantages. First, the trust-region restriction (1.3) can be added to the SQP problem (1.2a,b)without concern as to whether it will yield an infeasible subproblem. There is no need for a composite-step approach such as those mentioned above [3, 12]. Second, the objective function f can itself be used as a merit function. Third, if the algorithm is terminated early, we will be able to use the latest iterate z^k as a *feasible* suboptimal point, which in many applications is preferable to an infeasible suboptimum.

The advantages stated above are, of course, shared by other feasible SQP methods. The FSQP approach described in Lawrence and Tits [10] (based on an earlier version of Panier and Tits [13] and also using ideas from Birge, Qi, and Wei [2]) calculates the main search direction via a modified SQP subproblem that includes a parameter for "tilting" the search direction toward the interior of the set defined by the inequality constraints. A second subproblem is solved to obtain a second-order correction, and an "arc search" is performed along these two directions to find a new iterate that satisfies feasibility as well as a sufficient decrease condition in the objective f. The approach can also handle nonlinear equality constraints, but feasibility is not maintained with respect to these constraints in general. Our algorithm below differs in that it uses a trust region rather than arc searches to attain global convergence, it requires feasibility with respect to both inequality and equality constraints at each iteration, and it is less specific than in [10] about how the step is calculated. In this sense, Algorithm FP-SQP represents an algorithmic *framework* rather than a specific algorithm.

Heinkenschloss [8] considers projected SQP methods for problems with equality constraints in addition to bounds on a subset of the variables. He specifically targets optimal control problems with bounds on the controls—a set of problems similar to those we discuss in a companion paper [15]. The linearized equality constraints are used to express the free variables in terms of the bounded variables, and a projected Newton direction (see [1]) is constructed for the bounded variables. The step is computed by performing a line search along this direction with projection onto the bound constraints. This method contrasts with ours not only because it uses a line search rather than a trust region, but also because feasibility is not enforced with respect to the equality constraints; thus an augmented Lagrangian merit function must be used to determine an acceptable step length.

Other related work includes the feasible algorithm for problems with convex constraints discussed in Conn, Gould, and Toint [5]. At each iteration, this algorithm seeks an approximate minimizer of the model function over the intersection of the trust region with the original feasible set. The algorithm is targeted to problems in which the constraint set is simple (especially bound-constrained problems with ∞-norm trust regions, for which the intersection is defined by componentwise bounds). Aside from not requiring convexity, our method could be viewed as a particular instance of Algorithm 12.2.1 of [5, p. 452], in which the model function approximates the Lagrangian and the trial step is a perturbed SQP step. It may then be possible to apply the analysis of [5], once we show that the step generated in this fashion satisfies the assumptions in [5], at least for sufficiently small values of the trust-region radius. It appears nontrivial, however, to put our algorithm firmly into the framework of [5] and to extend the latter algorithm to handle a class of problems (featuring nonconvexity)

which its designers did not have in mind. Therefore, we present an analysis that was developed independently of that in [5]. We note that several features of the analysis in [5, Chapter 12] are similar to ours; for instance, χ in [5, p. 452] is similar to $C(z,1)$ defined below in (3.1), except that minimization in χ is taken over the original feasible set rather than over its linearized approximation, as in (3.1). Other aspects of the analysis in [5] and this paper are different; for instance, the generalized Cauchy point in [5, section 12.2.1] is defined in a much more complex fashion with respect to the projected-gradient path, rather than along the straight line as in Lemma 3.3 below.

The remainder of the paper is structured as follows. The algorithm is specified in section 2, and in section 2.1 we show that it is possible to find a feasible perturbation of the SQP step that satisfies requirements (1.4) and (1.5). We present global convergence results in section 3. After some basic lemmas in section 3.1, we describe in section 3.2 conditions under which the algorithm has at least one limit point that either fails a constraint qualification or satisfies KKT conditions. In particular, we assume in this section that the approximate Hessian H_k in (1.2) satisfies the bound $\|H_k\|_2 \leq \sigma_0 + \sigma_1 k$ for some constant σ_0 and σ_1—a type of bound often satisfied by quasi-Newton update formulae. In section 3.3, we make the stronger assumption that $\|H_k\|$ is uniformly bounded and prove the stronger result that *all* limit points of the algorithm either fail a constraint qualification or else satisfy KKT conditions. Under stronger assumptions on the limit point z^* and the feasibility projection technique, we prove fast local convergence in section 4. Some final comments appear in section 5.

A companion report of Tenny, Wright, and Rawlings [15] describes application of the algorithm to nonlinear optimization problems arising in model predictive control.

1.1. Optimality results and notation. The Lagrangian function for (1.1) is

$$\mathcal{L}(z,\mu,\lambda) \stackrel{\text{def}}{=} f(z) + \mu^T c(z) + \lambda^T d(z), \tag{1.7}$$

where $\mu \in \mathsf{R}^m$ and $\lambda \in \mathsf{R}^r$ are Lagrange multipliers for the constraints. The KKT conditions for (1.1) are as follows:

$$\nabla_z \mathcal{L}(z,\mu,\lambda) = \nabla f(z) + \nabla c(z)\mu + \nabla d(z)\lambda = 0, \tag{1.8a}$$

$$c(z) = 0, \tag{1.8b}$$

$$0 \geq d(z) \perp \lambda \geq 0, \tag{1.8c}$$

where $\perp$ indicates that $\lambda^T d(z) = 0$. We refer to any point z such that there exist μ and λ satisfying the conditions (1.8) as a *KKT point*.

For any feasible point z, we denote the *active set* $\mathcal{A}(z)$ as follows:

$$\mathcal{A}(z) \stackrel{\text{def}}{=} \{i = 1, 2, \ldots, r \,|\, d_i(z) = 0\}. \tag{1.9}$$

To ensure that the tangent cone to the constraint set at a feasible point z adequately captures the geometry of the feasible set near z, a constraint qualification must be satisfied at z. In the global convergence analysis of section 3, we use the Mangasarian–Fromovitz constraint qualification (MFCQ), which requires that

$$\nabla c(z) \text{ has full column rank, and} \tag{1.10a}$$

there exists a vector $v \in \mathsf{R}^n$ such that

$$\nabla c(z)^T v = 0 \text{ and } v^T \nabla d_i(z) < 0 \text{ for all } i \in \mathcal{A}(z). \tag{1.10b}$$

A more stringent constraint qualification, used in the local convergence analysis of section 4, is the linear independence constraint qualification (LICQ), which requires that

$$\{\nabla c_i(z),\ i = 1, 2, \ldots, m\} \cup \{\nabla d_i(z),\ i \in \mathcal{A}(z)\} \text{ is linearly independent.} \tag{1.11}$$

If z is a solution of (1.1), at which a constraint qualification such as (1.10) or (1.11) is satisfied, there exist vectors μ and λ such that (1.8) is satisfied by the triplet (z, μ, λ).

We say that the *strict complementarity* condition is satisfied at the KKT point z if, for some choice of the Lagrange multiplier vectors μ and λ satisfying the conditions (1.8), we have

$$\lambda - d(z) > 0. \tag{1.12}$$

That is, $\lambda_i > 0$ for all $i \in \mathcal{A}(z)$.

We use $\mathcal{B}(z, t)$ to denote the open ball (in the Euclidean norm) of radius t about z. When the subscript on the norm $\|\cdot\|$ is omitted, the Euclidean norm $\|\cdot\|_2$ is to be understood. The closure of a set L is denoted by $\mathrm{cl}(L)$.

We use order notation in the following way: If two matrix, vector, or scalar quantities M and A are functions of a common quantity, we write $M = O(\|A\|)$ if there is a constant β such that $\|M\| \le \beta\|A\|$ whenever $\|A\|$ is sufficiently small. We write $M = \Omega(\|A\|)$ if there is a constant β such that $\|M\| \ge \beta^{-1}\|A\|$ whenever $\|A\|$ is sufficiently small. We write $M = o(\|A\|)$ if there is a continuous, increasing function $\phi : \mathsf{R} \to \mathsf{R}$ with $\phi(0) = 0$ such that $\|M\| \le \phi(\|A\|)\|A\|$ for all $\|A\|$ sufficiently small.

2. The algorithm. In specifying the algorithm, we assume only that the perturbed step $\widetilde{\Delta z}$ satisfies (1.4) and (1.5), without specifying how it is calculated. As with all trust-region algorithms, a critical role is played by the ratio of actual to predicted decrease, which is defined for a given SQP step Δz^k and its perturbed counterpart $\widetilde{\Delta z}^k$ as follows:

$$\rho_k = \frac{f(z^k) - f\left(z^k + \widetilde{\Delta z}^k\right)}{-m_k(\Delta z^k)}. \tag{2.1}$$

The algorithm is specified as follows.

ALGORITHM 2.1 (FP-SQP). *Given starting point $z_0 \in \mathcal{F}$, initial radius $\Delta_0 \in (0, \bar{\Delta}]$, initial scaling matrix D_0, trust-region upper bound $\bar{\Delta} \ge 1$, $\eta \in (0, 1/4)$, and $p \in [1, \infty]$;*

for $k = 0, 1, 2, \ldots$

Obtain Δz^k by solving (1.2), (1.3)*;*

if $m_k(\Delta z^k) = 0$

STOP;

Seek $\widetilde{\Delta z}^k$ with the properties (1.4) *and* (1.5)*;*

if *no such $\widetilde{\Delta z}^k$ is found;*

$\Delta_{k+1} \leftarrow (1/2)\|D_k \Delta z^k\|_p$*;*

$z^{k+1} \leftarrow z^k$*;* $D_{k+1} \leftarrow D_k$*;*

else

Calculate ρ_k using (2.1)*;*

if $\rho_k < 1/4$

$\Delta_{k+1} \leftarrow (1/2)\|D_k \Delta z^k\|_p$;
else **if** $\rho_k > 3/4$ *and* $\|D_k \Delta z^k\|_p = \Delta_k$
$\Delta_{k+1} \leftarrow \min(2\Delta_k, \bar{\Delta})$;
else
$\Delta_{k+1} \leftarrow \Delta_k$;
if $\rho_k \geq \eta$
$z^{k+1} \leftarrow z^k + \widetilde{\Delta z}^k$;
choose new scaling matrix D_{k+1};
else
$z^{k+1} \leftarrow z^k$; $D_{k+1} \leftarrow D_k$;

end (for).

We now state some assumptions that are used in the subsequent analysis. We start by defining the level set L_0 as follows:

$$L_0 \stackrel{\text{def}}{=} \{z \,|\, c(z) = 0,\ d(z) \leq 0,\ f(z) \leq f(z_0)\} \subset \mathcal{F}.$$

Our assumption on the trust-region bound (1.3) is as follows.

ASSUMPTION 1. *There is a constant* δ *such that, for all points* $z \in L_0$ *and all scaling matrices* D *used by the algorithm, the following conditions hold:*

(a) D *is uniformly bounded.*

(b) *We have for any* Δz *satisfying the constraints*

$$c(z) + \nabla c(z)^T \Delta z = 0, \quad d(z) + \nabla d(z)^T \Delta z \leq 0$$

that

$$\delta^{-1}\|\Delta z\|_2 \leq \|D\Delta z\|_p \leq \delta \|\Delta z\|_2. \tag{2.2}$$

In this assumption, the constant that relates $\|\cdot\|_2$ with the equivalent norms $\|\cdot\|_p$ for all p between 1 and ∞ is absorbed into δ. Note that the right-hand inequality in (2.2) is implied by part (a) of the assumption.

Note that for unconstrained problems (in which c and d are vacuous), the left-hand inequality in (2.2) is satisfied when all scaling matrices D used by the algorithm have bounded inverse. Another special case relevant to optimal control problems occurs when the constraints have the form

$$c(u, v) = 0, \quad c : \mathsf{R}^{n-m} \times \mathsf{R}^m \to \mathsf{R}^m, \tag{2.3}$$

(that is, $u \in \mathsf{R}^{n-m}$ and $v \in \mathsf{R}^m$), and the trust-region constraint is imposed only on the u variables; that is,

$$\|D_u \Delta u\|_p \leq \Delta, \tag{2.4}$$

where D_u is a diagonal matrix with positive diagonal elements. The linearized constraints (1.2b) then have the form

$$\nabla_u c(u, v)^T \Delta u + \nabla_v c(u, v)^T \Delta v = 0 \tag{2.5}$$

which, if $\nabla_v c(u, v)$ is invertible, leads to

$$\Delta v = -\left(\nabla_v c(u, v)\right)^{-T} \nabla_u c(u, v)^T \Delta u.$$

If we assume that $\nabla_v c(u,v)$ is invertible for all points (u,v) in the region of interest, with $\|(\nabla_v c(u,v))^{-1}\|$ bounded, we can define a constant $\hat{\delta} > 0$ such that $\|\Delta v\|_p \leq \hat{\delta}\|\Delta u\|_p$. We then have

$$\|(\Delta u, \Delta v)\|_p \leq (1+\hat{\delta})\|\Delta u\|_p \leq (1+\hat{\delta})D_{\min}^{-1}\|D_u \Delta u\|_p = (1+\hat{\delta})D_{\min}^{-1}\|D(\Delta u, \Delta v)\|_p,$$

where we define $D_{\min}$ to be a lower bound on the diagonals of D_u, and $D = \mathrm{diag}(D_u, 0)$. On the other hand, we have

$$\|D(\Delta u, \Delta v)\|_p = \|D_u \Delta u\|_\infty \leq D_{\max}\|\Delta u\|_p \leq D_{\max}\|(\Delta u, \Delta v)\|_p,$$

where $D_{\max}$ is an upper bound on the diagonals of D_u. It follows from the last two expressions that Assumption 1 is satisfied in this situation.

For some results we make an assumption on the boundedness of the level set L_0 and on the smoothness of the objective and constraint functions.

ASSUMPTION 2. *The level set L_0 is bounded, and the functions f, c, and d in* (1.1) *are twice continuously differentiable in an open neighborhood $\mathcal{N}(L_0)$ of this set.*

Note that L_0 is certainly closed so that, if Assumption 2 holds, it is also compact.

2.1. Algorithm FP-SQP is well defined. We show first that the algorithm is well defined, in the sense that given a feasible point z_k, a step $\widetilde{\Delta z}_k$ satisfying (1.4) and (1.5) can be found for all sufficiently small Δ_k, under certain assumptions.

We note first that whenever $z = z_k$ is feasible and Assumption 1 holds, the subproblem (1.2), (1.3) has a solution. This fact follows from nonemptiness, closedness, and boundedness of the feasible set for the subproblem. To show that there exists $\widetilde{\Delta z}_k$ satisfying (1.4) and (1.5), we use the following assumption.

ASSUMPTION 3. *For every point $\hat{z} \in L_0$, there are positive quantities ζ and $\hat{\Delta}_3$ such that, for all $z \in \mathrm{cl}(\mathcal{B}(\hat{z}, \delta\hat{\Delta}_3))$, we have*

$$\min_{v \in \mathcal{F}} \|v - z\| \leq \zeta\left(\|c(z)\| + \|[d(z)]_+\|\right), \tag{2.6}$$

where δ is the constant from Assumption 1 *and $[d(z)]_+ = [\max(d_i(z), 0)]_{i=1}^r$. (Recall our convention that $\|\cdot\|$ denotes $\|\cdot\|_2$.)*

This assumption requires the constraint system to be regular enough near each feasible point that a bound like that of Hoffman [9] for systems of linear equalities and inequalities is satisfied. Assumption 3 is essentially the same as Assumption C of Lucidi, Sciandrone, and Tseng [11]. A result of Robinson [14, Corollary 1] shows that Assumption 3 is satisfied whenever MFCQ is satisfied at all points in L_0. The following result shows that a bound similar to (2.6) also holds *locally* in the vicinity of a feasible point satisfying MFCQ.

LEMMA 2.1. *Let $\hat{z}$ be a feasible point for* (1.1) *at which MFCQ is satisfied. Then there exist positive quantities ζ and $\hat{R}_1$ such that for all $z \in \mathrm{cl}(\mathcal{B}(\hat{z}, \hat{R}_1))$, the bound* (2.6) *is satisfied.*

Proof. We first choose $\bar{R}_1$ small enough such that for all $\tilde{z} \in \mathrm{cl}(\mathcal{B}(\hat{z}, \bar{R}_1)) \cap \mathcal{F}$, we have that $\mathcal{A}(\tilde{z}) \subset \mathcal{A}(\hat{z})$, where $\mathcal{A}(\cdot)$ is defined by (1.9). Let v be a vector satisfying (1.10) at $z = \hat{z}$ and assume, without loss of generality, that $\|v\|_2 = 1$. Because $\nabla c(\hat{z})$ has full column rank, we have, by decreasing $\bar{R}_1$ if necessary, that for any $\tilde{z} \in \mathrm{cl}(\mathcal{B}(\hat{z}, \bar{R}_1))$, $\nabla c(\tilde{z})$ also has full column rank. Moreover, using the full rank of $\nabla c(\tilde{z})$, we can find a perturbation $\tilde{v}$ of v satisfying $\|\tilde{v} - v\| = O(\|\tilde{z} - \hat{z}\|)$ and (after possibly decreasing $\bar{R}_1$ again) $\|\tilde{v}\| \geq 0.5$, such that

$$\nabla c(\tilde{z})^T \tilde{v} = 0 \text{ and } \tilde{v}^T \nabla d_i(\tilde{z}) < 0 \text{ for all } i \in \mathcal{A}(\hat{z}) \supset \mathcal{A}(\tilde{z}), \text{ for all } \tilde{z} \in \mathrm{cl}(\mathcal{B}(\hat{z}, \bar{R}_1)) \cap \mathcal{F}.$$

Hence, the MFCQ condition is satisfied for all $\tilde{z} \in \mathrm{cl}(\mathcal{B}(\hat{z}, \bar{R}_1)) \cap \mathcal{F}$.

We now appeal to Corollary 1 of Robinson [14]. From this result, we have that there is $\zeta > 0$ (depending on $\hat{z}$ but not on $\tilde{z}$) and an open neighborhood $M(\tilde{z})$ of each $\tilde{z} \in \mathrm{cl}(\mathcal{B}(\hat{z}, \bar{R}_1)) \cap \mathcal{F}$ such that (2.6) holds for all $z \in M(\tilde{z})$. Since

$$\hat{M}(\hat{z}) \stackrel{\mathrm{def}}{=} \cup_{\tilde{z}} \{M(\tilde{z}) \,|\, \tilde{z} \in \mathrm{cl}(\mathcal{B}(\hat{z}, \bar{R}_1)) \cap \mathcal{F}\}$$

is an open neighborhood of the compact set $\mathrm{cl}(\mathcal{B}(\hat{z}, \bar{R}_1)) \cap \mathcal{F}$, we can define $\hat{R}_1 \leq \bar{R}_1$ small enough that $\mathrm{cl}(\mathcal{B}(\hat{z}, \hat{R}_1)) \subset \hat{M}(\hat{z})$. Thus, since (2.6) holds for all $z \in M(\tilde{z})$, our proof is complete. □

We observed above that, under Assumption 1, the solution Δz of (1.2a,b), (1.3) is well defined. Using the other assumptions, we now show that $\widetilde{\Delta z}$ satisfying the properties (1.4) and (1.5) can also be found, so that Algorithm FP-SQP is well defined.

THEOREM 2.2. *Suppose that Assumptions* 1, 2, *and* 3 *are satisfied. Then there is a positive constant* Δ_{def} *such that for any* $z \in L_0$ *and any* $\Delta \leq \Delta_{\mathrm{def}}$, *there is a step* $\widetilde{\Delta z}$ *that satisfies the properties* (1.4) *and* (1.5), *where* Δz *is the solution of* (1.2a,b), (1.3) *for the given values of* z *and* δ.

Proof. We show that the result holds for the function $\phi(t) = \min(1/2, \sqrt{t})$ in (1.5).

We first choose $\hat{\Delta}_0$ small enough that $\mathcal{B}(z, \delta\hat{\Delta}_0) \subset \mathcal{N}(L_0)$ for all $z \in L_0$, where $\mathcal{N}(L_0)$ is defined in Assumption 2. Thus, for $\Delta \leq \hat{\Delta}_0$ and Δz solving (1.2a,b), (1.3), we have for all $\alpha \in [0, 1]$ that

$$\|\alpha \Delta z\| \leq \|\Delta z\| \leq \delta \|D\Delta z\|_p \leq \delta \hat{\Delta}_0, \tag{2.7}$$

so that $z + \alpha\Delta z \in \mathcal{N}(L_0)$.

Given any $\hat{z} \in L_0$, we seek a positive constant $\hat{\Delta}$ such that for all $z \in \mathrm{cl}(\mathcal{B}(\hat{z}, \delta\hat{\Delta}/2)) \cap \mathcal{F}$, and all $\Delta \leq \hat{\Delta}/2$, there is a step $\widetilde{\Delta z}$ that satisfies properties (1.4) and (1.5).

We choose initially $\hat{\Delta} = \hat{\Delta}_0$ and assume that Δz satisfies $\|D\Delta z\|_p \leq \Delta$, which implies from Assumption 1 and the definitions of Δ and $\hat{\Delta}$ that

$$\|\Delta z\| \leq \delta \|D\Delta z\|_p \leq \delta\Delta \leq \delta\hat{\Delta}/2 < \delta\hat{\Delta}_0.$$

From feasibility of z, (2.7), (1.2b), and the fact that c and d are twice continuously differentiable in $\mathcal{N}(L_0)$, we have that

$$c(z + \Delta z) = c(z) + \nabla c(z)^T \Delta z + O(\|\Delta z\|^2) = O(\|\Delta z\|^2)$$

and

$$[d(z + \Delta z)]_+ = \left[d(z) + \nabla d(z)^T \Delta z + O(\|\Delta z\|^2)\right]_+ = O(\|\Delta z\|^2).$$

We now set $\hat{\Delta} \leftarrow \min(\hat{\Delta}, \hat{\Delta}_3)$ and apply Assumption 3. Since

$$\|(z + \Delta z) - \hat{z}\| \leq \|z - \hat{z}\| + \|\Delta z\| \leq \delta\hat{\Delta}/2 + \delta\hat{\Delta}/2 \leq \delta\hat{\Delta}_3,$$

we have from Assumption 3 and the estimates above that

$$\min_{v \in \mathcal{F}} \|v - (z + \Delta z)\| \leq \zeta \left(\|c(z + \Delta z)\| + \|[d(z + \Delta z)]_+\|\right) = O(\zeta \|\Delta z\|^2), \tag{2.8}$$

where ζ may depend on $\hat{z}$. Since $v = z$ is feasible for (2.8), we have that any solution of this projection problem satisfies $\|v - (z + \Delta z)\| \leq \|\Delta z\|$. Hence, the minimization

on the left-hand side of (2.8) may be restricted to the nonempty compact set $\mathrm{cl}(\mathcal{B}(z+\Delta z, \|\Delta z\|)) \cap \mathcal{F}$, so the minimum is attained. If we use the minimizer v to define $\widetilde{\Delta z} = v - z$, then from (2.8) we have

$$\|\widetilde{\Delta z} - \Delta z\| = O(\zeta \|\Delta z\|^2).$$

Therefore, by decreasing $\hat{\Delta}$ if necessary, we find that (1.5) is satisfied for our choice $\phi(t) = \min(1/2, \sqrt{t})$.

The set of open Euclidean balls $\mathcal{B}(\hat{z}, \delta\hat{\Delta}/2)$, $\hat{z} \in L_0$, forms an open cover of L_0. Since L_0 is compact, we can define a finite subcover. By defining Δ_{def} to be the minimum of the $\hat{\Delta}/2$ over the subcover, we have that Δ_{def} is positive and has the desired property. ∎

3. Global convergence. In this section, we prove convergence to KKT points of (1.1). Our results are of two types. We show first in section 3.2 that if Algorithm FP-SQP does not terminate finitely (at a KKT point), it has a limit point that either satisfies the MFCQ and KKT conditions or else fails to satisfy MFCQ. In section 3.3, we show, under a stronger assumption on the approximate Hessian H_k, that *all* limit points either fail to satisfy MFCQ or else satisfy both MFCQ and KKT.

We start with some technical results.

3.1. Technical results. The first result concerns the solution of a linear programming variant of the SQP subproblem (1.2a,b), (1.3). Its proof appears in the appendix.

LEMMA 3.1. *Let f, c, and d be as defined in* (1.1)*, and let $C(z,\tau)$ denote the* negative *of the value function of the following problem, for some $z \in \mathcal{F}$ and $\tau > 0$:*

$$\text{(3.1a)} \qquad \mathrm{CLP}(z,\tau): \qquad \min_w \nabla f(z)^T w \ \textit{subject to}$$

$$\text{(3.1b)} \qquad c(z) + \nabla c(z)^T w = 0, \ \ d(z) + \nabla d(z)^T w \le 0, \ \ w^T w \le \tau^2.$$

For any point $\bar{z} \in \mathcal{F}$, we have $C(\bar{z}, 1) \ge 0$ with $C(\bar{z}, 1) = 0$ if and only if $\bar{z}$ is a KKT point (1.8)*.*

When the MFCQ conditions (1.10a,b) *are satisfied at $\bar{z}$, but $\bar{z}$ is not a KKT point, there exist positive quantities R_2 and ϵ such that for any $z \in \mathcal{B}(\bar{z}, R_2) \cap \mathcal{F}$, we have $C(z,1) \ge \epsilon$.*

An immediate consequence of this result is that for any subsequence $\{z^k\}_{k\in\mathcal{K}}$ such that $z^k \to \bar{z}$ and $C(z^k, 1) \to 0$, where $\bar{z}$ satisfies the MFCQ conditions, we must have that $\bar{z}$ is a KKT point for (1.1).

Note that $C(z,\tau)$ is an increasing concave function of $\tau > 0$. In particular, if $w(z,\tau)$ attains the optimum in $\mathrm{CLP}(z,\tau)$, the point $\alpha w(z,\tau)$ is feasible in $\mathrm{CLP}(z,\alpha\tau)$ for all $\alpha \in [0,1]$, so that

$$\text{(3.2)} \qquad C(z,\alpha\tau) \ge \alpha C(z,\tau), \quad \text{for all } \tau > 0, \text{ for all } \alpha \in [0,1].$$

For convenience, we restate the subproblem (1.2a,b), (1.3) at an arbitrary feasible point z as follows:

$$\text{(3.3a)} \qquad \min_{\Delta z} m(\Delta z) \stackrel{\mathrm{def}}{=} \nabla f(z)^T \Delta z + \tfrac{1}{2}\Delta z^T H \Delta z \ \ \text{subject to}$$

$$\text{(3.3b)} \qquad c(z) + \nabla c(z)^T \Delta z = 0, \ \ d(z) + \nabla d(z)^T \Delta z \le 0,$$

$$\text{(3.3c)} \qquad \|D\Delta z\|_p \le \Delta,$$

where D satisfies Assumption 1. Consider now the following problem, obtained by omitting the quadratic term from (3.3a):

$$\min_{\Delta z^{\mathrm{L}}} \nabla f(z)^T \Delta z^{\mathrm{L}} \quad \text{subject to} \tag{3.4a}$$

$$c(z) + \nabla c(z)^T \Delta z^{\mathrm{L}} = 0, \quad d(z) + \nabla d(z)^T \Delta z^{\mathrm{L}} \leq 0, \tag{3.4b}$$

$$\|D\Delta z^{\mathrm{L}}\|_p \leq \Delta. \tag{3.4c}$$

Denote the *negative* of the value function for this problem by $V(z, D, \Delta)$. Referring to (3.1) and Assumption 1, we see that the feasible region for $\mathrm{CLP}(z, \delta^{-1}\Delta)$ is contained in the feasible region for (3.4), and the objectives are the same. Hence for $\Delta \in (0, 1]$, we have from (3.2) that

$$V(z, D, \Delta) \geq C(z, \delta^{-1}\Delta) \geq \delta^{-1} C(z, 1)\Delta.$$

For $\Delta > 1$, on the other hand, we have

$$V(z, D, \Delta) \geq C(z, \delta^{-1}\Delta) \geq \delta^{-1} C(z, \Delta) \geq \delta^{-1} C(z, 1).$$

Hence, by combining these observations, we obtain that

$$V(z, D, \Delta) \geq \delta^{-1} C(z, 1) \min(1, \Delta). \tag{3.5}$$

The following result, together with Lemma 3.1, is an immediate consequence of (3.5).

LEMMA 3.2. *Suppose that Assumption* 1 *holds. Let* $\bar{z} \in L_0$ *satisfy the MFCQ conditions* (1.10a,b) *but not the KKT conditions* (1.8a–c). *Then there exist positive quantities* R_2 *and* ϵ *such that for any* $z \in \mathcal{B}(\bar{z}, R_2) \cap \mathcal{F}$ *and any* $\Delta > 0$, *we have*

$$C(z, 1) \geq \epsilon, \tag{3.6a}$$

$$V(z, D, \Delta) \geq \delta^{-1} \epsilon \min(1, \Delta), \tag{3.6b}$$

where $V(\cdot, \cdot, \cdot)$ *is the negative of the value function for* (3.4).

If Assumption 1 holds, we have that

$$\|\Delta z^{\mathrm{L}}\|_2 \leq \delta \|D\Delta z^{\mathrm{L}}\|_p \leq \delta\Delta. \tag{3.7}$$

Hence, since Δz is optimal for (3.3), and since Δz^{L} that solves (3.4) is feasible for this problem, we have

$$\begin{aligned} m(\Delta z) &\leq m(\Delta z^{\mathrm{L}}) \\ &= (\Delta z^{\mathrm{L}})^T \nabla f(z) + \tfrac{1}{2}(\Delta z^{\mathrm{L}})^T H(\Delta z^{\mathrm{L}}) \\ &\leq -V(z, D, \Delta) + \tfrac{1}{2}\delta^2 \|H\| \Delta^2 \\ &\leq -\delta^{-1} \min(1, \Delta) C(z, 1) + \tfrac{1}{2}\delta^2 \|H\| \Delta^2, \end{aligned} \tag{3.8}$$

where the last inequality follows from (3.5).

We now define the Cauchy point for problem (3.3a–c) as

$$\Delta z^{\mathrm{C}} = \alpha^{\mathrm{C}} \Delta z^{\mathrm{L}}, \tag{3.9}$$

where

$$\alpha^{\mathrm{C}} = \arg \min_{\alpha \in [0,1]} \alpha \nabla f(z)^T \Delta z^{\mathrm{L}} + \tfrac{1}{2}\alpha^2 (\Delta z^{\mathrm{L}})^T H \Delta z^{\mathrm{L}}. \tag{3.10}$$

We show that Δz^{C} has the following property:

$$m(\Delta z^{\mathrm{C}}) \leq -\tfrac{1}{2} C(z,1) \min\left[\delta^{-1}, \delta^{-1}\Delta, (\delta^4 \bar{\Delta}^2 \|H\|_2)^{-1} C(z,1)\right], \tag{3.11}$$

where $\bar{\Delta}$ is defined in Algorithm FP-SQP. We prove (3.11) by considering two cases. First, when $(\Delta z^{\mathrm{L}})^T H \Delta z^{\mathrm{L}} \leq 0$, we have $\alpha^{\mathrm{C}} = 1$ in (3.10), and hence $\Delta z^{\mathrm{C}} = \Delta z^{\mathrm{L}}$. Similarly to (3.8), but using $(\Delta z^{\mathrm{L}})^T H \Delta z^{\mathrm{L}} \leq 0$ together with (3.5), we have

$$m(\Delta z^{\mathrm{C}}) = m(\Delta z^{\mathrm{L}}) \leq -V(z, D, \Delta) \leq -\delta^{-1} C(z,1) \min(1, \Delta),$$

so result (3.11) holds in this case. In the alternate case $(\Delta z^{\mathrm{L}})^T H \Delta z^{\mathrm{L}} > 0$, we have

$$\alpha = \min\left(1, \frac{-\nabla f(z)^T \Delta z^{\mathrm{L}}}{(\Delta z^{\mathrm{L}})^T H \Delta z^{\mathrm{L}}}\right). \tag{3.12}$$

If the minimum is achieved at 1, we have from $(\Delta z^{\mathrm{L}})^T H \Delta z^{\mathrm{L}} \leq -\nabla f(z)^T \Delta z^{\mathrm{L}}$ and (3.5) that

$$m(\Delta z^{\mathrm{C}}) = m(\Delta z^{\mathrm{L}}) \leq \tfrac{1}{2} \nabla f(z)^T \Delta z^{\mathrm{L}} \leq -\tfrac{1}{2} \delta^{-1} C(z,1) \min(1, \Delta), \tag{3.13}$$

and therefore (3.11) again is satisfied. If the min in (3.12) is achieved at $-\nabla f(z)^T \Delta z^{\mathrm{L}} / (\Delta z^{\mathrm{L}})^T H \Delta z^{\mathrm{L}}$, we have from (3.5) that

$$m(\Delta z^{\mathrm{C}}) = m(\alpha \Delta z^{\mathrm{L}}) = -\frac{1}{2} \frac{(\nabla f(z)^T \Delta z^{\mathrm{L}})^2}{(\Delta z^{\mathrm{L}})^T H \Delta z^{\mathrm{L}}} \leq -\frac{1}{2} \frac{\delta^{-2} \min(1, \Delta^2) C(z,1)^2}{\|H\|_2 \|\Delta z^{\mathrm{L}}\|_2^2}. \tag{3.14}$$

Because of (3.7), we have from (3.14) that

$$\begin{aligned} m(\Delta z^{\mathrm{C}}) &\leq -\frac{1}{2} \frac{\delta^{-2} \min(1, \Delta^2) C(z,1)^2}{\delta^2 \Delta^2 \|H\|_2} \\ &= -\frac{1}{2} (\delta^4 \|H\|_2)^{-1} \min(1, \Delta^{-2}) C(z,1)^2 \leq -\frac{1}{2} (\delta^4 \bar{\Delta}^2 \|H\|_2)^{-1} C(z,1)^2, \end{aligned}$$

which again implies that (3.11) is satisfied.

Since Δz^{C} is feasible for (3.3), we have proved the following lemma.

LEMMA 3.3. *Suppose that $z \in L_0$ and that Assumption 1 holds. Suppose that Δz^C is obtained from* (3.4a–c), (3.9), *and* (3.10). *Then the decrease in the model function m obtained by the point Δz^C satisfies the bound* (3.11), *and therefore the solution Δz of* (3.3a–c) *satisfies the similar bound*

$$m(\Delta z) \leq -\tfrac{1}{2} C(z,1) \min\left[\delta^{-1}, \delta^{-1}\Delta, (\delta^4 \bar{\Delta}^2 \|H\|_2)^{-1} C(z,1)\right], \tag{3.15}$$

where $C(z,1)$ is the negative of the value function of CLP$(z,1)$ *defined in* (3.1).

Note that this lemma holds even when we assume only that Δz is feasible for (3.3a–c) and satisfies $m(\Delta z) \leq m(\Delta z^{\mathrm{C}})$. This relaxation is significant since, when H is indefinite, the complexity of finding a solution of (3.3a–c) is greater than the complexity of computing Δz^{C}.

3.2. Result I: At least one KKT limit point. We now discuss convergence of the sequence of iterates generated by the algorithm under the assumptions of section 2 and the additional assumption that the Hessians H_k of (1.2) are bounded as follows:

$$\|H_k\|_2 \leq \sigma_0 + \sigma_1 k, \quad k = 0, 1, 2 \ldots. \tag{3.16}$$

The style of analysis follows that of a number of earlier works on convergence of trust-region algorithms for unconstrained, possibly nonsmooth, problems, for example, Yuan [17], Wright [16]. However, many modifications are needed to adapt the algorithms to constrained problems and to the algorithm of section 2.

We first prove a key lemma as a preliminary to the global convergence result of this section. It finds a lower bound on the trust-region radii for the case when no subsequence of $\{C(z^k,1)\}$ approaches zero.

LEMMA 3.4. *Suppose that Assumptions* 1, 2, *and* 3 *are satisfied and that there are* $\epsilon > 0$ *and an index* K *such that*

$$C(z^k,1) \geq \epsilon \ \textit{for all}\ k \geq K,$$

Then there is a constant $T > 0$ *such that*

$$\Delta_k \geq T/N_k \ \textit{for all}\ k \geq K, \tag{3.17}$$

where

$$N_k \stackrel{\text{def}}{=} 1 + \max_{i=0,1,\ldots,k} \|H_k\|_2.$$

Proof. For $\Delta_k \geq 1$, claim (3.17) obviously holds with $T = 1$. Hence, we assume for the remainder of the proof that $\Delta_k \in (0,1]$.

From Lemma 3.3, we have

$$\begin{aligned} -m_k(\Delta z^k) &\geq \tfrac{1}{2}\epsilon \min\left[\delta^{-1}\Delta_k, (\delta^4\bar{\Delta}^2\|H_k\|_2)^{-1}\epsilon\right] \\ &\geq \tfrac{1}{2}\epsilon \min\left[\delta^{-1}\Delta_k, (\delta^4\bar{\Delta}^2 N_k)^{-1}\epsilon\right]. \end{aligned} \tag{3.18}$$

We define the constants $\bar{\sigma}$ and γ as follows:

$$\bar{\sigma} = \sup\{\|\nabla^2 f(z)\|_2 \,|\, z \in \mathcal{N}(L_0)\}, \qquad \gamma = \sup\{\|\nabla f(z)\|_2 \,|\, z \in L_0\}, \tag{3.19}$$

where $\mathcal{N}(L_0)$ is the neighborhood defined in Assumption 2. Suppose now that T is chosen small enough to satisfy the following conditions:

$$T \leq 1, \tag{3.20a}$$

$$\{z \,|\, \operatorname{dist}(z, L_0) \leq 2\delta T\} \subset \mathcal{N}(L_0), \tag{3.20b}$$

$$2T \leq \epsilon/(\delta^3\bar{\Delta}^2), \tag{3.20c}$$

$$(\gamma + 2\bar{\sigma}\delta)\phi(2\delta T)\delta^2 \leq (1/48)\epsilon, \tag{3.20d}$$

$$2\bar{\sigma}\delta^3 T \leq (1/48)\epsilon, \tag{3.20e}$$

$$\delta^3 T \leq (1/48)\epsilon, \tag{3.20f}$$

where $\phi(\cdot)$ is defined in (1.5).

For any k with

$$\|\Delta z^k\| \leq 2\delta T, \tag{3.21}$$

we have from Taylor's theorem and the definition of m_k that

(3.22)

$$\begin{aligned} &f(z^k) - f(z^k + \widetilde{\Delta z}^k) + m_k(\Delta z^k) \\ &= -\nabla f(z^k)^T \widetilde{\Delta z}^k - \tfrac{1}{2}(\widetilde{\Delta z}^k)^T \nabla^2 f(z_\theta^k)\widetilde{\Delta z}^k + \nabla f(z^k)^T \Delta z^k + \tfrac{1}{2}(\Delta z^k)^T H_k \Delta z^k \\ &= \left[\nabla f(z^k) + \nabla^2 f(z_\theta^k)\Delta z^k\right]^T (\Delta z^k - \widetilde{\Delta z}^k) \\ &\qquad -\tfrac{1}{2}(\widetilde{\Delta z}^k - \Delta z^k)^T \nabla^2 f(z_\theta^k)(\widetilde{\Delta z}^k - \Delta z^k) - \tfrac{1}{2}(\Delta z^k)^T(\nabla^2 f(z_\theta^k) - H_k)\Delta z^k, \end{aligned}$$

where z_θ^k lies on the line segment between z^k and $z^k + \widetilde{\Delta z}^k$. If k is an index satisfying (3.21), we have from feasibility of both z^k and $z^k + \widetilde{\Delta z}^k$ that

$$\begin{aligned}\operatorname{dist}(z_\theta^k, L_0) &\le \tfrac{1}{2}\|\widetilde{\Delta z}^k\|_2 \\ &\le \tfrac{1}{2}\left(\|\Delta z^k\|_2 + \|\Delta z^k - \widetilde{\Delta z}^k\|_2\right) \\ &\le \tfrac{1}{2}\left(\|\Delta z^k\|_2 + \phi(\|\Delta z^k\|_2)\|\Delta z^k\|_2\right) \\ &\le \tfrac{1}{2}(2\delta T + \phi(2\delta T)2\delta T) \le 2\delta T,\end{aligned}$$

and therefore from (3.20b) and (3.19) we have $\|\nabla^2 f(z_\theta^k)\|_2 \le \bar\sigma$. For k satisfying (3.21), we have from (3.22) that

$$\begin{aligned}&\left|f(z^k) - f(z^k + \widetilde{\Delta z}^k) + m_k(\Delta z^k)\right| \\ &\le \left(\|\nabla f(z^k)\|_2 + \|\nabla^2 f(z_\theta^k)\|_2\|\Delta z^k\|_2\right)\|\Delta z^k - \widetilde{\Delta z}^k\|_2 \\ &\qquad + \tfrac{1}{2}\|\nabla^2 f(z_\theta^k)\|_2\|\widetilde{\Delta z}^k - \Delta z^k\|_2^2 + \tfrac{1}{2}(\|\nabla^2 f(z_\theta^k)\|_2 + \|H_k\|_2)\|\Delta z^k\|_2^2 \\ &\le (\gamma + 2\bar\sigma\delta T)\|\Delta z^k - \widetilde{\Delta z}^k\|_2 + \tfrac{1}{2}\bar\sigma\|\Delta z^k - \widetilde{\Delta z}^k\|_2^2 + \tfrac{1}{2}(\bar\sigma + N_k)\|\Delta z^k\|_2^2. \qquad (3.23)\end{aligned}$$

Now using (1.5) and Assumption 1, we have for indices k satisfying (3.21) that

(3.24)

$$\begin{aligned}&\left|f(z^k) - f(z^k + \widetilde{\Delta z}^k) + m_k(\Delta z^k)\right| \\ &\le (\gamma + 2\bar\sigma\delta T)\phi(\|\Delta z^k\|_2)\|\Delta z^k\|_2 + \frac{1}{2}\bar\sigma\phi(\|\Delta z^k\|_2)^2\|\Delta z^k\|_2^2 + \frac{1}{2}(\bar\sigma + N_k)\|\Delta z^k\|_2^2 \\ &\le \left[(\gamma + 2\bar\sigma\delta T)\phi(2\delta T) + \frac{1}{2}\bar\sigma\phi(2\delta T)^2 2\delta T + \bar\sigma\delta T + \frac{1}{2}N_k\|\Delta z^k\|_2\right]\|\Delta z^k\|_2 \\ &\le \left[(\gamma + 2\bar\sigma\delta T)\phi(2\delta T) + \bar\sigma\delta T + \bar\sigma\delta T + \frac{1}{2}N_k\|\Delta z^k\|_2\right]\|\Delta z^k\|_2 \\ &\le \left[\frac{1}{48}\frac{\epsilon}{\delta^2} + \frac{1}{48}\frac{\epsilon}{\delta^2} + \frac{1}{2}N_k\|\Delta z^k\|_2\right]\|\Delta z^k\|_2,\end{aligned}$$

where we used $\phi \le 1/2$, (3.20d), and (3.20e) to derive the various inequalities.

Now suppose that (3.17) is not satisfied for all k and for our choice of T, and suppose that l is the first index at which it is violated, that is,

$$\Delta_l < T/N_l. \qquad (3.25)$$

We exclude the case $l = K$ (by decreasing T further, if necessary), and consider the index $l-1$. Since $\Delta_k \ge (1/2)\|D_{k-1}\Delta z^{k-1}\|_p$ for all k, and since $N_l \ge 1$, we have

$$\|\Delta z^{l-1}\|_2 \le \delta\|D_{l-1}\Delta z^{l-1}\|_p \le 2\delta\Delta_l < 2\delta T, \qquad (3.26)$$

so that $l-1$ satisfies (3.21). Hence, bound (3.24) applies with $k = l-1$, and we have

$$\begin{aligned}&\left|f(z^{l-1}) - f\left(z^{l-1} + \widetilde{\Delta z}^{l-1}\right) + m_{l-1}(\Delta z^{l-1})\right| \\ &\le \left[\frac{1}{24}\frac{\epsilon}{\delta^2} + \frac{1}{2}N_{l-1}\|\Delta z^{l-1}\|_2\right]\|\Delta z^{l-1}\|_2. \qquad (3.27)\end{aligned}$$

Since $N_{l-1} \leq N_l$, we have from (3.26) and (3.25) that

$$N_{l-1}\|\Delta z^{l-1}\|_2 \leq 2\delta N_l \Delta_l < 2\delta T. \tag{3.28}$$

Therefore by using (3.27) and (3.20f), we obtain

$$\begin{aligned} &\left|f(z^{l-1}) - f\left(z^{l-1} + \widetilde{\Delta z}^{l-1}\right) + m_{l-1}(\Delta z^{l-1})\right| \\ &\leq \left(\frac{1}{24}\frac{\epsilon}{\delta^2} + \delta T\right)\|\Delta z^{l-1}\|_2 \leq \frac{1}{16}\frac{\epsilon}{\delta^2}\|\Delta z^{l-1}\|_2. \end{aligned} \tag{3.29}$$

Returning to the right-hand side of (3.18), we have for $k = l-1$ that

$$\delta^{-1}\Delta_{l-1} \geq \delta^{-1}\|D_{l-1}\Delta z^{l-1}\|_p \geq \delta^{-2}\|\Delta z^{l-1}\|_2,$$

and using (3.28) and (3.20c), we have

$$\frac{\epsilon}{\delta^4 \bar{\Delta}^2 N_{l-1}} \geq \frac{\epsilon}{\delta^4 \bar{\Delta}^2}\frac{\|\Delta z^{l-1}\|_2}{2\delta T} \geq \delta^{-2}\|\Delta z^{l-1}\|_2.$$

Hence, from (3.18) and the last two inequalities, we have

$$-m_{l-1}(\Delta z^{l-1}) \geq \frac{1}{2}\frac{\epsilon}{\delta^2}\|\Delta z^{l-1}\|_2. \tag{3.30}$$

By comparing (3.29) and (3.30), we have from (2.1) that

$$\begin{aligned} \rho_{l-1} &= \frac{f(z^{l-1}) - f\left(z^{l-1} + \widetilde{\Delta z}^{l-1}\right)}{-m_{l-1}(\Delta z^{l-1})} \\ &\geq 1 - \frac{\left|f(z^{l-1}) - f\left(z^{l-1} + \widetilde{\Delta z}^{l-1}\right) + m_{l-1}(\Delta z^{l-1})\right|}{-m_{l-1}(\Delta z^{l-1})} \\ &\geq 1 - \frac{1}{8} = \frac{7}{8}. \end{aligned}$$

Hence, by the workings of the algorithm, we have $\Delta_l \geq \Delta_{l-1}$. But since $N_{l-1} \leq N_l$, we have $N_{l-1}\Delta_{l-1} \leq N_l\Delta_l$, so that $\Delta_{l-1} < T/N_{l-1}$, which contradicts the definition of l as the first index that violates (3.17). We conclude that no such l exists, and hence that (3.17) holds. □

The following technical lemma, attributed to M. J. D. Powell, is proved in Yuan [17, Lemma 3.4]. We modify the statement slightly to begin the sequence at the index K rather than at 0.

LEMMA 3.5. *Suppose $\{\Delta_k\}$ and $\{N_k\}$ are two sequences such that $\Delta_k \geq T/N_k$ for all $k \geq K$, for some integer K and constant $T > 0$. Let $\mathcal{K} \subset \{K, K+1, K+2, \dots\}$ be defined such that*

$$\Delta_{k+1} \leq \tau_0 \Delta_k \quad \text{if } k \in \mathcal{K}, \tag{3.31a}$$

$$\Delta_{k+1} \leq \tau_1 \Delta_k \quad \text{if } k \notin \mathcal{K}, \tag{3.31b}$$

$$N_{k+1} \geq N_k \quad \text{for all } k \geq K, \tag{3.31c}$$

$$\sum_{k \in \mathcal{K}} \min(\Delta_k, 1/N_k) < \infty, \tag{3.31d}$$

where τ_0 and τ_1 are constants satisfying $0 < \tau_1 < 1 < \tau_0$. Then

$$\sum_{k=K}^{\infty} 1/N_k < \infty. \tag{3.32}$$

Our main global convergence result for this section is as follows.

THEOREM 3.6. *Suppose that Assumptions* 1, 2, *and* 3 *are satisfied and that the approximate Hessians H_k satisfy* (3.16)*; that is, $\|H_k\|_2 \leq \sigma_0 + k\sigma_1$ for some nonnegative constants σ_0 and σ_1. Then Algorithm FP-SQP either terminates at a KKT point or else has at least one limit point that is either a KKT point or else fails to satisfy the MFCQ conditions* (1.10a,b).

Proof. Consider first the case in which the algorithm terminates finitely at some iterate z^k at which $m_k(\Delta z^k) = 0$. Then $\Delta z = 0$ is a solution of the subproblem (1.2a,b), (1.3) at $z = z^k$ at which the trust-region bound is inactive. The KKT conditions for the subproblem at $\Delta z = 0$ correspond exactly to the KKT conditions (1.8a–c) for the original problem (1.1) at z^k.

In the alternate case, the algorithm generates an infinite sequence $\{z^k\}$. Suppose first that it is possible to choose $\epsilon > 0$ and K such that the conditions of Lemma 3.4 are satisfied. We apply Lemma 3.5, choosing $\mathcal{K}$ to be the subsequence of $\{K, K+1, K+2, \dots\}$ at which the trust-region radius is *not* reduced. We can then set $\tau_0 = 2$, $\tau_1 = 0.5$, and define N_k as in Lemma 3.4. At the iterates $k \in \mathcal{K}$, the algorithm takes a step, and we have $\rho_k \geq \eta$. By using (3.15) and (3.18), we then have

$$\begin{aligned} f(z^k) - f\left(z^k + \widetilde{\Delta z}^k\right) &\geq -\eta m_k(\Delta z^k) \\ &\geq \frac{1}{2}\eta\epsilon \min\left(\delta^{-1}, \delta^{-1}\Delta_k, \delta^{-4}\bar{\Delta}^{-2}\frac{\epsilon}{N_k}\right) \\ &\geq \frac{1}{2}\eta\epsilon \min(\delta^{-1}, \delta^{-4}\bar{\Delta}^{-2}\epsilon) \min\left(\Delta_k, \frac{1}{N_k}\right), \end{aligned}$$

where the final inequality follows from $N_k \geq 1$. By summing both sides of this inequality over $k \in \mathcal{K}$ and using the fact that $f(z^k)$ is bounded below (since f is continuous on the compact level set L_0), we have that condition (3.31d) is satisfied. The conclusion (3.32) then holds. However, since from (3.16) we have $N_k \leq 1 + \sigma_0 + \sigma_1 k$, (3.32) cannot hold, so we have a contradiction. We conclude therefore that it is *not* possible to choose $\epsilon > 0$ and K satisfying the conditions of Lemma 3.4; that is, there is a subsequence $\mathcal{J} \subset \{0, 1, 2, \dots\}$ such that

$$\lim_{k \in \mathcal{J}} C(z^k, 1) = 0.$$

Since the points z^k, $k \in \mathcal{J}$, all belong to the compact set L_0, we can identify a limit point $\bar{z}$ and assume, without loss of generality, that $\lim_{k \in \mathcal{J}} z^k = \bar{z}$. From the observation immediately following the statement of Lemma 3.1, we have either that MFCQ conditions (1.10) fail to hold at $\bar{z}$ or else that $\bar{z}$ satisfies both the MFCQ conditions and the KKT conditions (1.8). ∎

3.3. Result II: All limit points are KKT points. In this section, we replace the bound (3.16) on the Hessians H_k with a uniform bound

$$\|H_k\|_2 \leq \sigma, \tag{3.33}$$

for some constant σ, and obtain a stronger global convergence result, namely, that every limit point of the algorithm either fails to satisfy MFCQ or else is a KKT point.

As a preliminary to the main result of this section, we show that for any limit point $\bar{z}$ of Algorithm FP-SQP at which MFCQ but not KKT conditions are satisfied, there is a subsequence $\mathcal{K}$ with $z^k \to_{k\in\mathcal{K}} \bar{z}$ and $\Delta_k \to_{k\in\mathcal{K}} 0$.

LEMMA 3.7. *Suppose that Assumptions* 1, 2, *and* 3 *are satisfied and that the Hessians* H_k *satisfy the bound* (3.33) *for some* $\sigma > 0$. *Suppose that* $\bar{z}$ *is a limit point of the sequence* $\{z^k\}$ *such that the MFCQ condition* (1.10a,b) *holds but the KKT conditions* (1.8) *are not satisfied at* $\bar{z}$. *Then there exists an (infinite) subsequence* $\mathcal{K}$ *such that*

$$\lim_{k\in\mathcal{K}} z^k = \bar{z}, \tag{3.34}$$

and

$$\lim_{k\in\mathcal{K}} \Delta_k = 0. \tag{3.35}$$

Proof. Since $\bar{z} \in L_0$, we can define ϵ and R_2 as in Lemma 3.1. From this lemma, we have that $C(z,1) \geq \epsilon$ for all $z \in \mathcal{B}(\bar{z}, R_2) \cap \mathcal{F}$. Hence, for such z, we have from Lemma 3.3 that the solution Δz of the trust-region subproblem at (3.3) with $\Delta \in (0,1]$ satisfies

$$\begin{aligned} m(\Delta z) &\leq -\tfrac{1}{2} C(z,1) \min\left[\delta^{-1}, \delta^{-1}\Delta, (\delta^4 \bar{\Delta}^2 \|H\|_2)^{-1} C(z,1)\right] \\ &\leq -\tfrac{1}{2}\epsilon \min\left[\delta^{-1}, \delta^{-1}\Delta, (\delta^4\bar{\Delta}^2\sigma)^{-1}\epsilon\right], \end{aligned} \tag{3.36}$$

where we used the bound (3.33) to obtain the second inequality.

Because $\bar{z}$ is a limit point, we can certainly choose a subsequence $\mathcal{K}$ satisfying (3.34). By deleting the elements from $\mathcal{K}$ for which $z_k \notin \mathcal{B}(\bar{z}, R_2)$, we have from (3.36) that

$$m_k(\Delta z^k) \leq -\tfrac{1}{2}\epsilon \min\left[\delta^{-1}, \delta^{-1}\Delta_k, (\delta^4\bar{\Delta}^2\sigma)^{-1}\epsilon\right] \quad \text{for all } k \in \mathcal{K}. \tag{3.37}$$

We prove the result (3.35) by modifying $\mathcal{K}$ and taking further subsequences as necessary. Consider first the case in which $\{z^k\}_{k\in\mathcal{K}}$ takes on only a finite number of distinct values. We then must have that $z^k = \bar{z}$ for all $k \in \mathcal{K}$ sufficiently large. Now, remove from $\mathcal{K}$ all indices k for which $z^k \neq \bar{z}$. Suppose for contradiction that some subsequent iterate in the full sequence $\{z^k\}$ is different from $\bar{z}$. If $\bar{k} \geq k$ is some iterate such that

$$f\left(z^{\bar{k}}\right) < f(z^k) = f(\bar{z}),$$

we have by monotonicity of $\{f(z^l)\}$ (for the full sequence of function values) that

$$f(z^l) \leq f\left(z^{\bar{k}}\right) < f(\bar{z})$$

for *all* $l > \bar{k}$. Hence the function values in the tail of the full sequence are bounded away from $f(\bar{z})$, so it is not possible to choose a subsequence $\mathcal{K}$ with the property (3.34). Therefore, we have that $z^l = \bar{z}$ for *all* $l \geq k$ so that all steps generated by Algorithm FP-SQP after iteration k fail the acceptance condition. We then have

that

$$\Delta_{l+1} = \tfrac{1}{2}\|D_l \Delta z^l\|_p \leq \tfrac{1}{2}\Delta_l \text{ for all } l \geq k,$$

so that $\Delta_l \to 0$ as $l \to \infty$ (for the full sequence). Hence, in particular, (3.35) holds.

We consider now the second case, in which $\{z^k\}_{k\in\mathcal{K}}$ takes on an infinite number of distinct values. Without loss of generality, we can assume that *all* elements z^k, $k \in \mathcal{K}$, are distinct (by dropping the repeated elements if necessary). Moreover, we can assume that $z^{k+1} \neq z^k$ for all $k \in \mathcal{K}$ by replacing k if necessary with the largest index $\bar{k}$ such that $\bar{k} \geq k$ and $z^{\bar{k}} = z^k$. Thus, we have that the sufficient decrease condition $\rho_k \geq \eta$ is satisfied at all $k \in \mathcal{K}$. Therefore from (2.1) and (3.36), and the easily demonstrated fact that $f(z^l) \geq f(\bar{z})$ for *all* $l = 0, 1, 2, \ldots$, we have

$$\begin{aligned} f(z^k) - f(\bar{z}) &\geq f(z^k) - f(z^{k+1}) \\ &\geq -\eta m_k(\Delta z^k) \\ &\geq \tfrac{1}{2}\eta\epsilon \min\left[\delta^{-1}, \delta^{-1}\Delta_k, (\delta^4 \bar{\Delta}^2 \sigma)^{-1}\epsilon\right] \geq 0. \end{aligned}$$

Since $f(z^k) \to_{k\in\mathcal{K}} f(\bar{z})$, we have from this chain of inequalities that (3.35) is satisfied in this case too. Hence, we have demonstrated (3.35). □

We now prove the main global convergence result of this section.

THEOREM 3.8. *Suppose that Assumptions* 1, 2, *and* 3 *are satisfied and that the Hessian approximations H_k satisfy* (3.33). *Then all limit points of Algorithm FP-SQP either are KKT points or else fail to satisfy the MFCQ conditions* (1.10a,b).

Proof. Suppose for contradiction that $\bar{z}$ is a limit point at which (1.10a,b) hold but (1.8a–c) are not satisfied, and let R_2 and ϵ be defined as in the proof of Lemma 3.7. We invoke Lemma 3.7 to define the subsequence $\mathcal{K}$ with the properties (3.34) and (3.35). The inequality (3.37) also holds for the subsequence $\mathcal{K}$.

Let $\bar{\sigma}$ and γ be defined as in (3.19). We now define the constants $R > 0$ and $\Delta_\phi > 0$ such that the following conditions hold:

$$R \leq R_2, \tag{3.38a}$$

$$\gamma\phi(\Delta_\phi) \leq \frac{1}{16}\frac{\epsilon}{\delta^2}, \tag{3.38b}$$

$$\mathcal{B}(\bar{z}, R + \Delta_\phi) \cap \mathcal{F} \subset \mathcal{N}(L_0), \tag{3.38c}$$

$$\Delta_\phi \leq \Delta_{\text{def}}, \tag{3.38d}$$

where Δ_{def} is defined in Theorem 2.2. Note, in particular from the latter theorem, that $\widetilde{\Delta z}$ satisfying (1.4) and (1.5) exists whenever $\|D\Delta z\|_2 \leq \Delta_\phi$.

Given R and Δ_ϕ, we can now define $\tilde{\Delta} > 0$ small enough to satisfy the following properties:

$$\tilde{\Delta} \leq 1, \tag{3.39a}$$

$$\left(2\bar{\sigma} + \frac{1}{2}\sigma\right)\delta\tilde{\Delta} \leq \frac{1}{16}\frac{\epsilon}{\delta^2}, \tag{3.39b}$$

$$\tilde{\Delta} \leq \frac{2\Delta_\phi}{3\delta}, \tag{3.39c}$$

$$\tilde{\Delta} \leq \frac{\epsilon}{\delta^3 \bar{\Delta}^2 \sigma}, \tag{3.39d}$$

where $\bar{\Delta}$ is the overall upper bound on a trust-region radius. We then define $\hat{\epsilon} > 0$ as

follows:

$$\hat{\epsilon} = \frac{1}{2}\eta\epsilon \min\left(\delta^{-1}, \frac{1}{4}\frac{R}{\delta^2}, (\delta^4\bar{\Delta}^2\sigma)^{-1}\epsilon\right). \tag{3.40}$$

Finally, we define an index $q \in \mathcal{K}$ sufficient large that

$$\|z^q - \bar{z}\|_2 < R/2, \tag{3.41a}$$

$$f(z^q) - f(\bar{z}) \le \hat{\epsilon}/2. \tag{3.41b}$$

(Existence of such an index q follows immediately from $z^k \to_{\mathcal{K}} \bar{z}$.)

Consider the neighborhood

$$\mathrm{cl}\,(\mathcal{B}(z^q, R/2)) \cap \mathcal{F}, \tag{3.42}$$

which is contained in $\mathcal{B}(\bar{z}, R) \cap \mathcal{F}$ because of (3.41a). We consider two cases.

Case I. All remaining iterates $z^{q+1}, z^{q+2}, \ldots$ of the full sequence remain inside the neighborhood (3.42). If

$$\|D_k \Delta z^k\|_p \le \tilde{\Delta} \quad \text{for any } k = q,\, q+1,\, q+2, \ldots, \tag{3.43}$$

we have from (1.6) and (3.39c) that

$$\|\widetilde{\Delta z}^k\|_2 \le (3/2)\|\Delta z^k\|_2 \le (3/2)\delta\|D_k\Delta z^k\|_p \le (3/2)\delta\tilde{\Delta} \le \Delta_\phi. \tag{3.44}$$

We now show that whenever (3.43) occurs, the ratio ρ_k defined by (2.1) is at least 3/4, so that the trust-region radius Δ_{k+1} for the next iteration is no smaller than the one for this iteration, Δ_k. As in the proof of Lemma 3.4, the relation (3.22) holds, with z_θ^k satisfying

$$\mathrm{dist}(z_\theta^k, L_0) \le \tfrac{1}{2}\|\widetilde{\Delta z}^k\|_2 \le \tfrac{1}{2}\Delta_\phi.$$

Hence, from (3.19) and (3.38c), we have $\|\nabla^2 f(z_\theta^k)\|_2 \le \bar{\sigma}$. Similarly to (3.23), we have

$$\begin{aligned}
&\left|f(z^k) - f\left(z^k + \widetilde{\Delta z}^k\right) + m_k(\Delta z^k)\right| \\
&\le \left(\|\nabla f(z^k)\|_2 + \|\nabla^2 f(z_\theta^k)\|_2\|\Delta z^k\|_2\right)\|\Delta z^k - \widetilde{\Delta z}^k\|_2 \\
&\qquad + \tfrac{1}{2}\|\nabla^2 f(z_\theta^k)\|_2\|\widetilde{\Delta z}^k - \Delta z^k\|_2^2 + \tfrac{1}{2}(\|\nabla^2 f(z_\theta^k)\|_2 + \|H_k\|_2)\|\Delta z^k\|_2^2 \\
&\le (\gamma + \bar{\sigma}\delta\tilde{\Delta})\phi(\|\Delta z^k\|_2)\|\Delta z^k\|_2 + \tfrac{1}{2}\bar{\sigma}\phi(\|\Delta z^k\|_2)^2\|\Delta z^k\|_2^2 + \tfrac{1}{2}(\bar{\sigma} + \sigma)\|\Delta z^k\|_2^2,
\end{aligned}$$

where we used (3.19) and $\|\Delta z^k\|_2 \le \delta\tilde{\Delta}$ from (3.44) in deriving the second inequality. Now using (3.44) again, together with monotonicity of ϕ, $\phi(\cdot) \le 1/2$, (3.38b), and

(3.39b), we have

$$\begin{aligned}
&\left|f(z^k) - f\left(z^k + \widetilde{\Delta z}^k\right) + m_k(\Delta z^k)\right| \\
&\leq (\gamma + \bar{\sigma}\delta\tilde{\Delta})\phi(\Delta_\phi)\|\Delta z^k\|_2 + \left[\frac{1}{2}\bar{\sigma}\phi(\Delta_\phi)^2\delta\tilde{\Delta} + \frac{1}{2}(\bar{\sigma}+\sigma)\delta\tilde{\Delta}\right]\|\Delta z^k\|_2 \\
&\leq \left[\gamma\phi(\Delta_\phi) + \left(\bar{\sigma}\delta\tilde{\Delta} + \frac{1}{2}\bar{\sigma}\delta\tilde{\Delta} + \frac{1}{2}(\bar{\sigma}+\sigma)\delta\tilde{\Delta}\right)\right]\|\Delta z^k\|_2 \\
&= \left[\gamma\phi(\Delta_\phi) + \left(2\bar{\sigma} + \frac{1}{2}\sigma\right)\delta\tilde{\Delta}\right]\|\Delta z^k\|_2 \\
&\leq \left(\frac{1}{16}\frac{\epsilon}{\delta^2} + \frac{1}{16}\frac{\epsilon}{\delta^2}\right)\|\Delta z^k\|_2 = \frac{1}{8}\frac{\epsilon}{\delta^2}\|\Delta z^k\|_2. \qquad (3.45)
\end{aligned}$$

Meanwhile, from (3.36) and, since $z^k \in \mathcal{B}(\bar{z}, R) \cap \mathcal{F}$ where $R \leq R_2$, we have

$$-m_k(\Delta z^k) \geq \tfrac{1}{2}\epsilon \min(\delta^{-1}, \delta^{-1}\Delta_k, (\delta^4\bar{\Delta}^2\sigma)^{-1}\epsilon). \qquad (3.46)$$

Now from Assumption 1, we have

$$\Delta_k \geq \|D_k\Delta z^k\|_p \geq \delta^{-1}\|\Delta z^k\|_2,$$

while from (3.39a) and (3.43), we have

$$1 \geq \tilde{\Delta} \geq \|D_k\Delta z^k\|_p \geq \delta^{-1}\|\Delta z^k\|_2.$$

From (3.39d) and Assumption 1, we have

$$\epsilon \geq \delta^3\bar{\Delta}^2\sigma\tilde{\Delta} \geq \delta^3\bar{\Delta}^2\sigma\|D_k\Delta z^k\|_p \geq \delta^2\bar{\Delta}^2\sigma\|\Delta z^k\|_2.$$

By substituting these last three expressions into (3.46), we obtain

$$-m_k(\Delta z^k) \geq \frac{1}{2}\frac{\epsilon}{\delta^2}\|\Delta z^k\|_2. \qquad (3.47)$$

We then have from (2.1), and using (3.45) and (3.47), that

$$\begin{aligned}
\rho_k &= \frac{f(z^k) - f\left(z^k + \widetilde{\Delta z}^k\right)}{-m_k(\Delta z^k)} \\
&\geq 1 - \frac{\left|f(z^k) - f\left(z^k + \widetilde{\Delta z}^k\right) + m_k(\Delta z^k)\right|}{-m_k(\Delta z^k)} \\
&\geq \frac{3}{4}.
\end{aligned}$$

It follows that the algorithm sets

$$\Delta_{k+1} \geq \Delta_k \qquad (3.48)$$

for all k satisfying (3.43). For $k = q, q+1, q+2, \ldots$ *not* satisfying (3.43), Algorithm FP-SQP may reduce the trust-region radius to

$$\Delta_{k+1} = (1/2)\|D_k\Delta z^k\|_p \geq (1/2)\tilde{\Delta}. \qquad (3.49)$$

By considering both cases, we conclude that

$$\Delta_k \geq \min(\Delta_q, (1/2)\tilde{\Delta}) \text{ for all } k = q, q+1, q+2, \ldots,$$

which contradicts (3.35). Hence, Case I cannot occur.

We now consider the alternative case.

Case II. Some subsequent iterate $z^{q+1}, z^{q+2}, \ldots$ leaves the neighborhood (3.42). If z^l is the first iterate outside this neighborhood, note that all iterates z^k, $k = q, q+1, q+2, \ldots, l-1$ lie inside the set $\mathcal{B}(\bar{z}, R) \cap \mathcal{F}$, within which (3.36) applies. By summing over the "successful" iterates in this span, we have the following:

$$\begin{aligned}
& f(z^q) - f(z^l) \\
&= \sum_{\substack{k=q \\ z^k \neq z^{k+1}}}^{l-1} f(z^k) - f(z^{k+1}) \\
&\geq \sum_{\substack{k=q \\ z^k \neq z^{k+1}}}^{l-1} -\eta m_k(\Delta z^k) \qquad \text{by (2.1) and Algorithm FP-SQP} \\
&\geq \eta \sum_{\substack{k=q \\ z^k \neq z^{k+1}}}^{l-1} \frac{1}{2}\epsilon \min\left[\delta^{-1}, \delta^{-1}\Delta_k, (\delta^4 \bar{\Delta}^2 \sigma)^{-1}\epsilon\right] \qquad \text{by (3.36)} \\
&\geq \frac{1}{2}\eta\epsilon \min\left[\delta^{-1}, \delta^{-1} \sum_{\substack{k=q \\ z^k \neq z^{k+1}}}^{l-1} \Delta_k, (\delta^4 \bar{\Delta}^2 \sigma)^{-1}\epsilon\right]. \qquad (3.50)
\end{aligned}$$

We have from Assumption 1 and (1.6) that

$$\Delta_k \geq \|D_k \Delta z^k\|_p \geq \delta^{-1}\|\Delta z^k\|_2 \geq \tfrac{1}{2}\delta^{-1}\|\widetilde{\Delta z}^k\|_2,$$

so that (3.50) becomes

$$f(z^q) - f(z^l) \geq \frac{1}{2}\eta\epsilon \min\left[\delta^{-1}, \sum_{\substack{k=q \\ z^k \neq z^{k+1}}}^{l-1} \frac{1}{2}\delta^{-2}\|\widetilde{\Delta z}^k\|_2, (\delta^4 \bar{\Delta}^2 \sigma)^{-1}\epsilon\right]. \qquad (3.51)$$

However, because z^l lies outside the neighborhood (3.42) we have that

$$R/2 \leq \|z^q - z^l\|_2 \leq \sum_{\substack{k=q \\ z^k \neq z^{k+1}}}^{l-1} \|\widetilde{\Delta z}^k\|_2,$$

so that (3.51) becomes

$$f(z^q) - f(z^l) \geq \tfrac{1}{2}\eta\epsilon \min\left[\delta^{-1}, \tfrac{1}{4}\delta^{-2}R, (\delta^4 \bar{\Delta}^2 \sigma)^{-1}\epsilon\right]. \qquad (3.52)$$

By using this estimate together with the definition of $\hat{\epsilon}$ in (3.40), we have

$$f(z^q) - f(z^l) \geq \hat{\epsilon}.$$

But since $f(z^l) \geq f(\bar{z})$ (since $\bar{z}$ is a limit point of the full sequence), this inequality contradicts (3.41b). Hence, Case II cannot occur either, and the proof is complete. ∎

4. Local convergence. We now examine local convergence behavior of the algorithm to a point z^* satisfying second-order sufficient conditions for optimality, under the assumption that $z^k \to z^*$. We do not attempt to obtain the most general possible superlinear convergence result, but rather make the kind of assumptions that are typically made in the local convergence analysis of SQP methods, in which second derivatives of the objective and constraint functions are available. We also make additional assumptions on the feasibility perturbation process that is used to recover $\widetilde{\Delta z}^k$ from Δz^k. Ultimately, we show that Algorithm FP-SQP converges Q-superlinearly.

We assume a priori that z^* satisfies the KKT conditions and define the active set $\mathcal{A}^*$ as follows:

$$\mathcal{A}^* \stackrel{\text{def}}{=} \mathcal{A}(z^*), \tag{4.1}$$

where $\mathcal{A}(\cdot)$ is defined in (1.9). In this section, we use the following subvector notation:

$$d_{\mathcal{I}}(z) \stackrel{\text{def}}{=} [d_i(z)]_{i\in\mathcal{I}}, \quad \text{where } \mathcal{I} \subset \{1, 2, \ldots, r\}.$$

Assumption 4.

(a) *The functions f, c, and d are twice continuously differentiable in a neighborhood of z^*.*

(b) *The LICQ* (1.11) *is satisfied at z^*.*

(c) *Strict complementarity holds; that is, for the (unique) multipliers (μ^*, λ^*) satisfying the KKT conditions* (1.8a–c) *at $z = z^*$, we have $\lambda_i^* > 0$ for all $i \in \mathcal{A}^*$.*

(d) *Second-order sufficient conditions are satisfied at z^*; that is, there is $\alpha > 0$ such that*

$$v^T \nabla^2_{zz} \mathcal{L}(z^*, \mu^*, \lambda^*) v \geq \alpha \|v\|^2 \text{ for all } v \text{ such that}$$
$$\nabla c(z^*)^T v = 0, \quad \nabla d_{\mathcal{A}^*}(z^*)^T v = 0,$$

where the Lagrangian function $\mathcal{L}$ is defined in (1.7).

Besides these additional assumptions on the nature of the limit point z^*, we make additional assumptions on the algorithm itself. As mentioned above, we start by assuming that $z^k \to z^*$. We further assume that estimates $\mathcal{W}_k$ of the active set $\mathcal{A}^*$ and estimates (μ^k, λ^k) of the optimal Lagrange multipliers (μ^*, λ^*) are calculated at each iteration k and that these estimates are asymptotically exact. It is known (see, for example, Facchinei, Fischer, and Kanzow [6]) that an asymptotically exact estimate $\mathcal{W}_k$ of $\mathcal{A}^*$ is available, given that $(z^k, \mu^k, \lambda^k) \to (z^*, \mu^*, \lambda^*)$, under weaker conditions than assumed here. On the other hand, it is also known that given an asymptotically exact $\mathcal{W}_k$, we can use a least-squares procedure to compute an asymptotically exact estimate (μ^k, λ^k) of (μ^*, λ^*). However, the *simultaneous* estimation of

$\mathcal{W}_k$ and (μ^k, λ^k) is less straightforward. We anticipate, however, that a procedure that works well in practice would be relatively easy to implement, especially under the LICQ and strict complementarity assumptions. Given an initial guess of $\mathcal{W}_k$, such a procedure would alternate between a least-squares estimate of (μ^k, λ^k) and an active-set identification procedure, like those in [6], until the estimate of $\mathcal{W}_k$ settles down. We note that the multipliers for the linearized constraints in the subproblem (1.2a,b), (1.3) (denoted in the analysis below by $\bar{\mu}^k$ and $\bar{\lambda}^k$) do not necessarily satisfy the asymptotic exactness condition, unless it is known a priori that the trust region is inactive for all k sufficiently large. Fletcher and Sainz de la Maza [7] have analyzed the behavior of these multipliers in the context of a sequential linear programming algorithm and show that, under certain assumptions, (μ^*, λ^*) is a limit point of the sequence $\{(\bar{\mu}^k, \bar{\lambda}^k)\}$.

We summarize the algorithmic assumptions as follows.

ASSUMPTION 5.

(a) $z^k \to z^*$.

(b) $\mathcal{W}_k = \mathcal{A}^*$ *for all* k *sufficiently large, where* $\mathcal{W}_k$ *is the estimate of the optimal active set.*

(c) $(\mu^k, \lambda^k) \to (\mu^*, \lambda^*)$.

(d) *In addition to* (1.4) *and* (1.5)*, Algorithm FP-SQP requires the perturbed step* $\widetilde{\Delta z}^k$ *to satisfy*

$$d_i(z^k + \widetilde{\Delta z}^k) = d_i(z^k) + \nabla d_i(z^k)^T \Delta z^k \quad \text{for all } i \in \mathcal{W}_k \tag{4.2}$$

and

$$\|\Delta z^k - \widetilde{\Delta z}^k\| = O(\|\Delta z^k\|^2). \tag{4.3}$$

We note the following about Assumption 5.

- For iterations k at which a step is taken (the "successful" iterations), we have that $\widetilde{\Delta z}^k = z^{k+1} - z^k$, which approaches zero by Assumption 5(a). Hence, by (1.6), and defining $\mathcal{K}$ to be the subsequence of successful iterations, we have that

$$\lim_{k \in \mathcal{K}} \|\Delta z^k\| = \lim_{k \in \mathcal{K}} \left\|\widetilde{\Delta z}^k\right\| = 0. \tag{4.4}$$

- The condition (4.2) is an explicit form of "second-order correction," a family of techniques that are often needed to ensure fast local convergence of SQP algorithms.
- It follows from (1.6) and (4.3) that

$$\left\|\Delta z^k - \widetilde{\Delta z}^k\right\| = O\left(\left\|\widetilde{\Delta z}^k\right\|^2\right). \tag{4.5}$$

We start with a technical result to show that the various requirements on the perturbed step $\widetilde{\Delta z}^k$ are consistent. Note that this result is merely an existence result. It is not intended to show a practical way of obtaining $\widetilde{\Delta z}^k$. There may be other (less expensive, problem-dependent) ways to calculate the perturbed step that result in satisfaction of all the required conditions.

LEMMA 4.1. *Suppose that Assumption* 4 *and Assumptions* 5(a),(b) *hold. Then for all sufficiently large* k*, it is possible to choose the trust-region radius* Δ_k *small enough that there exists* $\widetilde{\Delta z}^k$ *satisfying* (1.4), (1.5), (4.2)*, and* (4.3).

Proof. Assume first that k is chosen large enough that $\mathcal{W}_k = \mathcal{A}^*$. We prove the result constructively, generating $\widetilde{\Delta z}^k$ as the solution of the following problem:

$$\min_w \tfrac{1}{2}\|w - \Delta z^k\|_2^2 \text{ subject to} \tag{4.6a}$$

$$c(z^k + w) = 0, \tag{4.6b}$$

$$d_i(z^k + w) = d_i(z^k) + \nabla d_i(z^k)^T \Delta z^k \text{ for all } i \in \mathcal{W}_k. \tag{4.6c}$$

When the right-hand sides of (4.6b), (4.6c) are replaced with $c(z^k + \Delta z^k)$ and $d_i(z^k + \Delta z^k)$, respectively, the solution is $w = \Delta z^k$. By the smoothness assumptions on c and d, these modified right-hand sides represent only an $O(\|\Delta z^k\|^2)$ perturbation of the right-hand sides in (4.6b), (4.6c). Note that the Jacobian of the constraints (4.6b), (4.6c) has full row rank at $z^k + \Delta z^k$ because of Assumptions 4(b) and 5(a). Hence, the Jacobian matrix of the KKT conditions for problem (4.6) (which is a "square" system of nonlinear equations) is nonsingular at $z^k + \Delta z^k$, and a straightforward application of the implicit function theorem to this system yields that the solution $w = \widetilde{\Delta z}^k$ of (4.6) satisfies property (4.3) for all k sufficiently large. Condition (4.2) is an immediate consequence of (4.6c).

By decreasing Δ_k if necessary and using $\|\Delta z_k\| \le \delta \Delta_k$, we can derive (1.5) as a consequence of (4.3).

Because of (1.2b), we have

$$d_i\left(z^k + \widetilde{\Delta z}^k\right) = d_i(z^k) + \nabla d_i(z^k)^T \Delta z^k \le 0 \text{ for all } i \in \mathcal{A}^*,$$

while for $i \notin \mathcal{A}^*$ we have from $d_i(z^*) < 0$ and Assumption 5(a) that

$$d_i(z^k + \widetilde{\Delta z}^k) = d_i(z^k) + O(\Delta_k) \le (1/2) d_i(z^*) < 0$$

for all k sufficiently large and Δ_k sufficiently small. For the equality constraints, we have immediately from (4.6b) that $c(z^k + \widetilde{\Delta z}^k) = 0$. Hence $z^k + \widetilde{\Delta z}^k \in \mathcal{F}$, so condition (1.4) is also satisfied. ∎

We assume that the Hessian matrix H_k in the subproblem (1.2a,b), (1.3) at $z = z^k$ is the Hessian of the Lagrangian $\mathcal{L}$ evaluated at this point, with appropriate estimates of the multipliers μ^k and λ^k; that is,

$$H_k = \nabla^2_{zz} \mathcal{L}(z^k, \mu^k, \lambda^k) = \nabla^2 f(z^k) + \sum_{i=1}^m \mu_i^k \nabla^2 c_i(z^k) + \sum_{i=1}^r \lambda_i^k \nabla^2 d_i(z^k). \tag{4.7}$$

We show now that with this choice of H_k, the ratio ρ_k of actual to predicted decrease is close to 1 when k is sufficiently large and the steps Δz^k and $\widetilde{\Delta z}^k$ are sufficiently small. We prove the result specifically for the Euclidean-norm trust region; a minor generalization yields the proof for general $p \in [1, \infty]$.

LEMMA 4.2. *Suppose that $p = 2$ in (1.3), that Assumptions 1, 4, and 5 hold, and that H_k is defined by (4.7). Then there is a threshold value Δ_τ and an index K_1 such that if $k \ge K_1$ and $\|D_k \Delta z^k\|_2 \le \Delta_\tau$, we have $\rho_k \ge 1/2$, where ρ_k is defined by (2.1).*

Proof. Note first that we can use Δ_τ to control the size of both Δz^k and $\widetilde{\Delta z}^k$, since from Assumption 1 we have $\|\Delta z^k\| \le \delta \|D_k \Delta z^k\|_2 \le \delta \Delta_\tau$, while from (1.6) we have $\|\widetilde{\Delta z}^k\| \le (3/2)\|\Delta z^k\|$.

From (2.1) we have

$$\rho_k = 1 + \frac{f(z^k) - f\left(z^k + \widetilde{\Delta z}^k\right) + m_k(\Delta z^k)}{-m_k(\Delta z^k)}. \tag{4.8}$$

We prove the result by showing that the numerator of the final term in this expression is $o(\|\widetilde{\Delta z}^k\|^2)$, while the denominator is $\Omega(\|\Delta z^k\|^2)$.

We assume initially that K_1 is large enough that $\mathcal{W}_k = \mathcal{A}^*$ for all $k \geq K_1$. We work first with the numerator in (4.8). By elementary manipulation, using Taylor's theorem and the definition of $m_k(\cdot)$, we have for some $\theta^f \in (0,1)$ that

(4.9)

$$\begin{aligned}
&f(z^k) - f\left(z^k + \widetilde{\Delta z}^k\right) + m_k(\Delta z^k) \\
&= -\nabla f(z^k)^T \widetilde{\Delta z}^k - \tfrac{1}{2}\left(\widetilde{\Delta z}^k\right)^T \nabla^2 f\left(z^k + \theta^f \widetilde{\Delta z}^k\right) \widetilde{\Delta z}^k + \nabla f(z^k)^T \Delta z^k + \tfrac{1}{2}(\Delta z^k)^T H_k \Delta z^k \\
&= \left(\nabla f(z^k) + H_k \widetilde{\Delta z}^k\right)^T \left(\Delta z^k - \widetilde{\Delta z}^k\right) + \tfrac{1}{2}\left(\widetilde{\Delta z}^k\right)^T \left(H_k - \nabla^2 f\left(z^k + \theta^f \widetilde{\Delta z}^k\right)\right) \widetilde{\Delta z}^k \\
&\qquad + O\left(\left\|\Delta z^k - \widetilde{\Delta z}^k\right\|^2\right) \\
&= \nabla f(z^k)^T \left(\Delta z^k - \widetilde{\Delta z}^k\right) + \tfrac{1}{2}\left(\widetilde{\Delta z}^k\right)^T \left(H_k - \nabla^2 f(z^k)\right) \widetilde{\Delta z}^k + o\left(\left\|\widetilde{\Delta z}^k\right\|^2\right),
\end{aligned}$$

where we used (4.5), boundedness of H_k, and continuity of $\nabla^2 f$ to derive the final equality. Now from (1.2b) and continuity of $\nabla^2 c_i$ for all $i = 1, 2, \ldots, m$ (Assumption 4(a)), we have

$$\begin{aligned}
0 &= c_i\left(z^k + \widetilde{\Delta z}^k\right) \\
&= c_i(z^k) + \nabla c_i(z^k)^T \widetilde{\Delta z}^k + \tfrac{1}{2}\left(\widetilde{\Delta z}^k\right)^T \nabla^2 c_i(z^k) \widetilde{\Delta z}^k + o\left(\left\|\widetilde{\Delta z}^k\right\|^2\right) \\
&= \nabla c_i(z^k)^T \left(\widetilde{\Delta z}^k - \Delta z^k\right) + \tfrac{1}{2}\left(\widetilde{\Delta z}^k\right)^T \nabla^2 c_i(z^k) \widetilde{\Delta z}^k + o\left(\left\|\widetilde{\Delta z}^k\right\|^2\right).
\end{aligned} \tag{4.10}$$

From (4.2), we have for all $i \in \mathcal{A}^*$ that

$$\begin{aligned}
0 &= d_i\left(z^k + \widetilde{\Delta z}^k\right) - d_i(z^k) - \nabla d_i(z^k)^T \Delta z^k \\
&= \nabla d_i(z^k)^T \left(\widetilde{\Delta z}^k - \Delta z^k\right) + \tfrac{1}{2}\left(\widetilde{\Delta z}^k\right)^T \nabla^2 d_i(z^k) \widetilde{\Delta z}^k + o\left(\left\|\widetilde{\Delta z}^k\right\|^2\right).
\end{aligned} \tag{4.11}$$

For $i \notin \mathcal{A}^*$, we have from $\lambda_i^k \to \lambda_i^* = 0$ and (4.5) that

$$\lambda_i^k \nabla d_i(z^k)^T \left(\widetilde{\Delta z}^k - \Delta z^k\right) + \tfrac{1}{2}\lambda_i^k \left(\widetilde{\Delta z}^k\right)^T \nabla^2 d_i(z^k) \widetilde{\Delta z}^k = o\left(\left\|\widetilde{\Delta z}^k\right\|^2\right). \tag{4.12}$$

We now multiply (4.10) and (4.11) by their corresponding Lagrange multipliers (μ_i^k

and λ_i^k, respectively), and subtract them, together with (4.12), from (4.9) to obtain

(4.13)
$$\begin{aligned}
&f(z^k) - f\left(z^k + \widetilde{\Delta z}^k\right) + m_k(\Delta z^k) \\
&= \left(\nabla f(z^k) + \nabla c(z^k)\mu^k + \nabla d(z^k)\lambda^k\right)^T \left(\Delta z^k - \widetilde{\Delta z}^k\right) \\
&\qquad + \tfrac{1}{2}\left(\widetilde{\Delta z}^k\right)^T \left[H_k - \nabla^2 f(z^k) - \sum_{i=1}^m \mu_i^k \nabla^2 c_i(z^k) - \sum_{i=1}^r \lambda_i^k \nabla^2 d_i(z^k)\right] \widetilde{\Delta z}^k \\
&\qquad + o\left(\left\|\widetilde{\Delta z}^k\right\|^2\right) \\
&= O(\|(z^k,\mu^k,\lambda^k) - (z^*,\mu^*,\lambda^*)\|)\left\|\Delta z^k - \widetilde{\Delta z}^k\right\| + o\left(\left\|\widetilde{\Delta z}^k\right\|^2\right) \\
&= o\left(\left\|\widetilde{\Delta z}^k\right\|^2\right),
\end{aligned}$$

where we used the KKT condition (1.8a) at $(z,\mu,\lambda) = (z^*,\mu^*,\lambda^*)$ and the definition (4.7) to derive the second equality, and Assumption 5(a),(c), together with (4.5), to derive the third equality. Hence, we have shown that the numerator of the last term in (4.8) is $o(\|\widetilde{\Delta z}^k\|^2)$.

In the remainder of the proof we use the following shorthand notation for the Hessian of the Lagrangian:

$$(\nabla^2_{zz}\mathcal{L})_k = \nabla^2_{zz}\mathcal{L}(z^k,\mu^k,\lambda^k); \quad (\nabla^2_{zz}\mathcal{L})_* = \nabla^2_{zz}\mathcal{L}(z^*,\mu^*,\lambda^*). \tag{4.14}$$

Given $p = 2$ in (1.3), we see that the KKT conditions for Δz^k to be a solution of (1.2a,b), (1.3) at $z = z^k$ are that there exist Lagrange multipliers $\bar\mu^k$, $\bar\lambda^k$, and γ_k such that

$$\nabla f(z^k) + (\nabla^2_{zz}\mathcal{L})_k \Delta z^k + \nabla c(z^k)\bar\mu^k + \nabla d(z^k)\bar\lambda^k + \gamma_k D_k^T D_k \Delta z^k = 0, \tag{4.15a}$$
$$c(z^k) + \nabla c(z^k)^T \Delta z^k = 0, \tag{4.15b}$$
$$0 \ge d(z^k) + \nabla d(z^k)^T \Delta z^k \quad \perp \quad \bar\lambda^k \ge 0, \tag{4.15c}$$
$$0 \ge \|D_k \Delta z^k\|_2^2 - \Delta_k^2 \quad \perp \quad \gamma_k \ge 0, \tag{4.15d}$$

where γ_k is the Lagrange multiplier for the trust-region constraint $\|D_k\Delta z^k\|_2^2 \le \Delta_k^2$. From (4.15b), (4.15c), and feasibility of z^k, we have

$$(\bar\mu^k)^T \nabla c(z^k)^T \Delta z^k = -(\bar\mu^k)^T c(z^k) = 0, \tag{4.16a}$$
$$(\bar\lambda^k)^T \nabla d(z^k)^T \Delta z^k = -(\bar\lambda^k)^T d(z^k) \ge 0. \tag{4.16b}$$

We turn now to the denominator in (4.8) and show that it has size $\Omega(\|\Delta z^k\|^2)$ for all k sufficiently large. From the definition of $m_k(\cdot)$, (4.7), and (4.14), we have

$$\begin{aligned}
-m_k(\Delta z^k) &= -\nabla f(z^k)^T \Delta z^k - \tfrac{1}{2}(\Delta z^k)^T (\nabla^2_{zz}\mathcal{L})_k \Delta z^k \\
&= -(\Delta z^k)^T \left(\nabla f(z^k) + (\nabla^2_{zz}\mathcal{L})_k \Delta z^k\right) + \tfrac{1}{2}(\Delta z^k)^T (\nabla^2_{zz}\mathcal{L})_k \Delta z^k.
\end{aligned}$$

By substituting from (4.15a), then using (4.15b) and (4.16), we obtain

$$\begin{aligned}
-m_k(\Delta z^k) &= (\Delta z^k)^T \left(\nabla c(z^k)\bar\mu^k + \nabla d(z^k)\bar\lambda^k + \gamma_k D_k^T D_k \Delta z^k\right) \\
&\qquad + \tfrac{1}{2}(\Delta z^k)^T (\nabla^2_{zz}\mathcal{L})_k \Delta z^k \\
&= -d(z^k)^T \bar\lambda^k + \gamma_k \|D_k \Delta z^k\|_2^2 + \tfrac{1}{2}(\Delta z^k)^T (\nabla^2_{zz}\mathcal{L})_k \Delta z^k.
\end{aligned}$$

By using Assumption 1, we obtain

$$-m_k(\Delta z^k) \geq -d(z^k)^T \bar{\lambda}^k + \gamma_k \delta^{-2} \|\Delta z^k\|_2^2 + \tfrac{1}{2}(\Delta z^k)^T (\nabla^2_{zz}\mathcal{L})_k \Delta z^k. \tag{4.17}$$

We now define the constant $\bar{\gamma}$ as follows:

$$\bar{\gamma} \stackrel{\text{def}}{=} \max\left(2\delta^2 \left\|(\nabla^2_{zz}\mathcal{L})_*\right\|_2, 1\right). \tag{4.18}$$

By increasing K_1 if necessary, we have by smoothness of $\mathcal{L}$ together with Assumption 5(a),(c) that

$$\left\|(\nabla^2_{zz}\mathcal{L})_k\right\|_2 \leq 2\left\|(\nabla^2_{zz}\mathcal{L})_*\right\|_2 \leq \delta^{-2}\bar{\gamma} \quad \text{for all } k \geq K_1. \tag{4.19}$$

We derive the estimate for $-m_k(\Delta z^k)$ from (4.17) by considering two cases. In the first case, we assume that $\gamma_k \geq \bar{\gamma}$. We then have from (4.17), using (4.16b), that the following bound holds for all $k \geq K_1$:

$$\begin{aligned} -m_k(\Delta z^k) &\geq \gamma_k \delta^{-2}\|\Delta z^k\|_2^2 + \tfrac{1}{2}(\Delta z^k)^T(\nabla^2_{zz}\mathcal{L})_k \Delta z^k \\ &\geq \bar{\gamma}\delta^{-2}\|\Delta z^k\|_2^2 - \tfrac{1}{2}\|\Delta z^k\|^2 \left\|(\nabla^2_{zz}\mathcal{L})_k\right\|_2 \\ &\geq \tfrac{1}{2}\bar{\gamma}\delta^{-2}\|\Delta z^k\|_2^2, \end{aligned} \tag{4.20}$$

so we see that the estimate $-m_k(\Delta z^k) = \Omega(\|\Delta z^k\|^2)$ is satisfied in this case.

In the second case of $\gamma_k \leq \bar{\gamma}$, a little more analysis is needed. We show first that

$$\lim_{k\to\infty, \gamma_k \leq \bar{\gamma}} (\bar{\mu}^k, \bar{\lambda}^k) = (\mu^*, \lambda^*).$$

By choosing Δ_τ small enough and increasing K_1 if necessary, we have, when $\|D_k \Delta z^k\| \leq \Delta_\tau$ and $k \geq K_1$, that

$$\begin{aligned} &i \notin \mathcal{W}_k = \mathcal{A}^* \\ &\Rightarrow d_i(z^k) + \nabla d_i(z^k)^T \Delta z^k = d_i(z^*) + O(\|z^k - z^*\|) + O(\|\Delta z^k\|) \leq (1/2) d_i(z^*) < 0, \end{aligned}$$

where we used Assumption 5(a) for the first equality. Hence, from (4.15c), we have $\bar{\lambda}_i^k = 0$ for all $i \notin \mathcal{A}^*$. By rearranging (4.15a), we therefore have

$$\nabla c(z^k)\bar{\mu}^k + \nabla d_{\mathcal{A}^*}(z^k)\bar{\lambda}^k_{\mathcal{A}^*} = -\nabla f(z^k) - (\nabla^2_{zz}\mathcal{L})_k \Delta z^k - \gamma_k D_k^T D_k \Delta z^k.$$

By comparing this expression with the KKT condition for z^*, namely,

$$\nabla c(z^*)\mu^* + \nabla d_{\mathcal{A}^*}(z^*)\lambda^*_{\mathcal{A}^*} = -\nabla f(z^*),$$

and using the LICQ (Assumption 4(b)), Assumption 1, and $\gamma_k \leq \bar{\gamma}$, we obtain

$$\|(\bar{\mu}^k, \bar{\lambda}^k_{\mathcal{A}^*}) - (\mu^*, \lambda^*_{\mathcal{A}^*})\| = O(\|z^k - z^*\|) + O(\|\Delta z^k\|) \to 0.$$

Hence, by strict complementarity (Assumption 4(c)), and by increasing K_1 again if necessary, we can identify a constant $\bar{\lambda}_{\min} > 0$ such that

$$\bar{\lambda}_i^k \geq \bar{\lambda}_{\min} \text{ for all } i \in \mathcal{A}^*, \text{ for all } k \geq K_1 \text{ with } \gamma_k \leq \bar{\gamma}. \tag{4.21}$$

Therefore, by the complementarity condition (4.15c), we have that

$$\nabla d_{\mathcal{A}^*}(z^k)^T \Delta z^k = -d_{\mathcal{A}^*}(z^k).$$

Using this expression together with (4.15b), we deduce that

$$\begin{aligned}\begin{bmatrix} \nabla c(z^*)^T \\ \nabla d_{\mathcal{A}^*}(z^*)^T \end{bmatrix} \Delta z^k &= \begin{bmatrix} (\nabla c(z^*) - \nabla c(z^k))^T \Delta z^k \\ -d_{\mathcal{A}^*}(z^k) + (\nabla d_{\mathcal{A}^*}(z^*) - \nabla d_{\mathcal{A}^*}(z^k))^T \Delta z^k \end{bmatrix} \\ &= O(\|d_{\mathcal{A}^*}(z^k)\|) + O(\|z^k - z^*\| \|\Delta z^k\|). \end{aligned} \tag{4.22}$$

By full row rank of the coefficient matrix on the left-hand side of (4.22), we have that there exists a vector s^k with

$$\begin{bmatrix} \nabla c(z^*)^T \\ \nabla d_{\mathcal{A}^*}(z^*)^T \end{bmatrix} s^k = \begin{bmatrix} \nabla c(z^*)^T \\ \nabla d_{\mathcal{A}^*}(z^*)^T \end{bmatrix} \Delta z^k, \tag{4.23a}$$

$$\|s^k\| = O(\|d_{\mathcal{A}^*}(z^k)\|) + O(\|z^k - z^*\| \|\Delta z^k\|). \tag{4.23b}$$

Since the vector $\Delta z^k - s^k$ satisfies the conditions on v in the second-order sufficient conditions (Assumptions 4(d)), we have

$$(\Delta z^k - s^k)^T (\nabla^2_{zz} \mathcal{L})_* (\Delta z^k - s^k) \geq \alpha \|\Delta z^k - s^k\|_2^2,$$

so that by increasing K_1 again if necessary, we have by Assumption 5(a),(c) that

$$(\Delta z^k - s^k)^T (\nabla^2_{zz} \mathcal{L})_k (\Delta z^k - s^k) \geq \tfrac{1}{2} \alpha \|\Delta z^k - s^k\|_2^2 \quad \text{for all } k \geq K_1.$$

By using this inequality together with (4.23b) and Assumption 5(a), we obtain (again increasing K_1 if needed) that

$$\begin{aligned} &(\Delta z^k)^T (\nabla^2_{zz} \mathcal{L})_k \Delta z^k \\ &= (\Delta z^k - s^k)^T (\nabla^2_{zz} \mathcal{L})_k (\Delta z^k - s^k) + O(\|s^k\| \|\Delta z^k\|) + O(\|s^k\|^2) \\ &\geq \tfrac{1}{2} \alpha \|\Delta z^k - s^k\|_2^2 + O(\|\Delta z^k\| \|s^k\|) + O(\|s^k\|^2) \\ &= \tfrac{1}{2} \alpha \|\Delta z^k\|_2^2 + O(\|\Delta z^k\| \|s^k\|) + O(\|s^k\|^2) \\ &= \tfrac{1}{2} \alpha \|\Delta z^k\|_2^2 + O(\|d_{\mathcal{A}^*}(z^k)\| \|\Delta z^k\|) + O(\|d_{\mathcal{A}^*}(z^k)\|^2) \\ &\qquad + O(\|d_{\mathcal{A}^*}(z^k)\| \|z^k - z^*\| \|\Delta z^k\|) + o(\|\Delta z^k\|^2) \\ &\geq \tfrac{1}{4} \alpha \|\Delta z^k\|_2^2 + O(\|d_{\mathcal{A}^*}(z^k)\| \|\Delta z^k\|) + O(\|d_{\mathcal{A}^*}(z^k)\|^2) \end{aligned} \tag{4.24}$$

for all $k \geq K_1$ with $\gamma_k \leq \bar{\gamma}$. Because of (4.21), and since $\bar{\lambda}_i^k = 0$ for $i \notin \mathcal{A}^*$, we have

$$-(\bar{\lambda}^k)^T d(z^k) = \sum_{i \in \mathcal{A}^*} \bar{\lambda}_i^k (-d_i(z^k)) \geq \bar{\lambda}_{\min} \|d_{\mathcal{A}^*}(z^k)\|_1. \tag{4.25}$$

By substituting (4.24) and (4.25) into (4.17) and dropping the second term on the right-hand side of (4.17) (which is positive in any case), we obtain

$$\begin{aligned} -m_k(\Delta z^k) &\geq -d(z^k)^T \bar{\lambda}^k + \frac{1}{2} (\Delta z^k)^T (\nabla^2_{zz} \mathcal{L})_k \Delta z^k \\ &\geq \bar{\lambda}_{\min} \|d_{\mathcal{A}^*}(z^k)\|_1 + \left(\frac{1}{8}\right) \alpha \|\Delta z^k\|_2^2 + O(\|d_{\mathcal{A}^*}(z^k)\| \|\Delta z^k\|) + O(\|d_{\mathcal{A}^*}(z^k)\|^2) \\ &\geq \left(\frac{1}{8}\right) \alpha \|\Delta z^k\|_2^2 \quad \text{for all } k \geq K_1. \end{aligned} \tag{4.26}$$

The last inequality holds because the term $\bar{\lambda}_{\min} \|d_{\mathcal{A}^*}(z^k)\|_1$ dominates the remainder terms (after, possibly, another decrease of Δ_τ and increase of K_1).

We conclude from (4.20) and (4.26) that for all k sufficiently large, we have $-m_k(\Delta z^k) = \Omega(\|\Delta z^k\|^2)$. By combining this estimate with (4.13) and (4.8), and using (1.6), we obtain that

$$\rho_k = 1 + \frac{o\left(\left\|\widetilde{\Delta z}^k\right\|^2\right)}{\Omega(\|\Delta z^k\|^2)} = 1 + \frac{o(\|\Delta z^k\|^2)}{\Omega(\|\Delta z^k\|^2)}.$$

Hence by decreasing Δ_τ further if necessary, we have $\rho_k > 1/2$ whenever $k \geq K_1$ and $\|D_k \Delta z^k\| \leq \Delta_\tau$, as claimed. ∎

The next lemma takes a few more steps toward our superlinear convergence result.

LEMMA 4.3. *Suppose that $p = 2$ in (1.3), that Assumptions 1, 4, and 5 hold, and that H_k is defined by (4.7). Let K_1 and Δ_τ be as defined in Lemma 4.2. Then the following are true:*

(a) *For all $k \geq K_1$, we have $\Delta_k \geq \min(\Delta_{K_1}, \Delta_\tau/2)$.*

(b) *There is an index K_2 such that the trust-region bound (1.3) is inactive at all successful iterations k with $k \geq K_2$.*

Proof. For (a), Lemma 4.2 indicates that for $k \geq K_1$, the trust-region radius can be decreased only when $\|D_k \Delta z^k\|_2 > \Delta_\tau$. Since Algorithm FP-SQP decreases the trust region by setting it to $(1/2)\|D_k \Delta z^k\|_2$, we must have $\Delta_{k+1} \geq \Delta_\tau/2$ after any such decrease. On the other hand, if no decreases occur after iteration K_1, we have $\Delta_k \geq \Delta_{K_1}$ for all $k \geq K_1$. The claim follows by combining these two observations.

For (b), we observed in (4.4) that $\|\Delta z^k\| \to 0$ for the successful steps, while from part (a), the trust-region radius is bounded below by a positive quantity. Hence, we can identify an index K_2 with the required property. ∎

THEOREM 4.4. *Suppose that $p = 2$ in (1.3), that Assumptions 1, 4, and 5 hold, and that H_k is defined by (4.7). Then the sequence $\{z^k\}$ converges Q-superlinearly to z^*.*

Proof. At all successful iterations k with $k \geq K_2$, the step Δz^k is a (full) standard SQP step. Hence by the known local convergence properties of SQP with an exact Hessian, we have that

$$\|z^k + \Delta z^k - z^*\| \leq \beta \|z^k - z^*\| \left[\|z^k - z^*\| + \|(\mu^k, \lambda^k) - (\mu^*, \lambda^*)\|\right] = o(\|z^k - z^*\|),$$

where β is a constant, and we have used Assumption 5(a) and (c) to obtain the final equality. It follows from this expression that

$$\|\Delta z^k\| = O(\|z^k - z^*\|).$$

Using this estimate, together with (4.3), we have

$$\begin{aligned} \|z^{k+1} - z^*\| &= \left\|z^k + \widetilde{\Delta z}^k - z^*\right\| \\ &\leq \|z^k + \Delta z^k - z^*\| + \left\|\Delta z^k - \widetilde{\Delta z}^k\right\| \\ &= o(\|z^k - z^*\|) + O(\|\Delta z^k\|^2) = o(\|z^k - z^*\|), \end{aligned} \tag{4.27}$$

showing that Q-superlinear behavior occurs at all successful steps with $k \geq K_2$.

We show now that there is an index $K_3 \geq K_2$ such that *all* iterations $k \geq K_3$ are successful. If not, then there are infinitely many unsuccessful iterations, and the trust-region radius is reduced (by at least a factor of 2) at each such iteration. Since, by Lemma 4.3(b), the trust region is inactive at the successful steps, the radius is not increased at these steps. Hence, we have $\Delta_k \downarrow 0$, which contradicts Lemma 4.3(a). ∎

5. Conclusions. We have described a simple feasibility perturbed trust-region SQP algorithm for nonlinear programming with good global and local convergence properties. As discussed above, we believe that the feasibility perturbation often can be carried out efficiently when the constraints are separable or otherwise structured. The companion paper [15] describes application of the algorithm to optimal control problems with constraints on the inputs (controls).

We assert (without proof) the following result concerning global convergence to points satisfying second-order necessary conditions. When the assumptions used in section 3 are satisfied, z^* is a KKT limit point of the sequence $\{z^k\}$ at which LICQ and strict complementarity are satisfied, asymptotically exact estimates of (μ^k, λ^k) and $\mathcal{W}_k$ are available on the convergent subsequence $\mathcal{K}$, H_k is chosen as in (4.7), and Assumption 5(d) is satisfied, then the following second-order necessary condition holds at z^*:

$$v^T \nabla^2_{zz} \mathcal{L}(z^*, \mu^*, \lambda^*) v \geq 0 \ \text{ for all } v \text{ such that } \nabla c(z^*)^T v = 0, \ \ \nabla d_{\mathcal{A}^*}(z^*)^T v = 0,$$

We omit the proof, which uses many of the same techniques as in sections 3 and 4.

Appendix. Value function of a parametrized linear program. Here we prove Lemma 3.1.

Proof. Note first that $C(z, 1) \geq 0$ for any feasible z, since $w = 0$ is feasible for (3.1).

We have $C(z, 1) = 0$ if and only if $w = 0$ is a solution of problem (3.1). The bound $w^T w \leq 1$ is inactive at $w = 0$, and the optimality conditions for (3.1) are then identical to the KKT conditions (1.8a–c) for (1.1). Hence, $C(z, 1) = 0$ if and only if z satisfies the KKT conditions.

Suppose now that $\bar{z} \in \mathcal{F}$ satisfies MFCQ (1.10a,b) but not KKT conditions (1.8a–c). Suppose for contradiction that there exists a sequence $\{z^l\}$ with $z^l \to \bar{z}$, $z^l \in \mathcal{F}$, such that

$$0 \leq C(z^l, 1) \leq l^{-1}, \ \ l = 1, 2, 3, \ldots.$$

The KKT conditions for the solution w^l of (3.1) at $z = z^l$ are that there exist multipliers $\mu^l \in \mathsf{R}^m$, $\lambda^l \in \mathsf{R}^r$, and $\beta^l \in \mathsf{R}$ such that

$$\nabla f(z^l) + \nabla c(z^l)\mu^l + \nabla d(z^l)\lambda^l + 2\beta^l w^l = 0, \tag{A.1a}$$

$$c(z^l) + \nabla c(z^l)^T w^l = 0, \tag{A.1b}$$

$$d(z^l) + \nabla d(z^l)^T w^l \leq 0 \perp \lambda^l \geq 0, \tag{A.1c}$$

$$(w^l)^T w^l - 1 \leq 0 \perp \beta^l \geq 0. \tag{A.1d}$$

We now verify that these are in fact optimality conditions for (3.1) by showing that MFCQ holds at w^l. We define the "linearized" active indices at z^l as follows:

$$\mathcal{A}_l \stackrel{\text{def}}{=} \{i = 1, 2, \ldots, r \,|\, d_i(z^l) + \nabla d_i(z^l)^T w^l = 0\}.$$

Since MFCQ holds for the original problem (1.1) at $\bar{z}$, we have by the logic in the proof of Lemma 2.1 that MFCQ is also satisfied at z^l for all l sufficiently large. Hence, there is a vector v^l such that

$$\nabla c(z^l)^T v^l = 0 \text{ and } \nabla d_i(z^l)^T v^l < 0 \text{ for all } i \in \mathcal{A}(z^l).$$

Consider now the vector

$$u^l = -w^l + \epsilon v^l$$

for some $\epsilon > 0$ to be defined. We show that u^l is an "MFCQ direction" for (3.1) at w^l, that is,

$$\|w^l\|_2 = 1 \quad \Rightarrow \quad 2(w^l)^T u^l < 0, \tag{A.2a}$$

$$\nabla c(z^l)^T u^l = 0, \tag{A.2b}$$

$$i \in \mathcal{A}_l \cap \mathcal{A}(z^l) \quad \Rightarrow \quad \nabla d_i(z^l)^T u^l < 0, \tag{A.2c}$$

$$i \in \mathcal{A}_l \backslash \mathcal{A}(z^l) \quad \Rightarrow \quad \nabla d_i(z^l)^T u^l < 0. \tag{A.2d}$$

For (A.2a) we have, when $\|w^l\|_2 = 1$, that

$$(2w^l)^T u^l = -2\|w^l\|_2^2 + \epsilon (w^l)^T v^l = -2 + \epsilon (w^l)^T v^l < 0$$

for all $\epsilon > 0$ sufficiently small. The second condition (A.2b) obviously holds, since $\nabla c(z^l)^T w^l = 0$ and $\nabla c(z^l)^T v^l = 0$. For (A.2c), we have

$$\nabla d_i(z^l)^T u^l = -\nabla d_i(z^l)^T w^l + \epsilon \nabla d_i(z^l)^T v^l = d_i(z^l) + \epsilon \nabla d_i(z^l)^T v^l \leq \epsilon \nabla d_i(z^l)^T v^l < 0$$

for all $\epsilon > 0$, where the second equality follows from $i \in \mathcal{A}_l$ and the third equality from $z^l \in \mathcal{F}$. For (A.2d), we have from $i \notin \mathcal{A}(z^l)$ that $d_i(z^l) < 0$, and so

$$\nabla d_i(z^l)^T u^l = -\nabla d_i(z^l)^T w^l + \epsilon \nabla d_i(z^l)^T v^l = d_i(z^l) + \epsilon \nabla d_i(z^l)^T v^l < 0$$

for all $\epsilon > 0$ sufficiently small. It is clearly possibly to choose ϵ in such a way that all conditions (A.2a–d) are satisfied, so we conclude that (A.1a–d) are indeed optimality conditions for w^l.

From these relations, and using the fact that $z^l \in \mathcal{F}$, we have that

$$\begin{aligned} C(z^l, 1) &= -\nabla f(z^l)^T w^l \\ &= (w^l)^T \nabla c(z^l) \mu^l + (w^l)^T \nabla d(z^l) \lambda^l + 2\beta^l (w^l)^T w^l \\ &= -d(z^l)^T \lambda^l + 2\beta^l \geq 0. \end{aligned} \tag{A.3}$$

By taking limits as $l \to \infty$, and since $-d(z^l)^T \lambda^l$ and β^l are both nonnegative, we have from (A.3) that

$$\beta^l \to 0, \quad d(z^l)^T \lambda^l \to 0. \tag{A.4}$$

Consider first the case in which there is a subsequence $\mathcal{K}$ of multipliers from (A.1); that is, $\{\mu^l, \lambda^l\}_{l \in \mathcal{K}}$ is bounded. By compactness, and taking a further subsequence of $\mathcal{K}$ if necessary, we can identify $\bar{\mu}$ and $\bar{\lambda} \geq 0$ such that

$$(\mu^l, \lambda^l)_{l \in \mathcal{K}} \to (\bar{\mu}, \bar{\lambda}). \tag{A.5}$$

Then by taking limits in (A.1a), and using (A.4) and (1.7), we have that

$$\nabla_z \mathcal{L}(\bar{z}, \bar{\mu}, \bar{\lambda}) = 0, \quad d(\bar{z})^T \bar{\lambda} = 0. \tag{A.6}$$

By using these relations, together with feasibility of $\bar{z}$, we see that $\bar{z}$ is a KKT point, which is a contradiction.

In the other case, the sequence $\{\mu^l, \lambda^l\}$ has no bounded subsequence. By taking another subsequence $\mathcal{K}$, we can identify a vector $(\hat{\mu}, \hat{\lambda})$ with $\|(\hat{\mu}, \hat{\lambda})\|_2 = 1$ and $\hat{\lambda} \geq 0$ such that

$$\lim_{k \in \mathcal{K}} \frac{(\mu^l, \lambda^l)}{\|(\mu^l, \lambda^l)\|_2} = (\hat{\mu}, \hat{\lambda}), \quad \lim_{k \in \mathcal{K}} \|(\mu^l, \lambda^l)\|_2 = \infty.$$

By dividing both sides of (A.1a) by $\|(\mu^l, \lambda^l)\|_2$ and using (A.4), we obtain

$$\nabla c(\bar{z})\hat{\mu} + \nabla d(\bar{z})\hat{\lambda} = 0, \quad d(\bar{z})^T \hat{\lambda} = 0, \quad \hat{\lambda} \geq 0. \tag{A.7}$$

It is easy to show that (A.7), together with the MFCQ (1.10a,b), implies that $(\hat{\mu}, \hat{\lambda}) = 0$, which contradicts $\|(\hat{\mu}, \hat{\lambda})\|_2 = 1$ (see Clarke [4, pp. 235–236]).

Therefore, we obtain a contradiction, so that no sequence $\{z^l\}$ with the claimed properties exists, and therefore $C(z, 1)$ is bounded away from zero in a neighborhood of $\bar{z}$. □

Acknowledgments. We thank Paul Tseng for suggesting Assumption 3. We are extremely grateful to the editor and to two anonymous referees whose careful reading and penetrating insights greatly improved the paper. We gratefully acknowledge the financial support of the industrial members of the Texas-Wisconsin Modeling and Control Consortium.

REFERENCES

[1] D. P. Bertsekas, *Projected Newton methods for optimization problems with simple constraints*, SIAM J. Control Optim., 20 (1982), pp. 221–246.

[2] J. R. Birge, L. Qi, and Z. Wei, *A varaint of the Topkis–Veinott method for solving inequality constrained optimization problems*, J. Appl. Math. Optim., 41 (2000), pp. 309–330.

[3] M. R. Celis, J. E. Dennis, and R. A. Tapia, *A trust region strategy for nonlinear equality constrained optimization*, in Numerical Optimization, 1984: Proceeding of the SIAM Conference on Numerical Optimization (Boulder, CO, June 12–14, 1984), P. T. Boggs, R. H. Byrd, and R. B. Schnabel, eds., SIAM, Philadelphia, 1985, pp. 71–82.

[4] F. H. Clarke, *Optimization and Nonsmooth Analysis*, John Wiley, New York, 1983.

[5] A. R. Conn, N. I. M. Gould, and Ph. L. Toint, *Trust-Region Methods*, MPS-SIAM Series on Optimization 1, SIAM, Philadelphia, 2000.

[6] F. Facchinei, A. Fischer, and C. Kanzow, *On the accurate identification of active constraints*, SIAM J. Optim., 9 (1998), pp. 14–32.

[7] R. Fletcher and E. Sainz de la Maza, *Nonlinear programming and nonsmooth optimization by successive linear programming*, Math. Programming, 43 (1989), pp. 235–256.

[8] M. Heinkenschloss, *Projected sequential quadratic programming methods*, SIAM J. Optim., 6 (1996), pp. 373–417.

[9] A. J. Hoffman, *On approximate solutions of systems of linear inequalities*, J. Res. Nat. Bur. Standards, 49 (1952), pp. 263–265.

[10] C. T. Lawrence and A. L. Tits, *A computationally efficient feasible sequential quadratic programming algorithm*, SIAM J. Optim., 11 (2001), pp. 1092–1118.

[11] S. Lucidi, M. Sciandrone, and P. Tseng, *Objective-derivative-free methods for constrained optimization*, Math. Program. Ser. A, 92 (2002), pp. 37–59.

[12] E. O. Omojokun, *Trust-Region Algorithms for Optimization with Nonlinear Equality and Inequality Constraints*, Ph.D. thesis, University of Colorado, Boulder, 1989.

[13] E. R. Panier and A. L. Tits, *On combining feasibility, descent and superlinear convergence in inequality constrained optimization*, Math. Programming Ser. A, 59 (1993), pp. 261–276.

[14] S. M. Robinson, *Stability theory for systems of inequalities, Part* II*: Differentiable nonlinear systems*, SIAM J. Numer. Anal., 13 (1976), pp. 497–513.

[15] M. J. Tenny, S. J. Wright, and J. B. Rawlings, *Nonlinear model predictive control via feasibility-perturbed sequential quadratic programming*, Comput. Optim. Appl., 28 (2004), pp. 87–121.
[16] S. J. Wright, *Convergence of an inexact algorithm for composite nonsmooth optimization*, IMA J. Numer. Anal., 10 (1990), pp. 299–321.
[17] Y. Yuan, *On the superlinear convergence of a trust region algorithm for nonsmooth optimization*, Math. Programming, 31 (1985), pp. 269–285.

SIAM J. OPTIM.
Vol. 14, No. 4, pp. 1106–1127

SOLUTION STABILITY OF NONSMOOTH CONTINUOUS SYSTEMS WITH APPLICATIONS TO CONE-CONSTRAINED OPTIMIZATION*

V. JEYAKUMAR† AND N. D. YEN‡

Abstract. In this paper we establish conditions for stability, metric regularity, and a pseudo-Lipschitz property of the solution maps of parametric inequality systems involving nonsmooth (not necessarily locally Lipschitz) continuous functions and closed convex sets. We also derive open mapping and inverse mapping theorems for nonsmooth continuous functions, Lagrange multiplier rules for nonsmooth cone-constrained optimization problems, and conditions for the continuity of the optimal value functions of optimization problems. The main tool used is a generalized Jacobian, called approximate Jacobian. It provides a flexible nonsmooth local analysis of continuous functions and often gives sharp calculus rules for locally Lipschitz functions. The regularity condition, which plays a key role in the local analysis, is a new extension of the Robinson regularity condition for continuous functions.

Key words. inequality systems, continuous function, approximate Jacobians, pseudo-Lipschitz property, Lagrange multipliers

AMS subject classifications. Primary, 49J52, 49J53; Secondary, 90C31

DOI. 10.1137/S1052623402419236

1. Introduction.

Consider the generalized inequality system

$$0 \in f(x) + K, \quad x \in C, \tag{1.1}$$

where $C \subset R^n$ and $K \subset R^m$ are nonempty closed convex sets and $f : R^n \to R^m$ is a continuous function. A perturbation of (1.1) is a parametric inequality system of the form

$$0 \in f(x,p) + K, \quad x \in C, \tag{1.2}$$

where p is a parameter belonging to a set $P \subset R^r$, $f : R^n \times P \to R^m$ is a given function. We assume that for every $p \in P$ the function $f(\cdot, p)$ is continuous and there exists $p_0 \in P$ such that

$$f(x, p_0) = f(x) \quad \forall x \in R^n. \tag{1.3}$$

The perturbation (1.2) is denoted by $\{f(x,p), P, p_0\}$. For each $p \in P$, let

$$G(p) = \{x \in C : 0 \in f(x,p) + K\}$$

be the solution set of (1.2). Thus $G(\cdot)$ is the *implicit multifunction* defined by the parametric system (1.2). Note that if

$$\begin{aligned} K &= R^s_+ \times \{0\}_{m-s} \\ &:= \{y = (y_1, \ldots, y_m) \in R^m : y_1 \geq 0, \ldots, y_s \geq 0,\ y_{s+1} = \cdots = y_m = 0\}, \end{aligned}$$

*Received by the editors December 5, 2002; accepted for publication (in revised form) January 26, 2004; published electronically July 20, 2004. This work was supported by a research grant from the University of New South Wales.
http://www.siam.org/journals/siopt/14-4/41923.html
†Department of Applied Mathematics, University of New South Wales, Sydney, Australia (jeya@maths.unsw.edu.au).
‡Institute of Mathematics, 10307 Hanoi, Vietnam (ndyen@math.ac.vn). The work of this author was carried out while he was visiting the University of New South Wales.

then (1.1) (resp., (1.2)) is a system of s inequalities and $m - s$ equalities with the constraint set C. We say that (1.1) is a *smooth* (resp., *locally Lipschitz*, *continuous*) generalized inequality system if f is a C^1- function (resp., a locally Lipschitz function, a continuous function). Robinson [23] established a fundamental theorem on the stability of smooth generalized inequalities systems that states that if the system is regular at a certain solution, then this solution is stable when the system undergoes a small admissible perturbation. Robinson's result has been extended to systems including nonsmooth functions (see, for instance, [3, 7, 26, 27]) and to systems including normal-cone operators (see, for instance, [18, 20, 24]).

The aim of this paper is to establish general conditions for stability of solutions of nonsmooth (not necessarily locally Lipschitz) continuous generalized inequality system (1.1) and apply them to obtain inverse function and open mapping theorems, and Lagrange multiplier rules for cone-constrained optimization problems. This is achieved by employing the recent theory of approximate Jacobians (see [10, 11, 12, 13]) and using a new extension of the Robinson regularity condition for continuous functions. It turns out that approximate Jacobians provide a useful device for treating problems that have continuous, not necessarily locally Lipschitz functions. They enjoy rich calculus for continuous functions and frequently give sharp rules for locally Lipschitz functions as the Clarke generalized Jacobian may contain the closed convex hull of an approximate Jacobian. Moreover, several other known generalized derivatives of vector functions such as the Ioffe prederivative and the Warga unbounded derivative containers are examples of approximate Jacobians. On the other hand, the coderivative (see [19] and [24]) has been shown to be a useful tool for studying nonsmooth systems. However, the coderivative and the approximate Jacobian are not directly comparable. See [21] and also section 3 for a detailed comparison between our results and the corresponding results in [18, 20].

The organization of the paper is as follows. Section 2 presents basic results on approximate Jacobians and definitions of regularity, admissible perturbation, and stability of the continuous generalized inequality system (1.1). Section 3 gives sufficient conditions for the multifunction $p \mapsto G(p) \cap V$, where V is a neighborhood of x_0, to be lower semicontinuous on a neighborhood of p_0, for the metric regularity of $G(\cdot)$ at (p_0, x_0), and for the pseudo-Lipschitz property of $G(\cdot)$ at (p_0, x_0). It also provides two examples showing that, unlike the case of inverse multifunctions, for implicit multifunctions the metric regularity and the pseudo-Lipschitz property are two independent concepts. Section 4 gives open mapping and inverse mapping theorems and derives necessary optimality conditions for optimization problems with continuous data, as an application of the results of section 3.

2. Definitions and preliminaries. For a Euclidean space Z, $\|\cdot\|$, $\langle\cdot,\cdot\rangle$, B_Z, and S_Z denote, respectively, the norm, the inner product, the closed unit ball, and the unit sphere in Z. Subscripts will be deleted if no confusion is possible. The closed ball with center a and radius δ is denoted by $B(a, \delta)$. For a subset $M \subset Z$, we denote by $\mathrm{int}M$, $\overline{M}$, $\mathrm{co}M$, and $\mathrm{cone}M$ the interior, the closure, the convex hull, and the cone generated by M, respectively. For simplicity of notation, the closure of the last two sets are denoted, respectively, by $\overline{\mathrm{co}}M$ and $\overline{\mathrm{cone}}M$. The negative dual cone of M is denoted by M^*, that is, $M^* = \{w \in Z : \langle w, z\rangle \le 0\ \forall z \in M\}$. The distance from $a \in Z$ to $M \subset Z$ is denoted by $d(a, M)$. By convention, $d(a, \emptyset) = +\infty$. If A is a linear operator, then A^* stands for the conjugate of A. A multifunction $F : X \to 2^{R^s}$, where X is a subset in R^k, is said to be *upper semicontinuous* (usc) at $\overline{x} \in X$ if for any open set $V \subset R^s$ satisfying $F(\overline{x}) \subset V$ there exists $\delta > 0$ such that $F(x) \subset V$

for every $x \in X \cap B(\overline{x}, \delta)$. We say that F is *lower semicontinuous* (lsc) at $\overline{x} \in X$ if $F(\overline{x}) \neq \emptyset$ and for any open set $V \subset R^s$ with $F(\overline{x}) \cap V \neq \emptyset$ there exists $\delta > 0$ such that $F(x) \cap V \neq \emptyset$ for every $x \in X \cap B(\overline{x}, \delta)$. If F is usc (resp., lsc) at any point of X, then we say that F is usc (resp., lsc) on X. F is said to be *pseudo-Lipschitz* or *Aubin continuous* (see [24]) at $(\overline{x}, \overline{y})$, where $\overline{y} \in F(\overline{x})$, if there exist $\ell > 0$, $\varepsilon > 0$, and $\delta > 0$ such that

$$F(x') \cap B(\overline{y}, \varepsilon) \subset F(x) + \ell \|x' - x\| B_{R^s} \quad \forall x, x' \in X \cap B(\overline{x}, \delta).$$

For a subset $M \subset Z$, the recession cone M_∞ of M (see [12, 13, 24]) is the set of all $w \in Z$ for which there exists a sequence $\{t_k\}$ of positive numbers converging to 0 and sequence $\{z_k\} \subset M$ such that $w = \lim_{k\to\infty} t_k z_k$. For a cone $M \subset Z$ and for a number $\varepsilon \in (0, 1)$, the ε-conic neighborhood M^ε of M (see [12, 13]) is defined by the formula

$$M^\varepsilon = \{z + \varepsilon \|z\| B_Z : z \in M\}.$$

For simplicity, we abbreviate $(M_\infty)^\varepsilon$ to M_∞^ε.

We will need some facts concerning approximate Jacobians, which have been given in [10, 11, 12, 13].

Let $f : R^n \to R^m$ be a continuous map. A closed subset $Jf(x)$ of the space $L(R^n, R^m)$ of linear operators from R^n to R^m (which is identified with the set of $(m \times n)$-matrices) is called an *approximate Jacobian* of f at $\overline{x} \in R^n$ if, for every $u = (u_1, \ldots, u_n) \in R^n$ and $v = (v_1, \ldots, v_m) \in R^m$, one has

$$(vf)^+(\overline{x}, u) \leq \sup_{A \in Jf(\overline{x})} \langle v, Au \rangle,$$

where $(vf)(x) = v_1 f_1(x) + \cdots + v_m f_m(x)$ is the composite function of v and f, and

$$(vf)^+(\overline{x}, v) = \limsup_{t \downarrow 0} \frac{(vf)(\overline{x} + tu) - (vf)(\overline{x})}{t}$$

is the *upper Dini directional derivative* of vf at $\overline{x}$ in direction u. If $m = 1$, then one also writes $\partial f(\overline{x})$ for $Jf(\overline{x})$ and calls $\partial f(\overline{x})$ a *generalized subdifferential* of f at $\overline{x}$.

If f is Fréchet differentiable at $\overline{x}$ with the Fréchet derivative $f'(\overline{x})$, then $Jf(\overline{x}) = \{f'(\overline{x})\}$ is an approximate Jacobian of f at $\overline{x}$. If f is locally Lipschitz at $\overline{x}$, i.e., there exist $\ell > 0$ such that $\|f(x') - f(x)\| \leq \ell \|x' - x\|$ for all x, x' in a neighborhood of $\overline{x}$, then the generalized Jacobian in the sense of Clarke [5],

$$\partial^c f(\overline{x}) = \operatorname{co} \left\{ \lim_{k\to\infty} f'(x_k) : \{x_k\} \subset \Omega_f, x_k \to \overline{x} \right\},$$

is a compact, convex approximate Jacobian of f at $\overline{x}$. Here

$$\Omega_f = \{x \in R^n : \exists \text{ the Fréchet derivative } f'(x) \text{ of } f \text{ at } x\}.$$

If f is locally Lipschitz and $m = 1$, then the set $\partial^c f(\overline{x})$ collapses to the *Clarke generalized gradient* of f at $\overline{x}$ (see [5]).

Let us consider the following simple illustrative example [12] of approximate Jacobian of a non-Lipschitz function. Many other examples can be found in [10, 12].

Example 2.1. Let $f(x) = x^{1/3}$, $x \in R$. For $\overline{x} = 0$, it is easily verified that $Jf(\overline{x}) = [\alpha, +\infty)$, where $\alpha \in R$ is an arbitrary number, is an approximate Jacobian

of f at $\overline{x}$. For $\overline{x} \neq 0$, the set $Jf(\overline{x}) = \left\{\frac{1}{3}\overline{x}^{-2/3}\right\}$ is an approximate Jacobian of f at $\overline{x}$. It is clear that the approximate Jacobian mapping $x \mapsto Jf(x)$ is upper semicontinuous at $x = 0$.

The following chain rule plays a crucial role in deriving the main results. For completeness, the proof is given in the appendix.

PROPOSITION 2.2 (chain rule; see [12, Corollary 4.2]). *Let $f : R^n \to R^m$ be a continuous map, $g : R^m \to R$ a continuous function. Assume that*

(i) *f admits an approximate Jacobian mapping Jf which is upper semicontinuous at $\overline{x} \in R^n$;*

(ii) *g is Fréchet differentiable in a neighborhood of $f(\overline{x})$ and the gradient mapping ∇g is continuous at $f(\overline{x})$ with $\nabla g(f(\overline{x})) \neq 0$.*

Then, for every $\varepsilon > 0$, the closure of the set

$$\nabla g(f(\overline{x})) \circ \left[Jf(\overline{x}) + (Jf(\overline{x}))^{\varepsilon}_{\infty}\right]$$

is an approximate Jacobian of $g \circ f$ at $\overline{x}$.

DEFINITION 2.3 (surjectivity). *An operator $A \in L(R^n, R^m)$ is said to be* surjective *on a nonempty closed convex set $C \subset R^n$ at $x_0 \in C$ with respect to a nonempty closed set $K_0 \subset R^m$ with $0 \in K_0$ if*

$$0 \in \operatorname{int}(A[T_C(x_0)] + K_0), \tag{2.1}$$

where $T_C(x_0) = \overline{cone}(C - x_0)$ is the tangent cone of C at x_0 in the sense of convex analysis.

In the case where $K_0 = \{0\}$, it is easy to show that (2.1) is equivalent to the condition $0 \in \operatorname{int}(A[C - x_0])$. Thus the above definition is an extension of the notion given in [13]. (Note that in [13] the convex set C may not be closed, and instead of $x_0 \in C$ one uses the condition $x_0 \in \overline{C}$.)

The following necessary optimality condition, which easily follows from the definition of the generalized subdifferential, will be useful in the sequel.

PROPOSITION 2.4 (see [13, Proposition 2.1]). *Let $C \subset R^n$ be a convex set and let $\varphi : R^n \to R$ be continuous. If $\overline{x} \in C$ is a local minimum point of φ on C and if $\partial f(\overline{x})$ is a generalized subdifferential of φ at $\overline{x}$, then*

$$\sup_{\eta \in \partial f(\overline{x})} \langle \eta, u \rangle \geq 0 \quad \forall u \in T_C(\overline{x}).$$

We now turn our attention back to the generalized inequality system (1.1). Let us define an extension of the regularity concept introduced by Robinson [23] to the system. Let x_0 be a solution of (1.1).

DEFINITION 2.5 (regularity condition). *For the system* (1.1), *assume that f admits an approximate Jacobian mapping Jf. Then the system is said to be* regular *at x_0 if*

$$0 \in \operatorname{int}(A[T_C(x_0)] + f(x_0) + K) \quad \forall A \in \overline{\operatorname{co}} Jf(x_0) \cup \operatorname{co}((Jf(x_0))_{\infty} \setminus \{0\}). \tag{2.2}$$

In the next section it will be shown (see Lemma 3.1) that this regularity condition implies a uniform openness property of the operators $A \in Jf(x)$, where x belongs to a neighborhood of x_0. A comparison of (2.2) with (2.1) shows that (1.1) is regular at x_0 if and only if each operator A of the set $\overline{\operatorname{co}} Jf(x_0) \cup \operatorname{co}((Jf(x_0))_{\infty} \setminus \{0\})$ is surjective on C at x_0 w.r.t. $K_0 := f(x_0) + K$.

It is easily verified that the inequality system (1.1), where $n = m = 1$, $C = R$, $K = \{0\}$, and $f(x) = x^{1/3}$, is regular at the solution $x_0 = 0$. Note that the approximate Jacobian mapping Jf has been described in Example 2.1.

DEFINITION 2.6 (admissible perturbation). *A perturbation* $\{f(x,p), P, p_0\}$ *of* (1.1) *is said to be an admissible perturbation of the system at* x_0 *if*

(i) *the function* $f(x,p)$ *is continuous at* (x_0, p_0),

(ii) *for every* $x \in R^n$ *the function* $f(x,\cdot)$ *is continuous on* P,

(iii) *for every* $p \in P$ *the function* $f(\cdot,p)$ *admits an approximate Jacobian mapping denoted by* $J_1 f(\cdot,p)$,

(iv) *there exist a neighborhood* U_* *of* $p_0 \in P$ *and a number* $\delta_* > 0$ *such that, for every* $p \in U_*$, $J_1 f(\cdot,p)$ *is upper semicontinuous on* $B(x_0, \delta_*)$,

(v) *the multifunction* $(x,p) \mapsto J_1 f(x,p)$ *is upper semicontinuous at* (x_0, p_0).

DEFINITION 2.7 (stability). *We say that solution* x_0 *of* (1.1) *is* stable *under admissible perturbations if for every* $\varepsilon > 0$ *and for every admissible perturbation* $\{f(x,p), P, p_0\}$ *of* (1.1) *at* x_0, *there exists a neighborhood* U *of* p_0 *such that*

$$G(p) \cap B(x_0, \varepsilon) \neq \emptyset \quad \forall p \in U,$$

where $G(p)$ *is the solution set of* (1.2).

In the following example, we consider one special type of admissible perturbations of continuous generalized inequality systems.

Example 2.8. Suppose that $f : R^n \to R^m$ is a continuous function, $C \subset R^n$ a closed convex set. We put $P = R^m$, $p_0 = 0$, and consider the function $f : R^n \times P \to R^m$ defined by the formula $f(x,p) = f(x) - p$ for all $(x,p) \in R^n \times R^m$. It is clear that $\{f(x,p), P, p_0\}$ is a perturbation of (1.1). If, in addition, the function $f : R^n \to R^m$ admits an approximate Jacobian mapping Jf that is usc at any $x \in R^n$, then $\{f(x,p), P, p_0\}$ is an admissible perturbation of (1.1). Indeed, to verify this it suffices to note that, for every $p \in P$, formula $J_1 f(x,p) = Jf(x)$ $(x \in R^n)$ defines an approximate Jacobian mapping of the function $f(\cdot,p)$. It is also clear that the multifunction $(x,p) \mapsto J_1 f(x,p)$ is usc at (x_0, p_0). To have a concrete example, we define $f : R^2 \to R^2$ by setting $f(x_1, x_2) = (x_1^{2/3}, x_2)$ for all $(x_1, x_2) \in R^2$. Then the formulas

$$J_1 f(x,p) = \left\{ \begin{pmatrix} \frac{1}{3} x^{-2/3} & 0 \\ 0 & 1 \end{pmatrix} \right\} \quad (\forall x \neq 0) \quad \text{and} \quad J_1 f(0,p) = \left\{ \begin{pmatrix} \alpha & 0 \\ 0 & 1 \end{pmatrix} \right\},$$

where $\alpha > 0$, define an approximate Jacobian mapping of $f(\cdot,p)$, where $f(x,p) = f(x) - p$ $(p \in R^2)$.

3. Stability and implicit functions. Conditions for the solution stability of generalized inequality systems will be established in this section. Theorem 3.2 gives sufficient conditions for the truncated multifunction $p \mapsto G(p) \cap V$, where V is a neighborhood of x_0, is lsc on a neighborhood of p_0. Theorem 3.4 deals with the metric regularity of $G(\cdot)$ at (p_0, x_0), and Theorem 3.5 treats the pseudo-Lipschitz property of that implicit multifunction at (p_0, x_0).

Throughout this section it is assumed that $x_0 \in C$ is a solution of (1.1) and $\{f(x,p), P, p_0\}$ is an admissible perturbation of (1.1) at x_0.

The following lemma on uniform openness of a family of linear operators is crucial for obtaining the results of this section. This lemma is an extended version of Lemma 3.1 from [13] where, in our notation, the case $K = \{0\}$ and $P = \{p_0\}$ was treated.

LEMMA 3.1 (uniform openness). *If* (1.1) *is regular at* x_0*, then there exist* $\gamma > 0$ *and* $\delta > 0$ *such that*

$$B_{R^m} \subset \gamma \left(A \left[T_C(x) \cap B_{R^n} \right] + \left[\overline{\text{cone}}(K + f(x,p)) \cap B_{R^m} \right] \right) \tag{3.1}$$

for every $x \in B(x_0, \delta) \cap C$*,* $p \in B(p_0, \delta) \cap P$*, and*

$$A \in \bigcup_{x' \in B(x_0,\delta),\ p' \in B(p_0,\delta) \cap P} \overline{\text{co}} \left(J_1 f(x', p') + (J_1 f(x', p'))^{\delta}_{\infty} \right). \tag{3.2}$$

Proof. We will follow closely the proof scheme of Lemma 3.1 in [13]. Suppose our lemma were false. Then for each $k \geq 1$ and $\delta = 1/k$ we could find $v_k \in B_{R^m}$, $x_k, x_k' \in B(x_0, 1/k) \cap C$, $p_k, p_k' \in B(p_0, 1/k) \cap P$, and

$$A_k \in \overline{\text{co}} \left(J_1 f(x_k', p_k') + (J_1 f(x_k', p_k'))^{1/k}_{\infty} \right)$$

such that

$$v_k \notin k \left(A_k \left[T_C(x_k) \cap B_{R^n} \right] + \left[\overline{\text{cone}}(f(x_k, p_k) + K) \cap B_{R^m} \right] \right). \tag{3.3}$$

There is no loss of generality in assuming that

$$\lim_{k \to \infty} v_k = v_0 \in B_{R^m}.$$

We claim that, by taking a subsequence if necessary, it can be assumed that either

$$\lim_{k \to \infty} A_k = A_0 \in \overline{\text{co}} J_1 f(x_0, p_0) \tag{3.4}$$

or

$$\lim_{k \to \infty} t_k A_k = A_* \in \text{co} \left((J_1 f(x_0, p_0))_{\infty} \setminus \{0\} \right), \tag{3.5}$$

where $\{t_k\}$ is some sequence of positive numbers converging to 0.

We first show that (3.4) and (3.5) lead to a contradiction.

If (3.4) holds, then by (1.3) and the regularity condition (2.2) we have

$$0 \in \text{int} \left(A_0[T_C(x_0)] + f(x_0, p_0) + K \right).$$

Since $f(x_0, p_0) + K \subset \overline{\text{cone}}(f(x_0, p_0) + K)$, from the last inclusion we deduce that

$$R^m = A_0[T_C(x_0)] + \overline{\text{cone}}(f(x_0, p_0) + K). \tag{3.6}$$

It is clear that

$$\Omega := A_0[T_C(x_0) \cap B_{R^n}] + [\overline{\text{cone}}(f(x_0, p_0) + K) \cap B_{R^m}]$$

is a compact, convex set, and $0 \in \Omega$. If $0 \notin \text{int}\Omega$, then, by the separation theorem, there exists $\eta \in S_{R^m}$ such that

$$\Omega \subset \{y \in R^m : \langle \eta, y \rangle \geq 0\}.$$

For any $v \in R^m$, by (3.6) there exist $u \in T_C(x_0)$ and $v \in \overline{\text{cone}}(f(x_0, p_0) + K)$ such that $v = A_0 u + w$. If we select $t > 0$ as small as $tu \in B_{R^n}$ and $tw \in B_{R^m}$, then

$tv = A_0(tu) + tw \in \Omega$. Therefore $\langle \eta, tv\rangle \geq 0$, and hence $\langle \eta, v\rangle \geq 0$. Since the last inequality holds for any $v \in R^m$, we have arrived at a contradiction. Thus $0 \in \text{int}\Omega$. From this it follows that there exist $\varepsilon > 0$ and $k_0 > 1$ such that

$$B(v_0, \varepsilon) \subset k_0 \left(A_0 \left[T_C(x_0) \cap B_{R^n}\right] + \left[\overline{\text{cone}}(f(x_0, p_0) + K) \cap B_{R^m}\right]\right). \tag{3.7}$$

Since $A_k \to A_0$, there exists $k_1 \geq k_0$ such that

$$\|A_k - A_0\| < \varepsilon/4 \quad \text{for every } k \geq k_1. \tag{3.8}$$

We now show that there is $k_2 \geq k_1$ such that

$$B\left(v_0, \frac{\varepsilon}{2}\right) \subset k_0 \left(A_0 \left[T_C(x_k) \cap B_{R^n}\right] + \left[\overline{\text{cone}}(f(x_k, p_k) + K) \cap B_{R^m}\right]\right) \tag{3.9}$$

for every $k \geq k_2$. Indeed, if this is not valid, then we can assume that for each k there is an element $u_k \in B(v_0, \varepsilon/2)$ satisfying

$$u_k \notin k_0 \left(A_0 \left[T_C(x_k) \cap B_{R^n}\right] + \left[\overline{\text{cone}}(f(x_k, p_k) + K) \cap B_{R^m}\right]\right).$$

By the separation theorem, there exists $\xi_k \in S_{R^m}$ such that

$$\langle \xi_k, u_k\rangle \geq \langle \xi_k, k_0(A_0 z + w)\rangle \tag{3.10}$$

for every $z \in T_C(x_k) \cap B_{R^n}$ and $w \in \overline{\text{cone}}(f(x_k, p_k) + K) \cap B_{R^m}$. Using subsequences if necessary, we can assume that

$$\lim_{k\to\infty} u_k = u_0 \in B\left(v_0, \frac{\varepsilon}{2}\right), \quad \lim_{k\to\infty} \xi_k = \xi_0, \quad \text{where } \|\xi_0\| = 1.$$

From (3.10) we deduce that

$$\langle \xi_0, u_0\rangle \geq \langle \xi_0, k_0(A_0 z + w)\rangle \tag{3.11}$$

for all $z \in T_C(x_0) \cap B_{R^n}$ and $w \in \overline{\text{cone}}(f(x_0, p_0) + K) \cap B_{R^m}$. Indeed, to prove this claim it suffices to show that (3.11) is valid for any $z \in \text{cone}(C - x_0) \cap B_{R^n}$ and $w \in \text{cone}(f(x_0, p_0) + K) \cap B_{R^m}$. Let there be given any pair (z, w) satisfying the last two inclusions. Suppose that

$$z = t(c - x_0), \quad w = \tau(f(x_0, p_0) + v)$$

for some $c \in C$, $t, \tau \in [0, +\infty)$ and $v \in K$. For each k, we put

$$z_k = t(c - x_k), \quad w_k = \tau(f(x_k, p_k) + v).$$

Then $z_k \in T_C(x_k)$, $w_k \in \overline{\text{cone}}(f(x_k, p_k) + K)$, $z_k \to z$ and $w_k \to w$ as $k \to \infty$. If $z_k \in B_{R^n}$, then we set $z'_k = z_k$. If $z_k \notin B_{R^n}$, then we set $z'_k = (\|z\|/\|z_k\|)z_k$. Similarly, if $w_k \in B_{R^m}$, then we set $w'_k = w_k$. If $w_k \notin B_{R^m}$, then we set $w'_k = (\|w\|/\|w_k\|)w_k$. Clearly, $z'_k \in T_C(x_k) \cap B_{R^n}$ and $w'_k \in \overline{\text{cone}}(f(x_k, p_k) + K) \cap B_{R^m}$ for each k. Note that $z'_k \to z$ and $w'_k \to w$ as $k \to \infty$. By (3.10), we have

$$\langle \xi_k, u_k\rangle \geq \langle \xi_k, k_0(A_0 z'_k + w'_k)\rangle \qquad \forall k.$$

Letting $k \to \infty$ we obtain (3.11), as desired. Since $u_0 \in B(v_0, \varepsilon/2)$, combining (3.11) with (3.7) gives

$$\begin{aligned}\langle \xi_0, v_0\rangle + \frac{\varepsilon}{2} \geq \langle \xi_0, u_0\rangle &\geq \sup\big\{\langle \xi_0, k_0(A_0 z + w)\rangle : z \in T_C(x_0) \cap B_{R^n},\\ &\qquad w \in \overline{\text{cone}}(f(x_0, p_0) + K) \cap B_{R^m}\big\}\\ &\geq \sup\{\langle \xi_0, v\rangle : v \in B(v_0, \varepsilon)\}\\ &= \langle \xi_0, v_0\rangle + \varepsilon,\end{aligned}$$

a contradiction. We have thus proved that there is $k_2 \geq k_1$ such that (3.9) holds for every $k \geq k_2$. Using (3.8) and (3.9) we have

$$\begin{aligned}B\left(v_0, \frac{\varepsilon}{2}\right) &\subset k_0\left(A_0\left[T_C(x_k) \cap B_{R^n}\right] + \left[\overline{\text{cone}}(f(x_k, p_k) + K) \cap B_{R^m}\right]\right)\\ &\subset k_0\big(A_k\left[T_C(x_k) \cap B_{R^n}\right] + (A_0 - A_k)\left[T_C(x_k) \cap B_{R^m}\right]\\ &\qquad + \left[\overline{\text{cone}}(f(x_k, p_k) + K) \cap B_{R^m}\right]\big)\\ &\subset k_0\Bigg(A_k\left[T_C(x_k) \cap B_{R^n}\right] + B\left(0, \frac{\varepsilon}{4}\right)\\ &\qquad + \left[\overline{\text{cone}}(f(x_k, p_k) + K) \cap B_{R^m}\right]\Bigg).\end{aligned}$$

This implies that

$$B\left(v_0, \frac{\varepsilon}{4}\right) \subset k_0\left(A_k\left[T_C(x_k) \cap B_{R^n}\right] + \left[\overline{\text{cone}}(f(x_k, p_k) + K) \cap B_{R^m}\right]\right). \tag{3.12}$$

Choose $k \geq k_2$ sufficiently large; we have $v_k \in B(v_0, \varepsilon/4)$. Then (3.12) yields

$$v_k \in k\left(A_k\left[T_C(x_k) \cap B_{R^n}\right] + \left[\overline{\text{cone}}(f(x_k, p_k) + K) \cap B_{R^m}\right]\right), \tag{3.13}$$

contrary to (3.3).

We now suppose that (3.5) is valid. By the regularity condition, we have (3.6), where A_0 is replaced by A_*. Then there exist $\varepsilon > 0$ and $k_0 > 1$ such that (3.7), where A_0 is replaced by A_*, holds. The relations (3.8)–(3.10) remain true provided that A_0 is replaced by A_* and A_k by $t_k A_k$. Then relation (3.12) has the form

$$B\left(v_0, \frac{\varepsilon}{2}\right) \subset k_0\left(t_k A_k\left[T_C(x_k) \cap B_{R^n}\right] + \left[\overline{\text{cone}}(f(x_k, p_k) + K) \cap B_{R^m}\right]\right)$$

for all $k \geq k_2$. By choosing $k \geq k_2$ sufficiently large so that $v_k \in B(v_0, \varepsilon/4)$ and $0 < t_k \leq 1$ we obtain (3.13), which contradicts (3.3).

The proof of the lemma will be completed if we can show that either (3.4) or (3.5) holds. This part of the proof is omitted because it is a routine repetition of the second part of the proof of Lemma 3.1 in [13], where M_k is replaced by A_k, y_k by (x'_k, p'_k), $F(y_k)$ by $J_1 f(x'_k, p'_k)$, and $F(0)$ by $J_1 f(x_0, p_0)$. The upper semicontinuity of $F(\cdot)$ at 0 is now replaced by the upper semicontinuity of $J_1 f$ at (x_0, p_0). □

The next theorem will be obtained by the proof scheme of Theorem 3.1 in [27]. Unlike the generalized Jacobians in the sense of Clarke, approximate Jacobians, in general, are noncompact nonconvex sets of linear operators. Thus, some technical novelties are to be introduced. One of our key tools will be the lopsided minimax theorem.

THEOREM 3.2 (solution stability). *If (1.1) is regular at x_0 and $\{f(x,p), P, p_0\}$ is an admissible perturbation of the system at x_0, then there exist neighborhoods U of p_0 and V of x_0 such that $G(p) \cap V$ is nonempty for every $p \in U$, and the multifunction $\widetilde{G}(\cdot) := G(\cdot) \cap V$ is lower semicontinuous on U.*

Proof. Since (1.1) is regular at x_0 and $\{f(x,p), P, p_0\}$ is an admissible perturbation of (1.1) at x_0, by Lemma 3.1 there exist $\gamma > 0$ and $\delta \in (0, \delta_*)$ such that (3.1) holds for every $x \in B(x_0, \delta) \cap C$, $p \in B(p_0, \delta) \cap P$, and A satisfying (3.2). Here and in what follows, $\delta_* > 0$ and U_* are the number and the neighborhood specified by condition (iv) of Definition 2.6. Fix a number $\lambda \in (0, \gamma^{-1})$. Since $0 \in f(x_0, p_0) + K$ and the multifunction $p \mapsto f(x_0, p) + K$ is lsc at p_0, there exists $\delta_1 \in (0, \delta)$ such that

$$\forall p \in B(p_0, \delta_1) \cap P \;\; \exists y_p \in f(x_0, p) + K \;\; \text{satisfying} \;\; \|y_p\| < \lambda\delta.$$

Let $U = B(p_0, \delta_1) \cap U_*$. For every $p \in U$, we consider the restriction of the function

$$\nu_p(x) := d(0, f(x,p) + K) = \inf\{\|f(x,p) + v\| : v \in K\}$$

on the compact set $B(x_0, \delta) \cap C$. It is easily seen that $v_p(\cdot)$ is a continuous function. We have

$$\nu_p(x_0) = d(0, f(x_0, p)) \le \|y_p\| \le \lambda\delta'$$

for some $\delta' \in (0, \delta)$. By the Ekeland principle [8], there exists $\overline{x} \in B(x_0, \delta) \cap C$ such that

$$\nu_p(\overline{x}) \le \nu_p(x_0), \quad \|\overline{x} - x_0\| \le \delta', \tag{3.14}$$

$$\nu_p(\overline{x}) \le \nu_p(x) + \lambda\|x - \overline{x}\| \quad \forall x \in B(x_0, \delta) \cap C. \tag{3.15}$$

From (3.14) it follows that $\overline{x} \in \text{int}B(x_0, \delta)$. We have $0 \in f(\overline{x}, p) + K$, i.e., $\nu_p(\overline{x}) = 0$. Indeed, suppose to the contrary that $\nu_p(\overline{x}) \ne 0$. Since $f(\overline{x}, p) + K$ is a nonempty closed convex set, there exists a unique $\overline{y} \in f(\overline{x}, p) + K$ such that

$$\|\overline{y}\| = d(0, f(\overline{x}, p) + K) = \inf\{\|f(x,p) + v\| : v \in K\}, \quad \overline{y} \ne 0.$$

By the standard optimality condition of convex optimization we have

$$\|\overline{y}\|^{-1}\overline{y} \in -(f(\overline{x}, p) + K)^*.$$

We put $\overline{\eta} = \|\overline{y}\|^{-1}\overline{y}$. Let $\overline{w} = \overline{y} - f(\overline{x}, p)$. We have $\overline{w} \in K$, so $\nu_p(x) \le \|f(x,p) + \overline{w}\|$ for every $x \in R^n$. Define

$$\psi(x) = \|f(x,p) + \overline{w}\| \quad \text{and} \quad \varphi(x) = \psi(x) + \lambda\|x - \overline{x}\|$$

for every $x \in R^n$. From (3.15) we deduce that

$$\varphi(\overline{x}) \le \varphi(x) \quad \forall x \in B(x_0, \delta) \cap C.$$

Since $\overline{x} \in \text{int}B(x_0, \delta)$, the last property implies that $\overline{x}$ is a local minimum point of φ on C. By Proposition 2.4,

$$\sup_{\eta \in \partial f(\overline{x})} \langle \eta, u \rangle \ge 0 \quad \forall u \in T_C(\overline{x}), \tag{3.16}$$

where $\partial\varphi(\overline{x})$ is a generalized subdifferential of φ at $\overline{x}$. According to the chain rule formulated in Proposition 2.2, for any $\varepsilon \in (0, \delta)$, the closure of the set

$$\overline{\eta} \circ [J_1 f(\overline{x}, p) + (J_1 f(\overline{x}, p))^{\varepsilon}_{\infty}]$$

is a generalized subdifferential of ψ at $\overline{x}$. Applying the formula for computing the generalized subdifferential of the sum of two functions (see [14, Proposition 2.2]) we deduce that the closure of the set

$$\{\overline{\eta} \circ A + \lambda\xi : A \in J_1 f(\overline{x}, p) + (J_1 f(\overline{x}, p))^{\varepsilon}_{\infty},\ \xi \in B_{R^n}\}$$

is a generalized subdifferential of φ at $\overline{x}$. Then the larger set

$$\partial\varphi(\overline{x}) := \{\overline{\eta} \circ A + \lambda\xi : A \in \overline{\mathrm{co}}\,(J_1 f(\overline{x}, p) + (J_1 f(\overline{x}, p))^{\varepsilon}_{\infty}),\ \xi \in B_{R^n}\}, \tag{3.17}$$

which is closed and convex, is also a generalized subdifferential of φ at $\overline{x}$. Let

$$Q = \overline{\mathrm{co}}\,(J_1 f(\overline{x}, p) + (J_1 f(\overline{x}, p))^{\varepsilon}_{\infty}), \quad D = T_C(\overline{x}) \cap B_{R^n}.$$

We now show that

$$-\gamma^{-1} \geq \sup_{A \in Q} \inf_{v \in D} \langle \overline{\eta}, Av \rangle. \tag{3.18}$$

Indeed, for any given $A \in Q$ we observe that A satisfies (3.2) because $(J_1 f(\overline{x}, p))^{\varepsilon}_{\infty} \subset (J_1 f(\overline{x}, p))^{\delta}_{\infty}$, $\overline{x} \in \mathrm{int} B(x_0, \delta)$ and $p \in B(p_0, \delta) \cap P$. By (3.1), there exists $v \in T_C(\bar{x}) \cap B_{R^n}$ and $w \in \overline{\mathrm{cone}}(f(\overline{x}, p) + K) \cap B_{R^m}$ such that

$$-\overline{\eta} = \gamma(Av + w).$$

Then

$$-1 = -\langle \overline{\eta}, \overline{\eta} \rangle = \gamma \langle \overline{\eta}, Av + w \rangle.$$

Since $\langle \overline{\eta}, w \rangle \geq 0$, it follows that $-\gamma^{-1} \geq \langle \overline{\eta}, Av \rangle$. We have thus shown that $-\gamma^{-1} \geq \inf_{v \in D} \langle \overline{\eta}, Av \rangle$. Since the last inequality holds for any $A \in Q$, we conclude that (3.18) is valid. We next show that

$$\inf_{v \in D} \sup_{A \in Q} \langle \overline{\eta}, Av \rangle \geq -\lambda. \tag{3.19}$$

Indeed, let $v \in D$ be given arbitrarily. For any $\varepsilon_1 > 0$, from (3.16) and (3.17) it follows that there exist $A \in Q$ and $\xi \in B_{R^n}$ such that

$$(\overline{\eta} \circ A)(v) + \lambda \langle \xi, v \rangle \geq -\varepsilon_1.$$

So

$$\langle \overline{\eta}, Av \rangle \geq -\lambda \langle \xi, v \rangle - \varepsilon_1 \geq -\lambda - \varepsilon_1.$$

Hence $\sup_{A \in Q} \langle \overline{\eta}, Av \rangle \geq -\lambda - \varepsilon_1$. Since ε_1 can be chosen arbitrarily small, we conclude that $\sup_{A \in Q} \langle \overline{\eta}, Av \rangle \geq -\lambda$, and hence (3.19) is true. By the lopsided minimax theorem [1, p. 319], we have

$$\sup_{v \in D} \inf_{A \in Q} \langle \overline{\eta}, -Av \rangle = \inf_{A \in Q} \sup_{v \in D} \langle \overline{\eta}, -Av \rangle.$$

Therefore

$$\inf_{v\in D}\sup_{A\in Q}\langle\overline{\eta},Av\rangle=\sup_{A\in Q}\inf_{v\in D}\langle\overline{\eta},Av\rangle.$$

Combining this with (3.18) and (3.19) we get the inequality $-\gamma^{-1}\geq-\lambda$, which contradicts the inclusion $\lambda\in(0,\gamma^{-1})$. We have thus proved that $0\in f(\overline{x},p)+K$, and so $\overline{x}\in G(p)$.

We set $V=\mathrm{int}B(x_0,\delta)$ and $\widetilde{G}(p)=G(p)\cap V$. From what has already been proved we can assert that

$$\widetilde{G}(p)\neq\emptyset\quad\forall p\in U.$$

We now prove that the multifunction $\widetilde{G}(\cdot)$ is lsc on U. Let $p\in U$ and $x\in\widetilde{G}(p)$ be given arbitrarily. Given any $\varepsilon>0$ we chose $\tau\in(0,\varepsilon)$ so that $B(x,\tau)\subset V$. Repeating the above procedure with (x,p) taking the place of (x_0,p_0), we find a neighborhood U' of p in P such that

$$\forall p'\in U'\ \ \exists x'\in B(x,\tau)\quad\text{satisfying } 0\in f(x',p')+K.$$

The last inclusion shows that $x'\in G(p')$. Since $B(x,\tau)\subset V\cap B(x,\varepsilon)$, we have $x'\in\widetilde{G}(p')\cap B(x,\varepsilon)$. From this it follows that $\widetilde{G}(\cdot)$ is lsc at p. □

Observe that Theorem 3.2 shows that if the inequality system is regular at a solution, then this solution is stable under admissible perturbations. This implication is also typical in most of the studies on stability and sensitivity of optimization problems and variational inequalities. From the conclusions of Theorems 3.4 and 3.5 below it also follows that the solution x_0 is stable under admissible perturbations.

Lemma 3.1 and the procedure to show that the point $\overline{x}$ found by the Ekeland principle satisfies the inclusion $0\in f(\overline{x},p)+K$ in the preceding proof will enable us to obtain metric regularity and the pseudo-Lipschitz property of $G(\cdot)$. The methods of proof remain the same as in the proofs of Theorems 3.2 and 3.3 in [27]. The technique of taking some limit in an expression given by the first assertion of the Ekeland principle is originally due to Aubin and Frankowska [2]. Dien and Yen (see [7, 26, 27]) showed that the technique is useful not only for proving the pseudo-Lipschitz property but also for proving the metric regularity of implicit multifunctions.

DEFINITION 3.3 (see Borwein [3]). *The implicit multifunction $G(\cdot)$ defined by the generalized inequality system* (1.2) *is said to be* metrically regular *at (p_0,x_0) if there exist a constant $\mu>0$ and neighborhoods U_1 of p_0 and V_1 of x_0 such that*

$$d(x,G(p))\leq\mu d(0,f(x,p)+K)\quad\forall p\in U_1,\ \forall x\in V_1\cap C. \tag{3.20}$$

Metric regularity of inverse multifunctions (see section 4) is a special case of the above notion of the metric regularity of implicit multifunctions.

THEOREM 3.4 (metric regularity). *If* (1.1) *is regular at x_0 and $\{f(x,p),P,p_0\}$ is an admissible perturbation of the system at x_0, then $G(\cdot)$ is metrically regular at (p_0,x_0).*

Proof. Let constants $\gamma,\ \delta$ and neighborhoods U of p_0, V of x_0 be defined as in the proof of Theorem 3.2. Since the multifunction $(x,p)\mapsto f(x,p)+K$ is lsc at (x_0,p_0) and $0\in f(x_0,p_0)+K$, there exist neighborhoods U_1 of p_0 and V_1 of x_0 such that

$$U_1\subset U,\quad V_1\subset B\left(x_0,\frac{\delta}{2}\right)$$

and

$$d(0, f(x,p)+K) < \frac{\delta}{2\gamma} \quad \forall p \in U_1,\ \forall x \in V_1. \tag{3.21}$$

We will prove the inequality in (3.20) for $\mu = \gamma$. Fix any $x \in V_1 \cap C$ and $p \in U_1$. We put $\alpha = d(0, f(x,p)+K)$. By (3.21), $\alpha < 2^{-1}\gamma^{-1}\delta$. Hence the interval $\left(2\delta^{-1}\alpha, \gamma^{-1}\right)$ is nonempty. Let $\tau \in \left(2\delta^{-1}\alpha, \gamma^{-1}\right)$. We consider the function

$$\nu_p(z) = d(0, f(z,p)+K) \quad (z \in R^n).$$

Fix any $\tau' \in (\tau, \gamma^{-1})$. We have

$$\nu_p(x) = \alpha < \tau^{-1}\alpha\tau'.$$

By the Ekeland principle, there exists $\overline{x} \in B(x_0, \delta) \cap C$ such that

$$\|\overline{x} - x\| \le \tau^{-1}\alpha,$$

$$\nu_p(\overline{x}) \le \nu_p(z) + \tau'\|z - \overline{x}\| \quad \forall z \in B(x_0, \delta) \cap C.$$

Then

$$\|\overline{x} - x_0\| \le \|\overline{x} - x\| + \|x - x_0\| < \tau^{-1}\alpha + 2^{-1}\delta < \delta.$$

Since $0 < \tau' < \gamma^{-1}$, the arguments in the first part of the proof of Theorem 3.2 show that $0 \in f(\overline{x}, p) + K$. Hence $\overline{x} \in G(p)$ and we have

$$d(x, G(p)) \le \|x - \overline{x}\| \le \tau^{-1}\alpha.$$

Letting $\tau \to \gamma^{-1}$ we get the estimation $d(x, G(p)) \le \gamma\alpha$, which can be written equivalently as

$$d(x, G(p)) \le \gamma d(0, f(x,p) + K).$$

The proof is complete. □

THEOREM 3.5 (pseudo-Lipschitz property). *In addition to the assumptions of Theorem* 3.2, *suppose that there exist* $\kappa > 0$ *and neighborhoods* U_0 *of* p_0 *in* P *and* V_0 *of* x_0 *such that*

$$\|f(x,p') - f(x,p)\| \le \kappa\|p' - p\| \quad \forall p, p' \in U_0,\ \forall x \in V_0. \tag{3.22}$$

Then the multifunction $G(\cdot)$ *is pseudo-Lipschitz at* (p_0, x_0).

Proof. Let γ, δ, U, and V be defined as in the proof of Theorem 3.2. We choose $\theta > 0$ as small as

$$B(x_0, \theta\kappa) \subset V \cap V_0, \quad B(p_0, \gamma^{-1}\theta) \cap P \subset U \cap U_0.$$

Let

$$\ell = 2\gamma\kappa, \quad \widetilde{U} = \mathrm{int}B(p_0, 8^{-1}\gamma^{-1}\theta) \cap P, \quad \widetilde{V} = \mathrm{int}B(x_0, 2^{-1}\theta\kappa).$$

We claim that

$$G(p) \cap \widetilde{V} \subset G(p') + \ell\|p - p'\|B_{R^n} \quad \forall p, p' \in \widetilde{U}.$$

To prove this, it suffices to show that for any $p, p' \in \widetilde{U}$ and $x \in G(p) \cap \widetilde{V}$ we have

$$d(x, G(p')) \leq \ell \|p - p'\|. \tag{3.23}$$

Since $\|p - p'\| < 4^{-1}\gamma^{-1}\theta$, there exists an ε satisfying

$$2\theta^{-1}\|p - p'\| < \varepsilon < 2^{-1}\gamma^{-1}. \tag{3.24}$$

Let

$$\varphi(z) = \nu_{p'}(z) + \varepsilon\|z - x\| \quad \forall z \in R^n,$$

where $\nu_{p'}(z) = d(0, f(z, p') + K)$. By (3.22), $\|f(x, p') - f(x, p)\| \leq \kappa\|p' - p\|$. Hence, if $w \in K$ is such that $\nu_p(x) = \|f(x, p) + w\| = 0$, then

$$\begin{aligned}\varphi(x) = \nu_{p'}(x) &= \nu_{p'}(x) - \nu_p(x)\\ &\leq \|f(x, p') + w\| - \|f(x, p) + w\|\\ &\leq \kappa\|p - p'\|.\end{aligned}$$

Combining this with (3.24) we get

$$\varphi(x) \leq 2^{-1}\kappa\varepsilon\theta.$$

Applying the Ekeland principle we find $\overline{x} \in B(x_0, \theta\kappa) \cap C$ such that

$$\varphi(\overline{x}) \leq \varphi(x), \quad \|\overline{x} - x\| \leq 2^{-1}\theta\kappa,$$

and

$$\varphi(\overline{x}) \leq \varphi(z) + \varepsilon\|z - \overline{x}\| \quad \forall z \in B(x_0, \theta\kappa) \cap C.$$

Therefore

$$\nu_{p'}(\overline{x}) + \varepsilon\|\overline{x} - x\| \leq \nu_{p'}(x), \tag{3.25}$$

$$\|\overline{x} - x\| \leq 2^{-1}\theta\kappa, \tag{3.26}$$

$$\nu_{p'}(\overline{x}) \leq \nu_{p'}(z) + 2\varepsilon\|z - \overline{x}\| \quad \forall z \in B(x_0, \theta\kappa) \cap C. \tag{3.27}$$

Since $x \in \mathrm{int}B(x_0, 2^{-1}\theta\kappa)$, (3.26) yields $\overline{x} \in \mathrm{int}B(x_0, \theta\kappa)$. Since $0 < \varepsilon < 2^{-1}\gamma^{-1}$, we have $2\varepsilon \in (0, \gamma^{-1})$. By a procedure similar to that in the proof of Theorem 3.2, from (3.27) we deduce that $0 \in f(\overline{x}, p') + K$, and hence $\overline{x} \in G(p')$. Inequality (3.25) shows that

$$\|\overline{x} - x\| \leq \varepsilon^{-1}\nu_{p'}(x) \leq \varepsilon^{-1}\kappa\|p - p'\|;$$

hence

$$d(x, G(p')) \leq \varepsilon^{-1}\kappa\|p - p'\|.$$

Due to (3.24), letting $\varepsilon \to 2^{-1}\gamma^{-1}$ from the last inequality we obtain (3.23). The proof is complete. □

If f and $f(\cdot,p)$ $(p \in P)$ are locally Lipschitz functions, then as $Jf(x)$ and $J_1 f(x,p)$ we can choose the Clarke generalized Jacobian of $f(\cdot)$ and $f(\cdot,p)$, respectively, at x. Hence Theorems 3.1–3.3 in [27] follow from the above implicit function theorems provided that C is closed and convex. (In [27] it is assumed only that C is a closed subset of R^n. In this case, $T_C(x)$ stands for the Clarke tangent cone.)

Let us consider a simple example showing that, in general, the metric regularity of implicit multifunctions does not imply the pseudo-Lipschitz property.

Example 3.6. Let $n = m = r = 1$, $C = R$, $K = \{0\}$, $f(x,p) = x(p+1) - p^{1/3}$ for all $x, p \in R$. Let $p_0 = 0$ and $x_0 = 0$. Then the map $p \mapsto G(p)$, where $G(p) = \{x \in C : 0 \in f(x,p) + K\}$, is metrically regular at (p_0, x_0), but it is not pseudo-Lipschitz at (p_0, x_0). It is easily verified that the assumptions of Theorem 3.4 are satisfied, while the assumptions of Theorem 3.5 are not.

Here is another simple example showing that for implicit multifunctions the pseudo-Lipschitz property does not imply the metric regularity.

Example 3.7. Let $n = m = r = 1$, $C = R$, $K = \{0\}$, $f(x,p) = x^3 - p^3$, $p_0 = 0$, and $x_0 = 0$. Since $G(p) = \{x \in C : 0 \in f(x,p) + K\} = \{p\}$ for every p, $G(\cdot)$ is pseudo-Lipschitz at (p_0, x_0). However, there does not exist any $\mu > 0$ such that

$$d(x, G(p)) \le \mu d(0, f(x,p) + K)$$

for all (x,p) in a neighborhood of (x_0, p_0). Indeed, since

$$d(x, G(p)) = |x - p| \quad \text{and} \quad d(0, f(x,p) + K) = |x^3 - p^3|,$$

such a constant μ cannot exist.

So, *for implicit multifunctions, both statements "the metric regularity implies the pseudo-Lipschitz property" and "the pseudo-Lipschitz property implies the metric regularity" are not true in general*. Meanwhile, it is well known that for inverse multifunctions the metric regularity is equivalent to the pseudo-Lipschitz property (see [4, 17, 22]).

Effective sufficient conditions for the pseudo-Lipschitz property of implicit multifunctions in term of coderivatives have been given in [18, Theorems 4.1 and 5.1] and [20, Theorems 5.1, 5.8, and 6.1]. The above remark shows that these conditions may not guarantee the metric regularity of the implicit multifunctions. Under some restrictive assumptions (see [18, Theorem 4.9]), the metric regularity of implicit multifunctions is equivalent to the pseudo-Lipschitz property.

Relationships between the concept of approximate Jacobian and the concept of coderivative are discussed in detail in [21]. In particular, it has been shown that if $f : R^n \to R^m$ is a continuous vector valued function and $Jf(\overline{x})$ is a representative for the coderivative mapping $D^* f(\overline{x})(\cdot) : R^n \rightrightarrows R^m$, that is, $Jf(\overline{x})$ is a nonempty closed subset of $L(R^n, R^m)$ and

$$\sup_{x^* \in D^* f(\overline{x})(y^*)} \langle x^*, u \rangle = \sup_{A \in Jf(\overline{x})} \langle A^* y^*, u \rangle \quad \forall u \in R^n,\ \forall y^* \in R^m,$$

then f is locally Lipschitz at $\overline{x}$ and $Jf(\overline{x})$ is an approximate Jacobian of f at $\overline{x}$. Example 3.5 in [21] shows that, for continuous real functions, the Mordukhovich subdifferential, even if it is nonempty, may not be an approximate Jacobian. Conversely, there exist many examples showing that nontrivial approximate subdifferentials exist, but the Mordukhovich subdifferential is empty. Therefore one can assert that, for continuous vector valued mappings, the concepts of coderivative and approximate Jacobian are not comparable.

We conclude this section with a simple example to which the abovementioned implicit function theorems in [18] and [20] cannot be applied, while Theorems 3.2–3.5 are applicable.

Example 3.8. Let $f(x) = x^{1/3}$ for every $x \in R$ and $f(x,p) = (p+1)x^{1/3} - p$ for every $(x,p) \in R \times R$. Let $P = R$, $C = R$, $K = \{0\}$, $p_0 = 0$, and $x_0 = 0$. For every $p \in (-1,1)$, the solution set $G(p)$ of (1.2) is given by the formula $G(p) = \{p^3/(p+1)^3\}$. It is clear that

$$J_1 f(x,p) = \begin{cases} [\alpha, +\infty) & \text{if } x = 0 \\ \{\frac{1}{3}(p+1)x^{-2/3}\} & \text{if } x \neq 0, \end{cases}$$

where $\alpha > 0$ is chosen arbitrarily, is an approximate mapping of $f(\cdot,p)$. It is easily verified that $\{f(x,p), P, p_0\}$ is an admissible perturbation of the system (1.1) at x_0 in the sense of Definition 2.6. Note that (1.1) is regular at x_0 in the sense of Definition 2.5. Since the assumptions of Theorem 3.2 are satisfied, there exist neighborhoods U of p_0 and V of p_0 such that $G(p) \cap V$ is nonempty for every $p \in U$, and the multifunction $\widetilde{G}(\cdot) := G(\cdot) \cap V$ is lower semicontinuous on U. By Theorem 3.4, $G(\cdot)$ is metrically regular at (P_0, x_0), that is, there exist constant $\mu > 0$ and neighborhoods U_1 of p_0 and V_1 of x_0 such that (3.20) is valid. Since (3.22) is satisfied for $\kappa = 2$, $U_0 = R$, and $V_0 = (-1,1)$, Theorem 3.5 asserts that the multifunction $G(\cdot)$ is pseudo-Lipschitz at (p_0, x_0).

4. Open mappings and Lagrange multipliers. In this section, we will derive from the results of the preceding section a general open mapping theorem, an inverse mapping theorem, Lagrange multiplier rules for cone-constrained optimization problems, and sufficient conditions for the continuity and the locally Lipschitz property of optimal value functions in parametric optimization problems with continuous data.

THEOREM 4.1 (open mapping theorem). *Let $C \subset R^n$ and $K \subset R^m$ be nonempty closed convex sets, $f : R^n \to R^m$ a continuous function. Let $x_0 \in C$. Assume that f admits an approximate Jacobian mapping Jf which is upper semicontinuous on a neighborhood of x_0, and each $A \in \overline{\text{co}} Jf(x_0) \cup \text{co}((Jf(x_0))_\infty \setminus \{0\})$ is surjective on C at x_0 w.r.t. $f(x_0) + K$. Then*

$$0 \in \text{int}(f(C) + K). \tag{4.1}$$

Proof. Let $P = R^m$, $p_0 = 0$, $f(x,p) = f(x) - p$ $(x \in R^n)$. It is clear that x_0 is a solution of the generalized inequality system

$$0 \in f(x) + K, \quad x \in C, \tag{4.2}$$

and $\{f(x,p), P, p_0\}$ is a perturbation of (4.2) at x_0. Since $Jf(\cdot,p) := Jf(\cdot)$ is an approximate Jacobian mapping of $f(\cdot,p)$ for every $p \in P$, from the hypothesis it follows that $\{f(x,p), P, p_0\}$ is an admissible perturbation of (4.2) at x_0 and (4.2) is regular at x_0. It is clear that for each $x \in R^n$, $f(x,\cdot)$ is a continuous function on P. Moreover,

$$\|f(x,p') - f(x,p)\| \leq \|p' - p\| \quad \forall p, p' \in P.$$

Applying Theorem 3.2 to the system (4.2) we conclude that there exist a neighborhood U of $p_0 = 0$ and a neighborhood V of x_0 such that $G(p) := \{x \in C : p \in f(x) + K\} \cap V$ is nonempty for all $p \in U$. This implies that $U \subset f(C \cap V) + K$, and hence (4.1) is valid. ∎

THEOREM 4.2 (inverse mapping theorem). *Under the assumptions of Theorem* 4.1, *the multifunction* $p \mapsto G(p)$, *where* $G(p) := \{x \in C : p \in f(x) + K\}$, *is pseudo-Lipschitz at* $(0, x_0)$, *and there exist* $\mu > 0$ *and neighborhoods* U *of* $0 \in R^m$ *and* V *of* x_0 *such that*

$$d(x, G(p)) \le \mu d(p, f(x) + K) \quad \forall p \in U,\ \forall x \in V,$$

that is, the inverse multifunction $G(\cdot)$ *is metrically regular at* $(0, x_0)$.

Proof. Let $P = R^m$, p_0, and $f(x, p)$ be defined as in the preceding proof. Applying Theorems 3.4 and 3.5 to the system (4.2) with the admissible perturbation $\{f(x, p), P, p_0\}$ we obtain the desired conclusions. □

Theorem 4.1 specializes to the open mapping theorem in [13] if $K = \{0\}$. Here we have to assume additionally that C is closed. Note that in the formulation of Theorem 3.3 in [13] one has to assume that the approximate Jacobian mapping $Jf(\cdot)$ is upper semicontinuous on a neighborhood of x_0, because in the proof of the theorem one uses the chain rule for an arbitrary point from a neighborhood of x_0. In [12], an open mapping theorem under the assumption that every $A \in \overline{\text{co}} Jf(x_0) \cup \text{co}((Jf(x_0))_\infty \setminus \{0\})$ is an invertible operator, has been established for the case where $C = R^n$ and $K = \{0\}$.

Theorem 4.2 describes some local properties of the inverse multifunction of the map $x \mapsto f(x) + K$ with respect to the constraint set C. In the case where $K = \{0\}$, we have thus proved that under the hypothesis that every $A \in \overline{\text{co}} Jf(x_0) \cup \text{co}((Jf(x_0))_\infty \setminus \{0\})$ is surjective on C at x_0, the inverse multifunction is metric regular at $(0, x_0)$ and pseudo-Lipschitz at $(f(x_0), x_0)$. As noted in the preceding section, the last two properties are equivalent. Metric regularity of inverse multifunctions has been considered by Borwein and Zhuang [4], Ioffe [9], Jourani [15], Mordukhovich [17, 18], Penot [22], and many other authors (see the references given in [15, 18]).

From Theorem 3.2 and the separation theorem we can easily derive necessary optimality conditions for the optimization problem

$$\text{Minimize } \varphi(x) \quad \text{subject to } x \in C,\ 0 \in f(x) + K, \tag{4.3}$$

where $\varphi : R^n \to R$ and $f : R^n \to R^m$ are continuous functions and $C \subset R^n$ and $K \subset R^m$ are nonempty closed convex sets. Suppose that φ admits a generalized subdifferential mapping $\partial\varphi(\cdot)$ and f admits an approximate Jacobian mapping $Jf(\cdot)$. In the case where $K = R^m$, if $x_0 \in C$ is a local solution of (4.3), then from Proposition 2.4 and the separation theorem it follows that

$$0 \in \overline{\text{co}} \partial f(x_0) + N_C(x_0).$$

We now consider the case where $K \ne R^m$.

THEOREM 4.3 (generalized Fritz–John conditions). *Let* $x_0 \in C$ *be a local solution of* (4.3). *Assume that the multifunctions* $\partial\varphi(\cdot)$ *and* $Jf(\cdot)$ *are upper semicontinuous on a neighborhood of* x_0. *Then there exist a nonzero vector* $(\lambda_0, \lambda) \in R_+ \times (-(f(x_0) + K)^*)$, *a vector* $x^* \in \overline{\text{co}} \partial\varphi(x_0) \cup \text{co}((\partial\varphi(x_0))_\infty \setminus \{0\})$, *and an operator* $A \in \overline{\text{co}} Jf(x_0) \cup \text{co}((Jf(x_0))_\infty \setminus \{0\})$ *such that*

$$0 \in \lambda_0 x^* + A^*(\lambda) + N_C(x_0). \tag{4.4}$$

If K *is a cone, then* $\lambda \in -K^*$ *and* $\langle \lambda, f(x_0) \rangle = 0$.

Proof. Let $x_0 \in C$ be a local solution of (4.3). Define $\widetilde{f}(x) = (\varphi(x) - \varphi(x_0), f(x))$ for all $x \in R^n$. It is easily seen that the formula $J\widetilde{f}(x) = \partial\varphi(x) \times Jf(x)$ $(x \in R^n)$

defines an approximate Jacobian mapping of $\widetilde{f}$. We claim that there exists $\widetilde{A} \in \overline{\text{co}}\widetilde{J}f(x_0) \cup \text{co}((\widetilde{J}f(x_0))_\infty \setminus \{0\})$ such that

$$0 \notin \text{int}\left\{\widetilde{A}(T_C(x_0)) + \widetilde{f}(x_0) + \widetilde{K}\right\}, \tag{4.5}$$

where $\widetilde{K} := R_+ \times K$. Indeed, we have

$$0 \in \widetilde{f}(x_0) + \widetilde{K}, \quad x_0 \in C.$$

Since x_0 is a local solution of (4.3), there cannot exist any sequence $\{x_k\} \subset C$ satisfying

$$0 \in \widetilde{f}(x_k) - q_k + \widetilde{K} \quad (\forall k),$$

where $q_k := (-1/k, 0) \in R \times R^m$. This implies that for any neighborhood V of x_0, the multifunction $q \mapsto \widetilde{G}(q) \cap V$, where

$$\widetilde{G}(q) = \{x \in C : 0 \in \widetilde{f}(x) - q + K\} \quad (\forall q = (\alpha, p) \in R \times R^m),$$

is not lsc at $q_0 := (0, 0)$. According to Theorem 3.2, the inequality system

$$0 \in \widetilde{f}(x) + \widetilde{K}, \quad x \in C,$$

cannot be regular at x_0. Thus there must exist $\widetilde{A} \in \overline{\text{co}}\widetilde{J}f(x_0) \cup \text{co}((\widetilde{J}f(x_0))_\infty \setminus \{0\})$ satisfying (4.5). By the separation theorem, from (4.5) we can assert that there exists a nonzero vector $(\lambda_0, \lambda) \in R \times R^m$ satisfying

$$\langle(\lambda_0, \lambda), w\rangle \geq 0 \quad \forall w \in \widetilde{A}(T_C(x_0)) + \widetilde{f}(x_0) + \widetilde{K}. \tag{4.6}$$

Let $\widetilde{A} = (x^*, A)$, where $x^* \in \overline{\text{co}}\partial\varphi(x_0) \cup \text{co}((\partial\varphi(x_0))_\infty \setminus \{0\})$ and $A \in \overline{\text{co}}Jf(x_0) \cup \text{co}((Jf(x_0))_\infty \setminus \{0\})$. From (4.6) it follows that $\lambda_0\alpha \geq 0$ for every $\alpha \geq 0$ and $\langle\lambda, w\rangle \geq 0$ for all $w \in f(x_0) + K$. So $(\lambda_0, \lambda) \in R_+ \times (-(f(x_0) + K)^*)$. Since $0 \in \widetilde{f}(x_0) + \widetilde{K}$, (4.6) also implies that

$$\langle(\lambda_0, \lambda), w\rangle \geq 0 \quad \forall w \in \widetilde{A}(T_C(x_0));$$

hence (4.4) is valid. If K is a cone, then the inclusion $\lambda \in -(f(x_0) + K)^*$ yields $\lambda \in -K^*$ and $\langle\lambda, f(x_0)\rangle = 0$. The proof is complete. □

If $C = R^n$ and $K = R^s_+ \times \{0\}_{m-s}$, where $0 \leq s \leq m$, then Theorem 4.3 just describes the multiplier rule stated in [13, Theorem 5.1]. Other Lagrange multiplier rules using the concept of approximate Jacobian have been obtained in [16, 25].

THEOREM 4.4 (generalized Kuhn–Tucker conditions). *Suppose that $x_0 \in C$ is a local solution of* (4.3). *Assume that the multifunctions $\partial\varphi(\cdot)$ and $Jf(\cdot)$ are upper semicontinuous on a neighborhood of x_0. If the regularity condition* (2.2) *is satisfied, then there exist $\lambda \in -(f(x_0) + K)^*$, $x^* \in \overline{\text{co}}\partial\varphi(x_0) \cup \text{co}((\partial\varphi(x_0))_\infty \setminus \{0\})$, and $A \in \overline{\text{co}}Jf(x_0) \cup \text{co}((Jf(x_0))_\infty \setminus \{0\})$ such that*

$$0 \in x^* + A^*(\lambda) + N_C(x_0). \tag{4.7}$$

If K is a cone, then $\lambda \in -K^*$ and $\langle\lambda, f(x_0)\rangle = 0$.

Proof. Let $x_0 \in C$ be a local solution of (4.3). According to Theorem 4.3, there exist a nonzero vector $(\lambda_0, \lambda) \in R_+ \times (-(f(x_0) + K)^*)$, $x^* \in \overline{\text{co}}\partial\varphi(x_0) \cup \text{co}((\partial\varphi(x_0))_\infty \setminus$

$\{0\}$), and $A \in \overline{\text{co}} Jf(x_0) \cup \text{co}((Jf(x_0))_\infty \setminus \{0\})$ such that (4.4) holds. If $\lambda_0 = 0$, then (4.4) and the inclusion $\lambda \in -(f(x_0) + K)^*$ imply that

$$\langle \lambda, Au + v \rangle \geq 0 \quad \forall (u, v) \in T_C(x_0) \times (f(x_0) + K).$$

Then (2.2) cannot hold, because $\lambda \neq 0$. Thus $\lambda_0 > 0$. Dividing both sides of (4.4) by λ_0 and replacing λ by $\lambda_0^{-1}\lambda$ if necessary, we can assert that (4.7) holds for some $\lambda \in -(f(x_0) + K)^*$. □

If φ and f are locally Lipschitz functions, then as $\partial\varphi(x)$ and $Jf(x)$, respectively, one can choose the Clarke generalized gradient $\partial^c\varphi(x)$ of φ at x and the Clarke generalized Jacobian $\partial^c f(x)$ of f at x. In this case, the above Lagrange multiplier rule is stated as follows.

COROLLARY 4.5. *Suppose that $x_0 \in C$ is a local solution of* (4.3). *Assume that φ and f are locally Lipschitz functions. If the regularity condition*

$$0 \in \text{int}(A[T_C(x_0)] + f(x_0) + K) \quad \forall A \in \partial^c f(x_0)$$

is satisfied, then there exist $\lambda \in -(f(x_0) + K)^*$, $x^* \in \partial^c\varphi(x)$, and $A \in \partial^c f(x_0)$ such that (4.7) *holds. If K is a cone, then $\lambda \in -K^*$ and $\langle \lambda, f(x_0) \rangle = 0$.*

In passing, observe that the method for deriving the Kuhn–Tucker conditions for smooth cone-constrained optimization problems from a stability theorem based on Robinson's concept of regularity was given in [6, p. 60].

From Theorems 3.2 and 3.5 it is easy to derive sufficient conditions for the continuity and locally Lipschitz properties of the optimal value function of an optimization problem involving continuous functions.

Let C, K, P be as above. Let $f : R^n \times P \to R^m$ and $\varphi : R^n \times P \to R$ be continuous functions. Suppose that for each $p \in P$, the function $f(\cdot, p)$ has an approximate Jacobian $J_1 f(\cdot, p)$ which is usc on R^n. Consider the parametric optimization problem

$$\text{Minimize } \varphi(x, p) \quad \text{subject to } x \in C, \ \ 0 \in f(x, p) + K \tag{4.8}$$

depending on the parameter $p \in P$. Denote by $G(p)$, $\nu(p)$, and $Q(p)$ the constraint set, the optimal value, and the solution set of (4.8).

PROPOSITION 4.6 (continuity of the optimal value function). *Suppose that*

(a) *there exists a compact set $\Sigma \subset R^n$ such that $Q(p) \cap \Sigma \neq \emptyset$ for every p in a neighborhood of p_0;*

(b) *there exists $x_0 \in Q(p_0) \cap \Sigma$ such that the map $(x, p) \mapsto J_1 f(x, p)$ is upper semicontinuous at (x_0, p_0) and*

$$\begin{aligned} &0 \in \text{int}\{A[T_C(x_0)] + f(x_0, p_0) + K\} \\ &\quad \forall A \in \overline{\text{co}} J_1 f(x_0, p_0) \cup \text{co}((J_1(f(x_0, p_0))_\infty \setminus \{0\})). \end{aligned} \tag{4.9}$$

Then, ν is continuous at p_0.

The proof of this proposition is omitted because it is similar to the proof of Theorem 4.1 in [27]. Instead of using an implicit function with the Clarke generalized Jacobian, one can use Theorem 3.2.

PROPOSITION 4.7 (locally Lipschitz property of the optimal value function). *Let φ be locally Lipschitz on $R^n \times P$. Assume the fulfillment of* (a) *and the following condition:*

(c) *for each $x_0 \in Q(p_0) \cap \Sigma$, the multifunction $(x, p) \mapsto J_1 f(x, p)$ is upper semicontinuous at (x_0, p_0), and there exist $\kappa > 0$ and neighborhoods U_0 of p_0 in P and V_0 of x_0 such that* (3.22) *is valid.*

Then, ν is locally Lipschitz at p_0.

For proving this proposition, it suffices to use Theorem 3.5 and follow the scheme of the proof of Theorem 4.2 in [27].

The following statement describes a typical situation where condition (a) is satisfied.

PROPOSITION 4.8. *Suppose that*

(d) *there exists $x_0 \in Q(p_0)$ such that the multifunction $(x,p) \to J_1 f(x,p)$ is upper semicontinuous at (x_0,p_0) and condition* (4.9) *holds.*

If

$$\begin{array}{ll} \textit{either} & \liminf_{\|x\|\to+\infty;\ p\to p_0} \varphi(x,p) > \varphi(x_0,p_0), \\ \textit{or} & \lim_{\|x\|\to+\infty;\ p\to p_0} \varphi(x,p) = +\infty, \end{array}$$

then condition (a) *is fulfilled.*

This proposition can be proved similarly as Theorem 4.3 in [27].

Following the schemes developed by Borwein [3], one can derive from Theorems 3.2–3.5 some formulas for tangent cones of closed sets and for directional derivatives of the optimal value function.

Appendix. This appendix contains the proof of Proposition 2.2. For the benefit of the reader we present the details here. We need the following lemma for this proof.

LEMMA A.1 (see [12]). *Let $F : R^n \to 2^{R^s}$ be a multifunction that is upper semicontinuous at $x_0 \in R^n$. Let $t_i > 0$ converge to 0, $q_i \in \overline{\text{co}} F(x_0 + t_i B_{R^n})$ with $\lim_{i\to\infty} \|q_i\| = \infty$ and $\lim_{i\to\infty} q_i/\|q_i\| = q_*$ for some $q_* \in R^s$. Then $q_* \in (\text{co}\, F(x_0))_\infty$. Moreover, if $\text{co}(F(x_0))_\infty$ is pointed, then $q_* \in \text{co}(F(x_0))_\infty = (\text{co}F(x_0))_\infty$.*

Proof. By the upper semicontinuity of F at x_0, for every $\varepsilon > 0$, there is i_0 sufficiently large such that

$$F(x_0 + t_i B_{R^n}) \subset F(x_0) + \varepsilon B_{R^s} \quad \text{for all} \ \ i \geq i_0.$$

Hence

$$q_i \in \overline{\text{co}}(F(x_0) + \varepsilon B_{R^s}) \subset \text{co}(F(x_0) + \varepsilon B_{R^s}) + \varepsilon B_{R^s} \quad \text{for all} \ \ i \geq i_0.$$

Consequently,

$$\begin{aligned} q_* \in [\text{co}(F(x_0) + \varepsilon B_{R^s}) + \varepsilon B(0,1)]_\infty &\subset [\text{co}(F(x_0) + \varepsilon B_{R^s})]_\infty \\ &\subset (\text{co}F(x_0))_\infty. \end{aligned}$$

The inclusion $\text{co}(F(x_0))_\infty \subset (\text{co}F(x_0))_\infty$ always holds because $F(x_0) \subset \text{co}F(x_0)$ and $(\text{co}F(x_0))_\infty$ is a closed convex cone. We now prove the reverse inclusion. Let $p \in (\text{co}F(x_0))_\infty$, $p \neq 0$. By Caratheodory's theorem, there exist convex combinations $p_i = \sum_{j=1}^{s+1} \lambda_{ij} p_{ij}$ with $\lambda_{ij} \geq 0$, $p_{ij} \in F(x_0)$, and $\sum_{j=1}^{s+1} \lambda_{ij} = 1$ such that

$$p/\|p\| = \lim_{i\to\infty} p_i/\|p_i\| \quad \text{and} \quad \lim_{i\to\infty} \|p_i\| = \infty.$$

Without loss of generality we can assume that $\lim_{i\to\infty} \lambda_{ij} = \lambda_j \geq 0$ for $j = 1, \ldots, s+1$ and $\sum_{j=1}^{s+1} \lambda_j = 1$. For every j, consider the sequence $\{\lambda_{ij} p_{ij}/\|p_i\|\}_{i\geq 1}$. We claim that this sequence is bounded; hence we may assume that it converges to some $p_{0j} \in (F(x_0))_\infty$. Then $p = \sum_{j=1}^{s+1} p_{0j} \in \text{co}(F(x_0))_\infty$ as desired. To prove the claim we suppose to the contrary that $\{\lambda_{ij} p_{ij}/\|p_i\|\}_{i\geq 1}$ is unbounded. Let $a_{ij} = \lambda_{ij} p_{ij}/\|p_i\|$. By taking a subsequence if necessary, we can assume that

$$\|a_{ij_0}\| = \max\{\|a_{ij}\| : j = 1, \ldots, s+1\}$$

for every i. Hence $\lim_{i\to\infty} \|a_{ij_0}\| = \infty$. Since $p_i/\|p_i\| = \sum_{j=1}^{s+1} a_{ij}$, we have

$$0 = \lim_{i\to\infty} p_i/(\|p_i\|.\|a_{ij_0}\|) = \lim_{i\to\infty} \sum_{j=1}^{s+1} a_{ij}/\|a_{ij_0}\|.$$

Again we can assume that $\{a_{ij}/\|a_{ij_0}\|\}_{i\geq 0}$ converges to some $a_{0j} \in (F(x_0))_\infty$ for $j = 1, \ldots, s+1$ because these sequences are bounded. As $a_{0j_0} \neq 0$, the equality $0 = \sum_{j=1}^{s+1} a_{0j}$ shows that $\mathrm{co}(F(x_0))_\infty$ is not pointed, a contradiction. □

Using Lemma A.1, we now give a simplified form of the proof of Theorem 4.1 in [12].

Proof of Proposition 2.2. We wish to show that for every $u \in R^n$, $\alpha \in R$,

$$(\alpha g \circ f)^+(\overline{x}, u) \leq \sup_{q\in Q}(\alpha p_0 q u), \tag{A.1}$$

where $p_0 = \nabla g(f(\overline{x}))$ and $Q := Jf(\overline{x}) + (Jf(\overline{x}))^\varepsilon_\infty$. Since the case $u = 0$ or $\alpha = 0$ is obvious, we assume that $u \neq 0$ and $\alpha \neq 0$. Let $t_i > 0$ be a sequence of numbers converging to 0 such that

$$(\alpha g \circ f)^+(\overline{x}, u) = \lim_{i\to\infty} \frac{\alpha(g(f(\overline{x} + t_i u)) - g(f(\overline{x})))}{t_i}. \tag{A.2}$$

It follows from the mean value theorem [11, Corollary 5.1] that for each t_i, there exist $p_i \in \overline{\mathrm{co}}\, \nabla g[f(\overline{x}), f(\overline{x} + t_i u)]$ and $q_i \in \overline{\mathrm{co}}\, Jf[\overline{x}, \overline{x} + t_i u]$ such that

$$\begin{cases} f(\overline{x} + t_i u) - f(\overline{x}) = t_i q_i u, \\ g(f(\overline{x} + t_i u)) - g(f(\overline{x})) = p_i(f(\overline{x} + t_i u) - f(\overline{x})). \end{cases} \tag{A.3}$$

By our hypothesis, $\lim_{i\to\infty} p_i = p_0$. By taking a subsequence if necessary, we need to deal with two cases:

(a) $\{q_i\}$ converges to some q_0;

(b) $\lim_{i\to\infty} \|q_i\| = \infty$ with $\{q_i/\|q_i\|\}$ converging to some q_*.

It follows from (A.2) and (A.3) that

$$(\alpha g \circ f)^+(\overline{x}, u) = \lim_{i\to\infty} (\alpha p_i q_i u).$$

In case (a) we have $q_0 \in \overline{\mathrm{co}} Jf(\overline{x})$ by the upper semicontinuity of Jf at $\overline{x}$. Therefore

$$(\alpha g \circ f)^+(\overline{x}, u) = \alpha p_0 q_0 u \leq \sup_{q\in Q}(\alpha p_0 q u).$$

For case (b), by Lemma A.1, $q_* \in (\mathrm{co} Jf(\overline{x}))_\infty$. If $\mathrm{co}(Jf(\overline{x}))_\infty$ is not pointed, then it is easily seen that $\mathrm{co}(J(f(\overline{x}))^\varepsilon_\infty$ coincides with the whole space $L(R^n, R^m)$. Since $u \neq 0$, the last property and the assumption $p_0 \neq 0$ imply

$$\sup_{q\in Q}(\alpha p_0 q u) \geq \sup_{q\in L(R^n,R^m)} (\alpha p_0 q u) = +\infty;$$

hence (A.1) is valid. If the cone $\mathrm{co}(Jf(\overline{x}))_\infty$ is pointed, then by Lemma A.1 it contains q_*. Let $\beta := \alpha p_0 q_* u$. If $\beta > 0$, then from the fact that $\lambda q_* \in \mathrm{co}(Jf(\overline{x}))_\infty$ for all $\lambda \geq 0$ we deduce the following relation, which implies (A.1):

$$\sup_{q\in Q}(\alpha p_0 q u) \geq \sup_{q\in q_r + co(Jf(\overline{x}))^\varepsilon_\infty} (\alpha p_0 q u) \geq \limsup_{\lambda\to\infty} (\alpha p_0 (q_r + \lambda q_*) u) \geq +\infty,$$

where q_r is an arbitrary element of $Jf(\overline{x})$.

If $\beta < 0$, then for i sufficiently large, one has

$$\alpha p_i \frac{q_i}{\|q_i\|} u < \frac{\beta}{2} < 0.$$

Hence

$$(\alpha g \circ f)^+(\overline{x}, u) = \lim_{i\to\infty} (\alpha p_i q_i u) \leq \lim_{i\to\infty} \|q_i\|.\frac{\beta}{2} = -\infty.$$

This shows that (A.1) is true.

Now, suppose that $\beta = 0$. From the inclusion $q_* \in \mathrm{co}(Jf(\overline{x}))_\infty$ and the definition of the set $\mathrm{co}(Jf(\overline{x}))^\varepsilon_\infty = (\mathrm{co}(Jf(\overline{x}))_\infty)^\varepsilon$ it follows that $q_* \in \mathrm{int}(\mathrm{co}(Jf(\overline{x}))^\varepsilon_\infty)$. We claim that there exists $q_1 \in \mathrm{co}(Jf(\overline{x}))^\varepsilon_\infty$ such that

$$\alpha p_0 q_1 u > 0. \tag{A.4}$$

Indeed, consider the linear functional $\phi : L(R^n, R^m) \to R$ defined by setting $\phi(q) = \alpha p_0 q u$ for every $q \in L(R^n, R^m)$. If our claim is not true, then $\phi(q) \leq 0$ for every $q \in \mathrm{co}(Jf(\overline{x}))^\varepsilon_\infty$. Since $\phi(q_*) = \beta = 0$ and $q_* \in \mathrm{int}(\mathrm{co}(Jf(\overline{x}))^\varepsilon_\infty)$, we conclude that $\phi = 0$. As $u \neq 0$ and $p_0 \neq 0$, there exists $\overline{q} \in L(R^n, R^m)$ such that $\overline{q}u$ does not belong to the kernel of the functional p_0. Then we have $\alpha p_0 \overline{q} u \neq 0$, which is impossible because $\phi = 0$. Our claim has been proved. Fixing one element $q_r \in Jf(\overline{x})$, from (A.4) we deduce that

$$\sup_{q\in Q}(\alpha p_0 q u) \geq \sup_{q\in q_r + co(Jf(\overline{x}))^\varepsilon_\infty} (\alpha p_0 q u) \geq \lim_{\lambda\to\infty} (\alpha p_0 (q_r + \lambda q_1) u) \geq +\infty;$$

hence (A.1) holds. □

Acknowledgments. Helpful comments of the referees are gratefully acknowledged. The second author would like to thank Professor V. Jeyakumar for his hospitality at Sydney.

REFERENCES

[1] J.-P. Aubin and I. Ekeland, *Applied Nonlinear Analysis*, Wiley-Interscience, New York, 1984.
[2] J.-P. Aubin and H. Frankowska, *On inverse function theorem for set-valued maps*, J. Math. Pures Appl. (9), 66 (1987), pp. 71–89.
[3] J. M. Borwein, *Stability and regular points of inequality systems*, J. Optim. Theory Appl., 48 (1986), pp. 9–52.
[4] J. M. Borwein and D. M. Zhuang, *Verifiable necessary and sufficient conditions for regularity of set-valued and single-valued maps*, J. Math. Anal. Appl., 134 (1988), pp. 441–459.
[5] F. H. Clarke, *Optimization and Nonsmooth Analysis*, Wiley-Interscience, New York, 1983.
[6] B. D. Craven, *Mathematical Programming and Control Theory*, Chapman and Hall, London, 1978.
[7] P. H. Dien and N. D. Yen, *On implicit function theorems for set-valued maps and their applications to mathematical programming under inclusion constraints*, Appl. Math. Optim., 24 (1991), pp. 35–54.
[8] I. Ekeland, *On the variational principle*, J. Math. Anal. Appl., 47 (1974), pp. 324–353.
[9] A. D. Ioffe, *Codirectional compactness, metric regularity and subdifferential calculus*, Canadian Math. Soc. Conference Proc., 27 (2000), pp. 123–163.
[10] V. Jeyakumar and D. T. Luc, *Approximate Jacobian matrices for nonsmooth continuous maps and C^1-optimization*, SIAM J. Control Optim., 36 (1998), pp. 1815–1832.
[11] V. Jeyakumar and D. T. Luc, *Nonsmooth calculus, minimality, and monotonicity of convexificators*, J. Optim. Theory Appl., 101 (1999), pp. 599–621.

[12] V. Jeyakumar and D. T. Luc, *An open mapping theorem using unbounded generalized Jacobians*, Nonlinear Anal., 50 (2002), pp. 647–663.
[13] V. Jeyakumar and D. T. Luc, *Convex interior mapping theorems for continuous nonsmooth functions and optimization*, J. Nonlinear Convex Anal., 3 (2002), pp. 251–266.
[14] V. Jeyakumar and X. Wang, *Approximate Hessian matrices and second-order optimality conditions for nonlinear programming problems with C^1-data*, J. Austral. Math. Soc. Ser. B, 40 (1999), pp. 403–420.
[15] A. Jourani, *Hoffman's error bound, local controllability, and sensitivity analysis*, SIAM J. Control Optim., 38 (2000), pp. 947–970.
[16] D. T. Luc, *A multiplier rule for multiobjective programming problems with continuous data*, SIAM J. Optim., 13 (2002), pp. 168–178.
[17] B. S. Mordukhovich, *Complete characterization of openness, metric regularity, and Lipschitzian properties of multifunctions*, Trans. Amer. Math. Soc., 340 (1993), pp. 1–36.
[18] B. S. Mordukhovich, *Lipschitzian stability of constraint systems and generalized equations*, Nonlinear Anal., 22 (1994), pp. 173–206.
[19] B. S. Mordukhovich, *Generalized differential calculus for nonsmooth and set-valued mappings*, J. Math. Anal. Appl., 183 (1994), pp. 250–288.
[20] B. S. Mordukhovich, *Stability theory for parametric generalized equations and variational inequalities via nonsmooth analysis*, Trans. Amer. Math. Soc., 343 (1994), pp. 609–657.
[21] N. M. Nam and N. D. Yen, *Relationships between approximate Jacobians and coderivatives*, submitted.
[22] J.-P. Penot, *Metric regularity, openness, and Lipschitzian behavior of multifunctions*, Nonlinear Anal., 13 (1989), pp. 629–643.
[23] S. M. Robinson, *Stability theory for systems of inequalities, Part* II: *Differentiable nonlinear systems*, SIAM J. Numer. Anal., 13 (1976), pp. 497–513.
[24] R. T. Rockafellar and R. J.-B. Wets, *Variational Analysis*, Springer-Verlag, Berlin, Heidelberg, 1998.
[25] X. Wang and V. Jeyakumar, *A sharp Lagrange multiplier rule for nonsmooth mathematical programming problems involving equality constraints*, SIAM J. Optim., 10 (2000), pp. 1136–1148.
[26] N. D. Yen, *Implicit function theorems for set-valued maps*, Acta Math. Vietnam., 12 (1987), pp. 7–28.
[27] N. D. Yen, *Stability of the solution set of perturbed nonsmooth inequality systems and application*, J. Optim. Theory Appl., 93 (1997), pp. 199–225.

SIAM J. OPTIM.
Vol. 14, No. 4, pp. 1128–1139

CONTINUUM OF ZERO POINTS OF A MAPPING ON A COMPACT, CONVEX SET *

A. J. J. TALMAN† AND Y. YAMAMOTO‡

Abstract. Let X be a nonempty, compact and convex set in R^n and ϕ be an outer semicontinuous mapping from X to the collection of nonempty, compact convex subsets of R^n. We show that for any nonzero vector c in R^n there exists a set of stationary points of ϕ on X with respect to c connecting a point in the boundary of X at which $c^\top x$ is minimized on X to another point in the boundary of X at which $c^\top x$ is maximized on X. We provide several conditions on ϕ under which there exists a continuum of zero points of ϕ connecting two such points in the boundary of X, and an intersection result on a convex, compact set. An application to constrained equilibria is also given.

Key words. stationary point, continuum of zero points, variational inequality, intersection theorem, constrained equilibria

AMS subject classifications. 54H25, 54C60, 91B24, 91B50, 65K05

DOI. 10.1137/S1052623402415469

1. Introduction.

Whenever a mathematical model of some phenomenon is constructed either in engineering or in economics, the first question to ask is whether a solution to the model exists. A very powerful tool to this end is Brouwer's fixed point theorem; see Brouwer [2]. When the model is not a system of equations but a system of correspondences, Kakutani's fixed point theorem [12] is invoked. An alternative to fixed point theorems may be intersection theorems on polytopes, with the KKM theorem of Knaster, Kuratowski, and Mazurkiewicz [13] perhaps the most prominent example. A close relationship between fixed point theorems and intersection theorems is well known. Yet another alternative consists of results that claim the existence of solutions to variational inequality problems, the existence of stationary points, or the existence of zero points.

For certain models, it not only is important to know that there exists at least one solution, but one would like to show the existence of a continuum of solutions. In economics the existence of a continuum of solutions leads to difficulties in expectation formation of agents and as a consequence provides scope for endogenously generated fluctuations. A particular example comes from general equilibrium theory with price rigidities. In Herings [6], the existence of a continuum of zero points of the underlying constrained excess demand function on the unit cube is shown; see also [9]. The continuum contains all types of interesting equilibria. It is therefore important to have generally applicable tools that guarantee the existence of a continuum of solutions to a certain system of equations.

This leads us to the following problem: Given a point-to-set mapping $\varphi : X \rightrightarrows \mathbb{R}^n$,

*Received by the editors November 6, 2002; accepted for publication (in revised form) October 3, 2003; published electronically July 20, 2004.
http://www.siam.org/journals/siopt/14-4/41546.html

†Department of Econometrics and Operations Research and CentER, Tilburg University, P.O. Box 90153, 5000 LE Tilburg, The Netherlands (talman@uvt.nl). This author's research has been made possible by a fellowship of the Japanese Society of the Promotion of Sciences. This author thanks the society and also the University of Tsukuba for their hospitality.

‡Institute of Policy and Planning Sciences, University of Tsukuba, Tsukuba, Ibaraki 305-8573, Japan (yamamoto@sk.tsukuba.ac.jp). This author is supported by the Grant-in-Aid for Scientific Research B2 14380188 of the Ministry of Education, Culture, Sports, Science and Technology of Japan.

with X an arbitrary nonempty, convex, compact set, what reasonable conditions can guarantee the existence of a continuum of solutions to the system

$$0 \in \varphi(x)\,?$$

Our approach to prove the existence of a continuum of solutions is to show that there is a connected subset of solutions that links two distinct points in X, thereby guaranteeing the continuum.

It is well known that under certain conditions a point-to-set mapping defined on a nonempty, convex, compact set has a solution to the variational inequality or stationary point problem; e.g., see Eaves [4]. In this paper we generalize this problem and define a parametric stationary point problem with respect to some given nonzero vector $c \in \mathbb{R}^n$. We show that under the same conditions a point-to-set mapping defined on a nonempty, convex, compact set has a connected set of solutions to the parametric stationary point problem, called parameterized stationary points or stationary points with respect to the given vector c. The connected set contains two distinct points in the boundary of X. At one of these points the value $c^\top x$ is minimized for $x \in X$, while at the other point the value $c^\top x$ is maximized for $x \in X$. We give several conditions under which there exists a connected set of zero points linking these two distinct boundary points of X.

Intersection results with a continuum of intersection points can be found in Freidenfelds [5] on the unit simplex and Herings and Talman [8] on the unit cube. We provide sufficient conditions for a collection of closed sets covering a nonempty, convex, compact set to have a connected set of intersection points containing two distinct points in the boundary of the set.

The results in the paper generalize earlier results of Browder [3], Mas-Colell [14], and Herings, Talman, and Yang [10]. In the case of Browder's theorem, the compact, convex set is the Cartesian product of the unit interval $[0, 1]$ and a compact, convex set of one dimension less, while c is the unit vector with the one on the last position. Mas-Colell's result is an extension of Browder's result to deal with point-to-set mappings. Both Browder and Mas-Colell proved their results via a rather sophisticated machinery. Herings, Talman, and Yang [10] deal with a polytope. In Browder's theorem and in Mas-Colell's theorem a connected set of fixed points is obtained, connecting the levels 0 and 1, whereas the result on the polytope yields a connected set of zero points connecting two different faces of the polytope.

This paper is organized as follows. In section 2 we state the problem and give a general existence result. In section 3 we give sufficient conditions for the existence of a connected set of zero points of the mapping. Section 4 states the intersection result. Section 5 gives an application of the result to a pure exchange economy with restricted price set.

2. Continuum of parameterized stationary points. Let X be an arbitrary nonempty, convex, compact set of $\mathbb{R}^n$ and let c be an arbitrary nonzero vector in $\mathbb{R}^n$. Without loss of generality we assume that X is of full dimension and that $\min\{\, c^\top x \mid x \in X \,\} = 0$ and $\max\{\, c^\top x \mid x \in X \,\} = 1$. Let

$$\begin{aligned} H(\alpha) &= \{\, x \mid x \in \mathbb{R}^n;\ c^\top x = \alpha \,\} \quad \text{for } \alpha \in [0,1], \\ X(\alpha) &= X \cap H(\alpha) \quad \text{for } \alpha \in [0,1], \\ H &= \bigcup_{\alpha \in [0,1]} H(\alpha), \\ C &= \{\, \beta c \mid \beta \in \mathbb{R} \,\}. \end{aligned} \tag{2.1}$$

For notational convenience, we write $F : A \rightrightarrows B$ to denote that F is a point-to-set mapping from set A into the class of subsets of set B. The domain of $F : A \rightrightarrows B$, denoted by $\operatorname{dom} F$, is the set $\{ x \mid x \in A;\ F(x) \neq \emptyset \}$, and its graph, denoted by $\operatorname{gph} F$, is the set $\{ (x, y) \mid x \in A;\ y \in F(x) \}$.

DEFINITION 2.1. *For $F : \mathbb{R}^n \rightrightarrows \mathbb{R}^m$ and $\overline{x} \in \mathbb{R}^n$ let*

$$\limsup_{x \to \overline{x}} F(x) = \{ u \mid \exists x^\nu \to \overline{x}\ \exists u^\nu \to u \text{ such that } \forall \nu \in \mathbb{N}\ u^\nu \in F(x^\nu) \},$$
$$\liminf_{x \to \overline{x}} F(x) = \{ u \mid \forall x^\nu \to \overline{x}\ \exists u^\nu \to u \text{ such that } \exists N \in \mathbb{N}\ \forall \nu \geq N\ u^\nu \in F(x^\nu) \},$$

where $\mathbb{N}$ is the set of natural numbers. The sets $\limsup_{x \to \overline{x}} F(x)$ and $\liminf_{x \to \overline{x}} F(x)$ are called the outer limit and inner limit, respectively. The point-to-set mapping F is said to be outer semicontinuous at $\overline{x}$ if $\limsup_{x \to \overline{x}} F(x) \subseteq F(\overline{x})$ and inner semicontinuous at $\overline{x}$ if $\liminf_{x \to \overline{x}} F(x) \supseteq F(\overline{x})$. F is continuous at $\overline{x}$ if both conditions hold. A mapping F is outer or inner semicontinuous on $A \subseteq \mathbb{R}^n$ if F is outer or inner semicontinuous at every point of A.

We refer some basic results about the continuity in Rockafellar and Wets [15].

LEMMA 2.2 (see [15, Theorems 5.7 and 5.9]). *A point-to-set mapping $F : \mathbb{R}^n \rightrightarrows \mathbb{R}^m$ is outer semicontinuous on $\mathbb{R}^n$ if and only if its graph is a closed set of $\mathbb{R}^n \times \mathbb{R}^m$. It is inner semicontinuous on the interior of its domain if its graph is a convex set of $\mathbb{R}^n \times \mathbb{R}^m$.*

Since $X(\cdot)$ can be considered as a point-to-set mapping which assigns $\alpha \in [0, 1]$ to a convex subset $X(\alpha)$ of $\mathbb{R}^n$, we also use the symbol X to denote the mapping.

LEMMA 2.3. *The mapping $X : [0, 1] \rightrightarrows \mathbb{R}^n$ in (2.1) is continuous on $[0, 1]$.*

Proof. Clearly, the graph of the mapping X is a closed convex set in $\mathbb{R}^{n+1}$. Hence it is outer semicontinuous on $[0, 1]$ and inner semicontinuous on $(0, 1)$ by Lemma 2.2. We show that it is inner semicontinuous at $\alpha = 0$. Let y be an arbitrary point of $X(1)$. For a given point $x \in X(0)$ and a given sequence $\alpha^\nu \to 0$ define $x^\nu = \alpha^\nu y + (1 - \alpha^\nu)x$. Then clearly $x^\nu \in X(\alpha^\nu)$ and $x^\nu \to x$. This proves $X(0) \subseteq \liminf_{\alpha \to 0} X(\alpha)$. The case where $\alpha = 1$ is proved in exactly the same way. ∎

For $x \in \mathbb{R}^n$ let $S(x)$ be given by

$$S(x) = X(c^\top x);$$

then by Lemma 2.3 and, for example, Proposition 5.52 of [15] we have the following lemma.

LEMMA 2.4. *The mapping $S : H \rightrightarrows \mathbb{R}^n$ is a continuous point-to-set mapping on H.*

The normal cone is defined as follows.

DEFINITION 2.5. *Let $Y \subseteq \mathbb{R}^n$ and $y \in Y$. The normal cone $N_Y(y)$ of Y at y is the closed convex cone given by*

$$N_Y(y) = \{ v \in \mathbb{R}^n \mid (y' - y)^\top v \leq 0 \text{ for all } y' \in Y \}.$$

It should be noticed that convex sets enjoy the Clarke regularity of Definition 6.4 of [15]; i.e., regular normal cone and normal cone coincide. The former is denoted by $\hat{N}$ and the latter by N in [15], but we need not to distinguish them because of the convexity shared by the sets we consider in this paper.

In the following we denote $N_{S(x)}(x)$ simply by $N_S(x)$. We readily obtain the outer semicontinuity of N_S.

LEMMA 2.6. *The mapping $N_S : X \rightrightarrows \mathbb{R}^n$ is outer semicontinuous on X, and $N_S(x) = N_S(x) + C$ for each $x \in X$.*

Proof. The second assertion is straightforward from the definition of N_S and that $S(x) \subseteq H(c^\top x)$. We show that

$$\limsup_{x \to \overline{x}} N_S(x) \subseteq N_S(\overline{x})$$

holds for an arbitrary point $\overline{x} \in X$. Take a point $\overline{y}$ of $\limsup_{x \to \overline{x}} N_S(x)$. Then there are sequences $\{x^\nu\} \subseteq X$ and $\{y^\nu\} \subseteq \mathbb{R}^n$ such that $x^\nu \to \overline{x}$, $y^\nu \to \overline{y}$, and $y^\nu \in N_S(x^\nu)$ for each $\nu \in \mathbb{N}$. Let z be an arbitrary point of $S(\overline{x})$. Then by the continuity of S in Lemma 2.4 there is a sequence $\{z^\nu\}$ such that $z^\nu \to z$ and $z^\nu \in S(x^\nu)$ for each $\nu \in \mathbb{N}$. Note that $(y^\nu)^\top (z^\nu - x^\nu) \leq 0$. Taking the limit of this inequality yields $(\overline{y})^\top (z - \overline{x}) \leq 0$. Since z is an arbitrary point of $S(\overline{x})$, this inequality implies that $\overline{y} \in N_S(\overline{x})$. □

Remark. It is known that the normal cone mapping $N_Y : Y \rightrightarrows \mathbb{R}^n$ defined for a closed convex set Y is outer semicontinuous. See Proposition 6.6 of [15]. Furthermore, for two closed convex sets Y_1 and Y_2 Proposition 6.42 of [15] says that

$$N_{Y_1 \cap Y_2}(y) \supseteq N_{Y_1}(y) + N_{Y_2}(y)$$

holds for $y \in Y_1 \cap Y_2$, where $+$ means Minkowsky sum. If in addition Y_1 and Y_2 cannot be separated,

$$N_{Y_1 \cap Y_2}(y) = N_{Y_1}(y) + N_{Y_2}(y).$$

Applying this result to $S(x) = X \cap H(c^\top x)$ we see that

$$\begin{aligned} N_S(x) &\supseteq N_X(x) + C \quad \text{for } x \in X, \\ N_S(x) &= N_X(x) + C \quad \text{for } x \in X \setminus (X(0) \cup X(1)). \end{aligned}$$

See Figure 2.1.

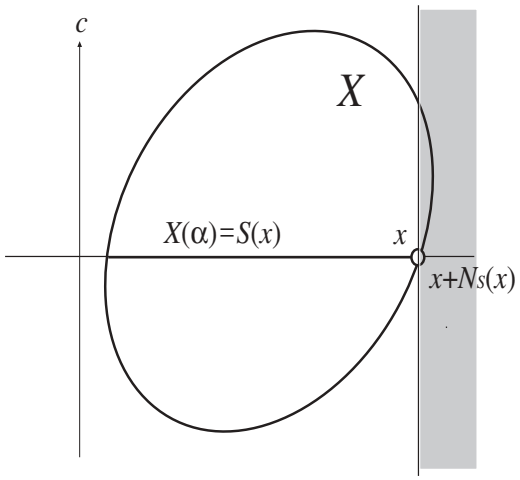


FIG. 2.1. *Section $X(\alpha) = S(x)$ and normal cone mapping N_S.*

For a subset Y of $\mathbb{R}^n$ and a mapping $\psi : Y \rightrightarrows \mathbb{R}^n$ we say that a point $y \in Y$ is a *stationary point* of ψ or a solution to the *variational inequality problem* for ψ on Y

when

$$\psi(y) \cap N_Y(y) \neq \emptyset.$$

In what follows, we consider an outer semicontinuous point-to-set mapping $\phi : X \rightrightarrows \mathbb{R}^n$ defined on X. We assume that the mapping ϕ is uniformly bounded, i.e., $\phi(X) = \bigcup_{x \in X} \phi(x)$ is bounded, and that for all $x \in X$ the set $\phi(x)$ is a nonempty, convex, compact subset of $\mathbb{R}^n$.

DEFINITION 2.7. *A point $x \in X$ is a stationary point of $\phi : X \rightrightarrows \mathbb{R}^n$ with respect to the nonzero vector c when x is a stationary point of ϕ on $S(x)$; i.e.,*

$$\phi(x) \cap N_S(x) \neq \emptyset.$$

A point $x \in X$ is a zero point of ϕ if $0 \in \phi(x)$.

We call the problem of finding a stationary point with respect to a nonzero vector a *parametric variational inequality problem*, and we call a solution to it a *parameterized stationary point*. It is known (e.g., see Eaves [4]) that for each $\alpha \in [0, 1]$ there exists a stationary point of ϕ on $X(\alpha)$; therefore there exists a stationary point x of ϕ on X with respect to c satisfying $c^\top x = \alpha$. Varying α from 0 to 1, we want to show that there exists a connected set of parameterized stationary points having a nonempty intersection with both $X(0)$ and $X(1)$, and we give conditions for the set of zero points of ϕ to connect these two sets.

For $x \in H$ let

$$p(x) = \operatorname{argmin} \{ \, \|x - y\|_2 \mid y \in S(x) \, \},$$

be the projection of x on $S(x)$, which is a singleton because of the convexity of $S(x)$, where $\| \cdot \|_2$ is the Euclidean norm. Clearly

$$x - p(x) \in N_S(p(x)) \quad \text{for each } x \in H. \tag{2.2}$$

LEMMA 2.8. *The function $p : H \to X$ is a continuous function.*

Proof. By Lemma 2.4 we have seen that $S : H \rightrightarrows \mathbb{R}^n$ is continuous. Applying Corollary 8.1 of Hogan [11] or Theorem 6, Section 1.2 of Aubin and Cellina [1] yields the continuity of p on H. □

Using the results above we are ready to prove the main result.

THEOREM 2.9. *Let X be a full-dimensional, compact, convex set in $\mathbb{R}^n$, let c be an arbitrary nonzero vector in $\mathbb{R}^n$, and let $\phi : X \rightrightarrows \mathbb{R}^n$ be an outer semicontinuous, uniformly bounded, nonempty, convex, compact-valued point-to-set mapping. Then there exists a connected set L of stationary points of ϕ on X with respect to c such that $L \cap X(0) \neq \emptyset$ and $L \cap X(1) \neq \emptyset$.*

Proof. Let r be the orthogonal projection from $\mathbb{R}^n$ onto $H(0)$; i.e., $r(x) = x - (c^\top x / c^\top c)c$. Since X is bounded and ϕ is uniformly bounded, the set $r(X + \phi(X)) = \{ \, y \mid y = r(x + f) \text{ for some } x \in X \text{ and } f \in \phi(X) \, \}$ is a bounded set in $H(0)$. Let D be a compact, convex subset of $H(0)$ containing $r(X + \phi(X))$ in its relative interior and let the mapping $\psi : D \times [0, 1] \rightrightarrows \mathbb{R}^n$ be defined by

$$\psi(y, \alpha) = r\Big(p(y + \alpha c) + \phi\big(p(y + \alpha c)\big)\Big). \tag{2.3}$$

Owing to the continuity of p and r and the outer semicontinuity of ϕ we yield the outer semicontinuity of ψ on $D \times [0, 1]$. With D being a nonempty, convex, compact

set, it follows from Mas-Colell [14] that there exists a connected set L' in $D \times [0,1]$ of fixed points of ψ satisfying $L' \cap (D \times \{0\}) \neq \emptyset$ and $L' \cap (D \times \{1\}) \neq \emptyset$, where $(y, \alpha) \in D \times [0,1]$ is said to be a fixed point of ψ if $y \in \psi(y, \alpha)$.

Let $y \in \psi(y, \alpha)$. Denoting $y+\alpha c$ by z, we have $y \in r(p(z)+\phi(p(z)))$ or equivalently $z-p(z)+(\beta-\alpha)c \in \phi(p(z))$ for some $\beta \in \mathbb{R}$. We also have $z-p(z)+(\beta-\alpha)c \in N_S(p(z))$ by (2.2) and Lemma 2.6. Therefore $\phi(p(z)) \cap N_S(p(z)) \neq \emptyset$, meaning that $x = p(y+\alpha c)$ is a stationary point of ϕ on X with respect to c. Finally, let the set L be defined by

$$L = \{\, x \mid x = p(y + \alpha c) \text{ for some } (y, \alpha) \in L' \,\}.$$

Since p is continuous and L' is connected, so is L. Moreover, $L' \cap (D \times \{0\}) \neq \emptyset$ implies $L \cap X(0) \neq \emptyset$ and $L' \cap (D \times \{1\}) \neq \emptyset$ implies $L \cap X(1) \neq \emptyset$. □

The theorem says that for any given nonzero vector c and any Kakutani-type point-to-set mapping on a full-dimensional compact, convex set, the set of stationary points with respect to c connects the pair of extreme sets $X(0)$ and $X(1)$.

3. Continuum of zero points. In this section we give sufficient conditions under which there exists in X a connected set of zero points of the mapping ϕ connecting $X(0)$ and $X(1)$.

THEOREM 3.1. *Let X, ϕ, and c satisfy the conditions of Theorem* 2.9. *If for each $x \in X$*

$$\phi(x) \cap N_S(x) = \emptyset \text{ or } \phi(x) \text{ contains } 0, \tag{3.1}$$

then there exists a connected set L of zero points of ϕ in X such that $L \cap X(0) \neq \emptyset$ and $L \cap X(1) \neq \emptyset$.

Proof. Let L be the connected set of parameterized stationary points in Theorem 2.9. For each point $x \in L$ we have $\phi(x) \cap N_S(x) \neq \emptyset$, which means that $0 \in \phi(x)$ by the above assumption. Therefore all elements of L are zero points of ϕ. □

The condition in the theorem says that at any $x \in X$ no nonzero element of the image $\phi(x)$ is allowed to lie in the normal cone $N_S(x)$, unless x is a zero point. Although the condition itself is rather weak, it has to hold for every element in $\phi(x)$. A sufficient condition that is much stronger but has only to hold for at least one element of the image set uses the notion of tangent cone.

DEFINITION 3.2. *For $x \in X$ the outer limit*

$$\limsup_{\tau \searrow 0} \frac{S(x) - x}{\tau}$$

is called the tangent cone of $S(x)$ at x and is denoted by $T_S(x)$.

Let Y be a cone of $\mathbb{R}^n$. The *polar cone* of Y, which is denoted by Y^*, is defined by

$$\{\, z \mid z \in \mathbb{R}^n;\ \ y^\top z \leq 0 \text{ for all } y \in Y \,\}.$$

Because of the convexity of $S(x)$, the tangent cone $T_S(x)$ of $S(x)$ at x coincides with the polar cone of the normal cone $N_S(x)$.

LEMMA 3.3 (see [15, Proposition 6.5]). *$T_S(x) = N_S^*(x)$ for every $x \in X$.*

THEOREM 3.4. *Let X, ϕ and c satisfy the conditions of Theorem* 2.9. *If*

$$\phi(x) \cap T_S(x) \neq \emptyset \tag{3.2}$$

for every $x \in X$, there exists a connected set L of zero points of ϕ in X such that $L \cap X(0) \neq \emptyset$ and $L \cap X(1) \neq \emptyset$.

The proof of this theorem does not follow immediately from Theorem 2.9, because the mapping T_S may not be outer semicontinuous on X. Take a polytope as X, and it is seen that T_S is not outer semicontinuous at vertices of X. To prove the theorem, let B denote the closed unit ball $\{ x \mid x \in \mathbb{R}^n;\ \|x\|_2 \leq 1 \}$ of $\mathbb{R}^n$, and let

$$\overline{X}(\alpha) = X(\alpha) + (B \cap H(0)) \quad \text{for } \alpha \in [0,1],$$
$$\overline{X} = \bigcup_{\alpha \in [0,1]} \overline{X}(\alpha).$$

The set $\overline{X}(\alpha)$ is the unit neighborhood of $X(\alpha)$ restricted to $H(\alpha)$. See Figure 3.1. The relative interior and relative boundary of $\overline{X}(\alpha)$ are denoted by $\operatorname{int} \overline{X}(\alpha)$ and $\operatorname{bd} \overline{X}(\alpha)$, respectively. By construction, $X(\alpha) \subseteq \overline{X}(\alpha)$ for each $\alpha \in [0,1]$, and $\overline{X}$ is a full-dimensional compact convex subset of H.

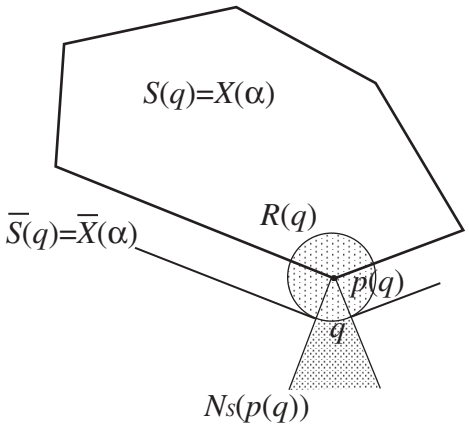


FIG. 3.1. *Unit neighborhood $\overline{X}(\alpha)$ and normal cone $N_S(p(q))$.*

For $q \in \overline{X}$ let us denote

$$R(q) = \{p(q)\} + (B \cap H(0)),$$
$$\overline{S}(q) = \overline{X}(c^\top q)$$

and employ the abbreviations

$$N_{\overline{S}(q)}(q) = N_{\overline{S}}(q),$$
$$N_{R(q)}(q) = N_R(q).$$

LEMMA 3.5. *For each $q \in \overline{X}$ it holds that*

$$N_{\overline{S}}(q) \subseteq N_R(q) \subseteq N_S(p(q)).$$

Proof. By construction $R(q)$ is a subset of $\overline{S}(q)$, which directly implies the first inclusion $N_{\overline{S}}(q) \subseteq N_R(q)$. When $q \in \operatorname{int} \overline{X}$, we have $N_R(q) = C$, which is clearly included in $N_S(p(q))$ by Lemma 2.6. When $q \in \operatorname{bd} \overline{X}$, we have $N_R(q) = \{ \mu(q - p(q)) \mid$

$\mu \geq 0\,\} + C$, which is again a subset of $N_S(p(q))$ by (2.2). See also Proposition 1, Section 0.6 of Aubin and Cellina [1]. □

LEMMA 3.6. *The point-to-set mapping $N_R^* : \overline{X} \rightrightarrows \mathbb{R}^n$ is outer semicontinuous on $\overline{X}$.*

Proof. As we have seen in the previous proof $N_R(q) = C$ when $q \in \operatorname{int} \overline{X}$ and $N_R(q) = \{\, \mu(q - p(q)) \mid \mu \geq 0\,\} + C$ when $q \in \operatorname{bd} \overline{X}$. Hence, $N_R^*(q) = \{\, y \mid c^\top y = 0\,\}$ when $q \in \operatorname{int} \overline{X}$ and $N_R^*(q) = \{\, y \mid c^\top y = 0;\ (q - p(q))^\top y \leq 0\,\}$ when $q \in \operatorname{bd} \overline{X}$. It is straightforward to see that the continuity of p yields the outer semicontinuity of N_R^*. □

For $q \in \overline{X}$ let

$$\psi(q) = \phi(p(q)) \cap T_S(p(q)) \tag{3.3}$$

and denote its closure mapping by Ψ; i.e., $\operatorname{gph} \Psi$ is the closure of $\operatorname{gph} \psi$. Then by Lemma 2.2 $\Psi : \overline{X} \rightrightarrows \mathbb{R}^n$ is an outer semicontinuous, uniformly bounded, nonempty, convex, compact-valued point-to-set mapping. Applying Theorem 2.9 to Ψ yields a connected set, say L', of stationary points of Ψ on $\overline{X}$ with respect to c having a nonempty intersection with both $\overline{X}(0)$ and $\overline{X}(1)$.

LEMMA 3.7. *Let $\hat{q}$ be a stationary point of Ψ on $\overline{X}$ with respect to c; i.e.,*

$$\Psi(\hat{q}) \cap N_{\overline{S}}(\hat{q}) \neq \emptyset. \tag{3.4}$$

Then $p(\hat{q})$ is a zero point of ϕ.

Proof. We start the proof by showing

$$\Psi(\hat{q}) \subseteq \phi(p(\hat{q})) \cap N_R^*(\hat{q}). \tag{3.5}$$

Let f be an arbitrary point of $\Psi(\hat{q})$. Then there are sequences $\{q^\nu\} \subseteq \overline{X}$ and $\{f^\nu\}$ such that $q^\nu \to \hat{q}$, $f^\nu \to f$ and $f^\nu \in \psi(q^\nu) = \phi(p(q^\nu)) \cap T_S(p(q^\nu))$ for each $\nu = 1, 2, \ldots$. Since ϕ is outer semicontinuous and p is continuous, we obtain $f \in \phi(p(\hat{q}))$. Applying Lemmas 3.3 and 3.5 we see $f^\nu \in T_S(p(q^\nu)) = N_S^*(p(q^\nu)) \subseteq N_R^*(q^\nu)$ for each ν, and hence by Lemma 3.6 $f \in N_R^*(\hat{q})$.

Furthermore, $N_{\overline{S}}(\hat{q}) \subseteq N_R(\hat{q})$ by Lemma 3.5. Thus we obtain from (3.4) and (3.5)

$$\emptyset \neq \Psi(\hat{q}) \cap N_{\overline{S}}(\hat{q}) \subseteq \phi(p(\hat{q})) \cap N_R^*(\hat{q}) \cap N_R(\hat{q}) \subseteq N_R^*(\hat{q}) \cap N_R(\hat{q}) = \{0\}.$$

This means that $0 \in \phi(p(\hat{q}))$. □

Proof of Theorem 3.4. Let L' be the connected set of stationary points of Ψ on $\overline{X}$ with respect to c and let $L = p(L')$. By the continuity of p and Lemma 3.7 L is a connected set of zero points of ϕ. Clearly, $L' \cap \overline{X}(0) \neq \emptyset$ implies $L \cap X(0) \neq \emptyset$ and $L' \cap \overline{X}(1) \neq \emptyset$ implies $L \cap X(1) \neq \emptyset$. □

The next theorem is a combination of the latter two theorems. It relaxes the rather strong condition of Theorem 3.4 to hold for at least one element of every image set and adds a condition for all elements in every image set, which is a weaker condition than that in Theorem 3.1.

THEOREM 3.8. *Let X, ϕ and c satisfy the conditions of Theorem* 2.9. *If for each $x \in X$ it holds that both*

$$\phi(x) \cap (T_S(x) + C) \neq \emptyset \tag{3.6}$$

and

$$\phi(x) \cap C = \emptyset \text{ or } \phi(x) \text{ contains } 0, \tag{3.7}$$

then there exists a connected set L of zero points of ϕ such that $L \cap X(0) \neq \emptyset$ and $L \cap X(1) \neq \emptyset$.

Proof. Let $r : \mathbb{R}^n \to H(0)$ denote the orthogonal projection onto the subspace $H(0)$, and let ψ be the composition $r\phi$. Then it is straightforward to see that ψ satisfies the conditions of Theorem 2.9 and (3.6) implies $\psi(x) \cap T_S(x) \neq \emptyset$; i.e., ψ satisfies the conditions of Theorem 3.4. Therefore we have a connected set of zero points of $\psi = r\phi$ in X having a nonempty intersection with $X(0)$ and $X(1)$. From (3.7) we readily see that every zero point of ψ is a zero point of ϕ. □

4. Intersection theorem. Suppose we are given k vectors $d^1, d^2, \ldots, d^k$ of $\mathbb{R}^n$, and k closed subsets $D^1, D^2, \ldots, D^k$ of X in $\mathbb{R}^n$ that cover X; i.e., $\bigcup_{j \in K} D^j = X$, where $K = \{1, 2, \ldots, k\}$. For each $x \in X$ let $d(x)$ denote the convex hull of $\{\, d^j \mid j \in K;\ x \in D^j \,\}$. We say that x is an *intersection point* with respect to a nonzero vector c in $\mathbb{R}^n$ when

$$d(x) \cap C \neq \emptyset,$$

or equivalently the orthogonal projection $r(d(x))$ of $d(x)$ onto $H(0)$ contains the origin.

THEOREM 4.1. *Let $\{\, D^j \mid j \in K \,\}$ be a closed covering of a full-dimensional, convex, compact set X in $\mathbb{R}^n$ and let $\{\, d^j \mid j \in K \,\}$ be a set of vectors in $\mathbb{R}^n$. Then with respect to any nonzero vector $c \in \mathbb{R}^n$, there exists a connected set L of intersection points in X satisfying $L \cap X(0) \neq \emptyset$ and $L \cap X(1) \neq \emptyset$ if one of the following two conditions is satisfied:*

1. *for every $x \in X$, $d(x) \cap N_S(x) = \emptyset$ or intersects C;*
2. *for every $x \in X$, $r(d(x)) \cap T_S(x) \neq \emptyset$.*

Proof. The proof follows from the fact that if we define ϕ be the composition rd, i.e., $\phi(x) = r(d(x))$ for each $x \in X$, the mapping ϕ satisfies the conditions of either Theorem 3.1 or 3.4 and, therefore, there exists a connected set L of zero points of ϕ in X having a nonempty intersection with both $X(0)$ and $X(1)$. Clearly, a zero point of ϕ is an intersection point in X. □

5. Application. The results in this paper will be used to show the existence of a connected set of constrained equilibria in a pure exchange economy with restricted price set. Let there be n commodities and m consumers in the economy. Consumer i initially owns $w^i = (w^i_1, \ldots, w^i_n)^\top \in \mathbb{R}^n_+$, where w^i_j is his endowment of commodity j, and $\mathbb{R}^n_+ = \{\, x \in \mathbb{R}^n \mid x \geq 0 \,\}$. Here we assume that $\sum_{i=1}^m w^i$ is a positive vector. Preference of consumer i on the commodity space $\mathbb{R}^n_+$ is represented by a continuous, strictly monotone, and strictly quasi-concave utility function u^i, where $u^i(y)$ denotes the utility level of consumer i when he consumes the commodity vector $y \in \mathbb{R}^n_+$. Given a price vector $\pi = (\pi_1, \ldots, \pi_n)^\top \in \mathbb{R}^n_+ \setminus \{0\}$ with π_j the price of commodity j, each consumer i maximizes his utility u^i over his budget set

$$B^i(\pi) = \{\, y \in \mathbb{R}^n_+ \mid \pi^\top y \leq \pi^\top w^i \,\}.$$

The solution $y^i(\pi)$, called the demand of consumer i at price vector π, is continuous and homogeneous of degree zero in π. Letting $z(\pi) = \sum_{i=1}^m (y^i(\pi) - w^i)$ denote the total excess demand at price vector π, the function $z : \mathbb{R}^n_+ \setminus \{0\} \to \mathbb{R}^n$ satisfies continuity, homogeneity of degree zero, and Walras's law, that is, $\pi^\top z(\pi) = 0$ at all π. When the excess demand is zero for all commodities at price vector π^*, i.e., $z(\pi^*) = 0$, an equilibrium is obtained and exchange of commodities between the consumers can take place. Such an equilibrium price vector always exists, and due to

the homogeneity of degree zero of z, it holds that if π^* is an equilibrium price vector, then $\lambda\pi^*$ is also an equilibrium price vector for any $\lambda > 0$. Hence, there exists a continuum $\{\, \lambda\pi^* \mid \lambda > 0 \,\}$ of equilibria.

When the set of feasible prices is restricted (e.g., minimum wage, price indexation, maximum commodity price) to some smaller set, the latter set may not contain an equilibrium price vector. To restore the equilibrium individual demand or supply could be rationed (e.g., quota); i.e., net demand and net supply of each consumer is constrained. When commodities are being rationed separately, a connected set of constrained equilibria is shown to exist in Herings, van der Laan, and Talman [7]. Instead of rationing commodities one by one, one may also constrain excess demand by one constraint for every consumer. Schalk [16] showed in this way that a constrained equilibrium always exist. We will now show that, in general, a connected set of such equilibria exists.

Let P denote the set of feasible prices. For simplicity we assume that P is a full-dimensional convex and compact subset of $\mathbb{R}^n_+ \setminus \{0\}$. A natural choice for the nonzero vector c is the total initial endowment vector $w = \sum_{i=1}^m w^i$. We define, in the same way as before, $\alpha_0 = \min\{\, w^\top\pi \mid \pi \in P \,\}$, $\alpha_1 = \max\{\, w^\top\pi \mid \pi \in P \,\}$, and $P(\alpha) = \{\, \pi \in P \mid w^\top\pi = \alpha \,\}$ for $\alpha \in [\alpha_0, \alpha_1]$. To determine the constraints on the individual excess demands, we extend the set $P(\alpha)$ to the set $X(\alpha)$, and then P to X by

$$X(\alpha) = P(\alpha) + (B \cap H(0)) \quad \text{and}$$
$$X = \bigcup_{\alpha\in[\alpha_0,\alpha_1]} X(\alpha),$$

or equivalently

$$X = \{\, x \in \mathbb{R}^n \mid \|x - \pi\|_2 \le 1 \text{ for some } \pi \in P \text{ with } w^\top x = w^\top\pi \,\},$$

where B is the closed unit ball of $\mathbb{R}^n$ and $H(0) = \{\, x \in \mathbb{R}^n \mid w^\top x = 0 \,\}$. Note that X is a full-dimensional compact and convex set in $\mathbb{R}^n$. Let $p(x)$ denote the projection of a point $x \in X$ on $P(w^\top x)$. Clearly, p is a continuous function on X. Then, for every $x \in X$, we define the constrained budget set of consumer i by

$$B^i(x) = \{\, y \in \mathbb{R}^n_+ \mid p^\top(x)y \le p^\top(x)w^i;\ (x - p(x))^\top(y - w^i) \le 1 - \|x - p(x)\|_2 \,\}.$$

The vector $p(x) \in P$ is the price vector and the vector $x - p(x)$ is the constraint vector induced by $x \in X$. When $x \in P$, then $p(x) = x$ and no constraint on the budget set is needed. When $x \notin P$, then x cannot be a price vector and rationing will take place. As price vector, the point in $P(w^\top x)$ closest to x is taken and the difference $x - p(x)$ becomes the vector of rationing. Notice that $x - p(x)$ is an element of the normal cone of the set $P(w^\top x)$ at the point $p(x)$.

Since the constrained budget set B^i is a continuous mapping on X for $i = 1, \ldots, m$, the solution $y^i(x)$ to the optimization problem of maximizing utility u^i over $B^i(x)$ is a continuous function of x and satisfies the budget constraint.

The aggregated constrained excess demand function is defined by

$$z(x) = \sum_{i=1}^m (y^i(x) - w^i).$$

We see that this function z is continuous and satisfies Walras's law; i.e., $p^\top(x)z(x) = 0$ for all $x \in X$. In fact, if the equality $p^\top(x)y = p^\top(x)w^i$ does not hold at $y \in B^i(x)$, we

can choose $\epsilon > 0$ such that $y+\epsilon w$ satisfies the equality, and it holds by $w^\top(x-p(x)) = 0$ that

$$\begin{aligned}(x-p(x))^\top(y+\epsilon w - w^i) &= (x-p(x))^\top(y-w^i) + \epsilon(x-p(x))^\top w \\ &= (x-p(x))^\top(y-w^i) \\ &\le 1 - \|x-p(x)\|_2.\end{aligned}$$

Hence $y+\epsilon w$ remains in $B^i(x)$. Since $w > 0$ and the utility function u^i is strictly monotone, we obtain that $y^i(x)$ satisfies $p^\top(x)y^i(x) = p^\top(x)w^i$. Summing up this equality for $i = 1, \dots, m$ yields Walras's law.

Moreover, if x lies on the boundary of X, $\|x-p(x)\|_2 = 1$, and by the second constraint of $B^i(x)$

$$(x-p(x))^\top z(x) \le 0 \tag{5.1}$$

holds. The latter property guarantees that the function z satisfies the conditions of Theorem 3.1. We denote $X(w^\top x)$ by $S(x)$ as in the preceding sections.

THEOREM 5.1. *The constrained excess demand function z satisfies $z(x) \notin N_S(x)$ unless $z(x) = 0$.*

Proof. We suppose $z(x) \in N_S(x)$ and show that $z(x) = 0$.

When $x \in \operatorname{int} X$, $N_S(x) = \{\, \beta w \mid \beta \in \mathbb{R} \,\}$, implying $z(x) = \beta w$ for some $\beta \in \mathbb{R}$. From Walras's law it follows that

$$0 = p^\top(x)z(x) = \beta w^\top p(x).$$

Since $w^\top p(x) > 0$, we have $\beta = 0$ and so $z(x) = 0$.

Suppose now that $x \in \operatorname{bd} X$. Then

$$z(x) = \beta w + \gamma(x-p(x))$$

for some $\gamma \ge 0$ and $\beta \in \mathbb{R}$. Moreover, $\|x-p(x)\|_2 = 1$ and $w^\top(x-p(x)) = 0$. Hence by (5.1)

$$0 \ge (x-p(x))^\top z(x) = \beta w^\top(x-p(x)) + \gamma = \gamma \ge 0.$$

Therefore, $\gamma = 0$. From Walras's law it now follows again that also $\beta = 0$ and so $z(x) = 0$. ∎

From Theorem 3.1 it now follows that there exists a continuum of zero points of z in X connecting $X(\alpha_0)$ and $X(\alpha_1)$. This connected set induces a continuum of constrained price equilibria connecting $P(\alpha_0)$, the set of feasible prices minimizing the value of total initial endowment, and $P(\alpha_1)$, the set of feasible prices maximizing the value of total initial endowment.

REFERENCES

[1] J. P. AUBIN AND A. CELLINA, *Differential Inclusions*, Springer-Verlag, Berlin, 1984.

[2] L. E. J. BROUWER, *Über Abbildung von Mannigfaltigkeiten*, Mathematische Annalen, 71 (1912), pp. 97–115.

[3] F. E. BROWDER, *On continuity of fixed points under deformation of continuous mapping*, Summa Brasiliensis Mathematicae, 4 (1960), pp. 183–191.

[4] B. C. EAVES, *On the basic theory of complementarity*, Math. Program., 1 (1971), pp. 68–75.

[5] J. FREIDENFELDS, *A set intersection theorem and applications*, Math. Program., 7 (1974), pp. 199-211.

[6] P. J.-J. Herings, *On the existence of a continuum of constrained equilibria*, J. Math. Econom., 30 (1998), pp. 257–273.
[7] P. J.-J. Herings, G. van der Laan, and A. J. J. Talman, *Quantity Constrained Equilibria*, CentER discussion paper 2001-93, Tilburg University, Tilburg, The Netherlands, 2001.
[8] P. J.-J. Herings and A. J. J. Talman, *Intersection theorems with a continuum of intersection points*, J. Optim. Theory Appl., 96 (1998), pp. 311–335.
[9] P. J.-J. Herings, A. J. J. Talman, and Z. Yang, *The computation of a continuum of constrained equilibria*, Math. Oper. Res., 21 (1996), pp. 675–696.
[10] P. J.-J. Herings, A. J. J. Talman, and Z. Yang, *Variational inequality problems with a continuum of solutions: Existence and computation*, SIAM J. Control Optim., 39 (2001), pp. 1852–1873.
[11] W. W. Hogan, *Point-to-set maps in mathematical programming*, SIAM Rev., 15 (1973), pp. 591–603.
[12] S. Kakutani, *A generalization of Brouwer's fixed point theorem*, Duke Math. J., 8 (1941), pp. 457–459.
[13] B. Knaster, C. Kuratowski, and C. Mazurkiewicz, *Ein Beweis des Fixpunktsatzes für n-dimensionale Simplexe*, Fundamenta Mathematicae, 14 (1929), pp. 132–137.
[14] A. Mas-Colell, *A note on a theorem of F. Browder*, Math. Program., 6 (1974), pp. 229–233.
[15] R. T. Rockafellar and R. J.-B. Wets, *Variational Analysis*, Springer-Verlag, Berlin, 1998.
[16] S. Schalk, *Equilibrium Theory: A Salient Approach*, Ph.D. thesis, CentER, Tilburg University, Tilburg, The Netherlands, 1999.

SIAM J. OPTIM.
Vol. 14, No. 4, pp. 1140–1162

MULTIVARIATE NONNEGATIVE QUADRATIC MAPPINGS*

ZHI-QUAN LUO†, JOS F. STURM‡, AND SHUZHONG ZHANG§

Abstract. In this paper, we study several issues related to the characterization of specific classes of multivariate quadratic mappings that are nonnegative over a given domain, with nonnegativity defined by a prespecified conic order. In particular, we consider the set (cone) of nonnegative quadratic mappings, defined with respect to the positive semidefinite matrix cone, and study when it can be represented by linear matrix inequalities. We also discuss the applications of the results in robust optimization, especially the robust quadratic matrix inequalities and the robust linear programming models. In the latter application the implementational errors of the solution are taken into account, and the problem is formulated as a semidefinite program.

Key words. linear matrix inequalities, convex cone, robust optimization, biquadratic functions

AMS subject classifications. 15A48, 90C22

DOI. 10.1137/S1052623403421498

1. Introduction. Let $\mathcal{C} \subset \Re^n$ be a closed and pointed convex cone. We can define a natural notion of conic ordering as follows: For vectors $x, y \in \Re^n$, we say $x \succeq_{\mathcal{C}} y$ if and only if $x - y \in \mathcal{C}$. Thus, $x \in \Re^n$ is nonnegative if and only if $x \in \mathcal{C}$. In this paper, we will be primarily interested in the conic ordering induced by the cone of positive semidefinite matrices, which is a very popular subject of study thanks to the recently developed high performance interior methods for conic optimization.

In general, given a closed and pointed convex cone $\mathcal{C}$, we wish to derive efficiently verifiable conditions under which a multivariate nonlinear mapping is nonnegative over a given domain (typically a unit ball), where nonnegativity is defined with respect to $\mathcal{C}$. In [17], Sturm and Zhang studied the problem of representing all nonnegative (defined with respect to the cone of nonnegative reals $\Re_+$) quadratic functions over a given domain. They showed that it is possible to characterize the set of nonnegative quadratic functions over some specific domains, e.g., the intersection of an ellipsoid and a half-space. Moreover, the characterization is a necessary and sufficient condition in the form of linear matrix inequalities (LMIs) which is easy to verify. This type of easily computable necessary and sufficient condition is particularly useful in systems theory and robust optimization, where the problem data themselves may contain certain design variables to be optimized. In particular, using these LMI conditions, many robust control or minimax-type of robust optimization problems can be reformulated as semidefinite programming (SDP) problems, which can be efficiently solved using modern interior point methods.

The problems to be studied in this paper belong to the same category as SDP problems. In particular, we show that it is possible to characterize, by LMIs, when a certain type of nonlinear matrix inequality holds over a domain. The first case of

*Received by the editors January 20, 2003; accepted for publication (in revised form) November 5, 2003; published electronically July 20, 2004.
http://www.siam.org/journals/siopt/14-4/42149.html

†Department of Electrical and Computer Engineering, University of Minnesota, Minneapolis, MN 55455 (luozq@ece.umn.edu). The research of this author was supported by a grant from NSERC, by the Canada Research Chair Program, and by National Science Foundation grant DMS-0312416.

‡This author is deceased.

§Department of Systems Engineering and Engineering Management, The Chinese University of Hong Kong, Shatin, Hong Kong (zhang@se.cuhk.edu.hk). The research of this author was supported by Hong Kong RGC earmarked grants CUHK4233/01E and CUHK4174/03E.

this type involves quadratic matrix inequalities (QMIs), where the quadratic matrix function is assumed to take a specific form. We prove that it is possible to give an LMI description, in terms of the problem data (i.e., the coefficients of the QMIs), for the quadratic matrix function to be positive semidefinite for all variables satisfying either a spectral or Frobenius norm bound. In fact, our methodology works for general quadratic matrix functions as well. What we derive is an equivalent condition in the dual conic space. However, the membership verification problem of this dual condition is NP-hard in general. There are several special cases in which the membership verification boils down to checking a system of LMIs, thus verifiable in polynomial time. The first such case is when the variable is one-dimensional (the dimension of the matrix-valued mapping is arbitrary). Alternatively, if the dimension of the matrix mapping is 2×2 (the variable can be in any dimension), then we prove that the QMI can again be characterized by LMIs of the problem data. We also show that our results can be applied to robust optimization. Specifically, we show that the robust linear programming models, where the implementational errors of the solution are taken into account, can be formulated as SDP problems.

This paper is organized as follows. In section 2, we introduce the general conic framework and problem formulation. In section 3, we present several results concerning the representation of matrix-valued quadratic matrix functions which are nonnegative over a domain. The discussion is continued in section 4 for the general matrix-valued mappings. A characterization for the nonnegativity of the mapping, in terms of the input data, over a general domain, is presented in the same section. This characterization is further shown to reduce to an LMI system in several special cases when the underlying variable is one-dimensional—or two-dimensional in the homogeneous case. Similarly, and in fact equivalently, we obtain LMI characterizations when the underlying variable is n-dimensional, but the mapping is 2×2 matrix valued. Particular attention is given to the case where the domain is an n-dimensional unit ball. In section 5 we discuss the applications of our results in robust optimization.

The notation we use is fairly standard. Vectors are lowercase letters, and matrices are capital letters. The transpose is expressed by ^{T}. The set of $n \times n$ symmetric matrices is denoted by $\mathcal{S}^n$; the set of $n \times n$ positive (semi)definite matrices is denoted by ($\mathcal{S}^n_+$) $\mathcal{S}^n_{++}$. For two given matrices A and B, we use $A \succ B$ ($A \succeq B$) to indicate that $A - B$ is positive (semi)definite, $A \otimes B$ to indicate the Kronecker product between A and B, and $A \bullet B := \sum_{i,j} A_{ij} B_{ij} = \mathrm{Tr}(AB^T)$ to indicate the matrix inner-product. For a given matrix A, $\|A\|_F$ stands for its Frobenius norm, and $\|A\|_2$ stands for its spectrum norm. By cone $\{x \mid x \in S\}$ (span $\{x \mid x \in S\}$) we mean the convex cone (respectively, linear subspace) generated by the set S. The acronym SOC stands for the second order cone $\{(t, x) \in \Re^n \mid t \geq \|x\|\}$, and $\|\cdot\|$ represents the Euclidean norm. Given a Euclidean space $\mathcal{L}$ with an inner-product $X \bullet Y$ and a cone $\mathcal{K} \subseteq \mathcal{L}$, the dual cone $\mathcal{K}^*$ is defined as

$$\mathcal{K}^* = \{Y \in \mathcal{L} \mid X \bullet Y \geq 0 \text{ for all } X \in \mathcal{K}\}.$$

Since the choice of $\mathcal{L}$ can be ambiguous, we call $\mathcal{K}^*$ the dual cone of $\mathcal{K}$ in $\mathcal{L}$. Often, $\mathcal{L}$ is chosen as span($\mathcal{K}$).

2. Cones of nonnegative mappings. One fundamental problem in optimization is checking the membership with respect to a given cone. Any polynomial-time ϵ-approximation procedure for the membership problem will lead to a polynomial-time ϵ-approximation algorithm for optimizing a linear function over the cone intersected with some affine subspace; see [11] for a precise statement. Checking the membership

for the dual cone is equivalent to asking whether a linear function is nonnegative over the whole cone itself. In Sturm and Zhang [17], a problem of this nature is surveyed and investigated in detail. In particular, the authors studied the structure of all quadratic functions that are nonnegative over a certain domain D. Such functions are characterized by the well-known S-lemma in the special case where D is the level set of a quadratic function; see Polyak [16] for a good survey on the S-lemma and its relation to range convexity. As a consequence, the cone generated by all nonnegative quadratic functions over this domain can be described using LMIs; see [5]. Moreover, it is shown in [17] that if D either is the contour of a strictly convex quadratic function or is the intersection of the level set of a convex quadratic function with a half-space, then the cone generated by all nonnegative quadratic functions over this domain can again be described using LMIs (even though the S-procedure is inexact in the second case). A consequence of this result is that the robust quadratic inequality over D can be converted equivalently to a single LMI.

If we consider a general vector-valued mapping, then questions such as the one posed in [17] can be generally formulated as follows:

> Determine a finite convex representation for the cone
>
> $$\mathcal{K} = \{f : \Re^n \to \Re^m \mid f \in \mathcal{F},\ f(D) \subseteq \mathcal{C}\}$$
>
> where $\mathcal{F}$ is a certain vector space of functions, $D \subseteq \Re^n$ is a given domain, and $\mathcal{C} \subseteq \Re^m$ is a given closed convex cone.

Solutions to problems of this type are essential ingredients in robust optimization [4], since they allow conversion of semi-infinite constraints into finite convex ones. To appreciate the difficulty of these problems, let us quote a useful result from [4] as follows.

PROPOSITION 2.1. *Let $\mathcal{F}$ be the set of all affine linear mappings, D be a unit sphere, $\mathcal{C}$ be the cone of positive semidefinite matrices. Then, it is NP-complete to decide the membership problem for $\mathcal{K}$. More explicitly, for given symmetric matrices $A_0, A_1, \ldots, A_n$ of size $m \times m$, it is NP-complete to test whether the following implication holds:*

$$\sum_{i=1}^{n} x_i^2 \le 1 \Longrightarrow A_0 + \sum_{i=1}^{n} x_i A_i \succeq 0.$$

However, there exist positive results as well. It is known [7, 14] that if $\mathcal{F}$ is the set of polynomials of order no more than d, $D = \Re^1$, and $\mathcal{C}$ is the cone of positive semidefinite matrices, then there is a polynomial reduction of $\mathcal{K}$ to an LMI. In other words, $\mathcal{K}$ can be described by a reasonably sized LMI. In the next section, we will show that if $\mathcal{F}$ is a certain quadratic matrix function set, and D is a unit ball defined by either the spectrum norm or the Frobenius norm, then $\mathcal{K}$ can still be described by reasonably sized LMIs. Before we discuss specific results, we need to introduce some definitions.

Let $D \subseteq \Re^n$ be a given domain. Then, its homogenization is given as

$$\mathcal{H}(D) = \mathrm{cl}\left\{ \left[\begin{array}{c} t \\ x \end{array} \right] \;\middle|\; x/t \in D \right\} \subseteq \Re^{1+n}.$$

We consider the cone of copositive matrices over D to be

$$\mathcal{C}_+(D) = \{Z \in \mathcal{S}^n \mid x^T Z x \ge 0 \text{ for all } x \in D\}. \tag{2.1}$$

Let $D_1 \subseteq \Re^n$ and $D_2 \subseteq \Re^m$ be two domains. The bilinear positive cone is defined as

$$\mathcal{B}_+(D_1, D_2) = \{Z \in \Re^{n\times m} \mid x^T Z y \geq 0 \text{ for all } x \in D_1,\, y \in D_2\}.$$

Obviously, the descriptions of $\mathcal{C}_+$ and $\mathcal{B}_+$ are the same as that of $\mathcal{K}$, where $\mathcal{F}$ is taken as the set of quadratic forms, and $\mathcal{C}$ is simply $\Re_+$.

If we have a general nonhomogeneous quadratic function $q(x) = c + 2b^T x + x^T A x$, then we introduce

$$M(q(\cdot)) = \begin{bmatrix} c & b^T \\ b & A \end{bmatrix}.$$

Consider

$$\mathcal{FC}_+(D) = \{M(q(\cdot)) \mid q(x) \geq 0 \text{ for all } x \in D\}.$$

It can be shown [17] that

$$\mathcal{FC}_+(D) = \mathcal{C}_+(\mathcal{H}(D)).$$

This implies that we need only concentrate on the homogeneous form.

The following lemma plays a key role in our analysis.

LEMMA 2.2. *Let $\mathcal{K}$, $\mathcal{K}_1$, and $\mathcal{K}_2$ be closed cones. It holds that*

$$\mathcal{C}_+^*(\mathcal{K}) = \text{cone } \{xx^T \mid x \in \mathcal{K}\}$$

and

$$\mathcal{B}_+^*(\mathcal{K}_1, \mathcal{K}_2) = \text{cone } \{xy^T \mid x \in \mathcal{K}_1,\, y \in \mathcal{K}_2\}.$$

Proof. Let us consider only the second assertion. It can be shown [17, Lemma 1] that cone $\{xy^T \mid x \in \mathcal{K}_1,\, y \in \mathcal{K}_2\}$ is convex. Using the bipolar theorem, it therefore suffices to prove that

$$\mathcal{B}_+(\mathcal{K}_1, \mathcal{K}_2) = (\text{cone } \{xy^T \mid x \in \mathcal{K}_1,\, y \in \mathcal{K}_2\})^*. \tag{2.2}$$

It is clear that

$$\mathcal{B}_+(\mathcal{K}_1, \mathcal{K}_2) \subseteq (\text{conv } \{xy^T \mid x \in \mathcal{K}_1,\, y \in \mathcal{K}_2\})^*.$$

We now show the inclusion in the reverse direction. Suppose, by contradiction, that there is

$$Z \in (\text{conv } \{xy^T \mid x \in \mathcal{K}_1,\, y \in \mathcal{K}_2\})^* \setminus \mathcal{B}_+(\mathcal{K}_1, \mathcal{K}_2).$$

Then, since $Z \notin \mathcal{B}_+(\mathcal{K}_1, \mathcal{K}_2)$, by definition there exist $u \in \mathcal{K}_1$ and $v \in \mathcal{K}_2$ such that $u^T Z v < 0$. We arrive now at a contradiction, namely,

$$0 > u^T Z v = Z \bullet (uv^T) \geq 0,$$

where the latter inequality holds, since $Z \in (\text{conv } \{xy^T \mid x \in \mathcal{K}_1,\, y \in \mathcal{K}_2\})^*$. For a proof of the first statement of the lemma, see Proposition 1 in [17]. ∎

We note that, although we are primarily interested in $\mathcal{C}_+(\mathcal{K})$ and $\mathcal{B}_+(\mathcal{K}_1, \mathcal{K}_2)$, it can be advantageous to work with their dual counterparts first and then dualize to get the original cone. For instance, in [17], Sturm and Zhang used this technique to show that

$$\mathcal{C}_+^*(\text{SOC}(1+n)) = \left\{ \begin{bmatrix} z_{11} & z^T \\ z & Z \end{bmatrix} \succeq 0 \;\middle|\; z_{11} \geq \text{Tr}(Z) \right\}, \tag{2.3}$$

which is an explicit LMI system [1]. Relation (2.3) is dual to the S-lemma [19]; see Proposition 3.1 below.

3. Robust QMIs. Suppose that we consider an ordinary inequality, say $f(x) \geq 0$, where x can be viewed as a parameter, which is uncertain. Assume that this uncertain parameter x can attain any value within a set D. We call the inequality $f(x) \geq 0$ *robust* if $f(x) \geq 0$ for all $x \in D$.

In this regard, the S-lemma of Yakubovich [19] plays a key role in robust analysis, where f is quadratic and D is given as either the level set or the contour of a quadratic function. Actually, there are several variants of the S-lemma of Yakubovich, of which we list two. For proofs, see, e.g., [17].

PROPOSITION 3.1 (S-lemma, level set). *Let $f : \Re^n \to \Re$ and $g : \Re^n \to \Re$ be quadratic functions with $g(\bar{x}) > 0$ for some $\bar{x}$. It holds that*

$$f(x) \geq 0 \text{ for all } x \text{ such that } g(x) \geq 0$$

if and only if there exists $t \geq 0$ such that

$$f(x) - tg(x) \geq 0 \text{ for all } x \in \Re^n.$$

PROPOSITION 3.2 (S-lemma, contour). *Let $f : \Re^n \to \Re$ and $g : \Re^n \to \Re$ be quadratic forms with $g(x^{(1)}) < 0$ and $g(x^{(2)}) > 0$ for some $x^{(1)}$ and $x^{(2)}$. It holds that*

$$f(x) \geq 0 \text{ for all } x \text{ such that } g(x) = 0$$

if and only if there exists $t \in \Re$ such that

$$f(x) + tg(x) \geq 0 \text{ for all } x \in \Re^n.$$

In this section, we derive extensions of the S-lemma to the matrix case, namely, the robust QMI.

Our first extension of Proposition 3.1 concerns the following robust QMI:

$$(S_1): \qquad C + X^T B + B^T X + X^T A X \succeq 0 \text{ for all } X \text{ with } I - X^T D X \succeq 0.$$

We show that this robust QMI holds if and only if the data matrices (A, B, C, D) satisfy a certain LMI relation.

THEOREM 3.3. *The robust QMI (S_1) is equivalent to*

$$\begin{bmatrix} C & B^T \\ B & A \end{bmatrix} \in \left\{ Z \;\middle|\; Z - t \begin{bmatrix} I & 0 \\ 0 & -D \end{bmatrix} \succeq 0,\ t \geq 0 \right\}. \tag{3.1}$$

Proof. We first show that the robust QMI (S_1) is equivalent to the following robust quadratic inequality (S_2):

$$(S_2): \qquad \xi^T C \xi + 2\eta^T B \xi + \eta^T A \eta \geq 0 \text{ for all } \xi, \eta \text{ with } \xi^T \xi - \eta^T D \eta \geq 0.$$

To see that (S_2) implies (S_1), we fix an X satisfying

$$I - X^T D X \succeq 0.$$

Then, by letting ξ be an arbitrary vector, and $\eta := X\xi$, we see that

$$\xi^T \xi - \eta^T D \eta \geq 0,$$

which, in light of (S_2), implies

$$\xi^T C \xi + 2\eta^T B \xi + \eta^T A \eta \geq 0,$$

or, equivalently,

$$\xi^T(C + X^TB + B^TX + X^TAX)\xi \geq 0.$$

This shows that

$$C + X^TB + B^TX + X^TAX \succeq 0.$$

Next we shall show that (S_1) implies (S_2). Suppose that (S_1) holds, and let ξ and η be such that

$$\xi^T\xi - \eta^TD\eta \geq 0. \tag{3.2}$$

Consider first the case that $\xi = 0$, and let $X(u) = \eta u^T/u^Tu$ for $u \neq 0$. Due to (3.2), we have

$$X(u)^TDX(u) = \frac{\eta^TD\eta}{(u^Tu)^2}uu^T \preceq 0 \prec I \text{ for all } u \neq 0.$$

It thus follows from (S_1) that

$$0 \leq u^T\left(C + X(u)^TB + B^TX(u) + X(u)^TAX(u)\right)u = \eta^TA\eta + o(\|u\|),$$

and hence $\eta^TA\eta \geq 0$. This establishes (S_2) for the case where $\xi = 0$. If $\xi \neq 0$, we let $X = \eta\xi^T/\xi^T\xi$. Due to (3.2), we have

$$X^TDX = \frac{\eta^TD\eta}{(\xi^T\xi)^2}\xi\xi^T \preceq \frac{1}{\xi^T\xi}\xi\xi^T \preceq I.$$

Then, by (S_1) we have

$$C + X^TB + B^TX + X^TAX \succeq 0.$$

By pre- and postmultiplying on both sides of the above matrix inequality by ξ^T and ξ, respectively, we get

$$\xi^TC\xi + 2\eta^TB\xi + \eta^TA\eta \geq 0.$$

This establishes the equivalence between (S_1) and (S_2). Now, applying Proposition 3.1 to (S_2), Theorem 3.3 follows. □

Theorem 3.3 may be applied with $D = I$ (or a multiple of the identity matrix) to yield a robust QMI, where the uncertainty set is a level set of the spectral radius. At first sight, this is a more conservative robustness than one based on the Frobenius norm, since $\|X\|_2 \leq \|X\|_F$ with a strict inequality if the rank of X is more than one. Nevertheless, these uncertainty sets turn out to be equivalent for the form of QMIs treated in this section. More precisely, we have the following.

PROPOSITION 3.4. *If $D \succeq 0$, then (S_1) is equivalent to the following robust QMI:*

$$(S_3): \quad C + X^TB + B^TX + X^TAX \succeq 0 \text{ for all } X \text{ with } \mathrm{Tr}(D(XX^T)) \leq 1.$$

Proof. Observe first that if X is such that $1 \geq \mathrm{Tr}(D(XX^T)) = \mathrm{Tr}(X^TDX)$ with $D \succeq 0$, then also $I - X^TDX \succeq 0$. Therefore, (S_1) implies (S_3).

Now we wish to show the converse. Suppose that (S_3) holds, and let ξ and η be such that

$$\xi^T\xi - \eta^T D\eta \geq 0.$$

Then by letting $X = \eta\xi^T/\xi^T\xi$, we have $\mathrm{Tr}(D(XX^T)) = \eta^T D\eta/\xi^T\xi \leq 1$. It thus follows from (S_3) that

$$C + X^T B + B^T X + X^T AX \succeq 0.$$

By pre- and postmultiplying both sides by ξ^T and ξ, we further get

$$\xi^T C\xi + 2\eta^T B\xi + \eta^T A\eta \geq 0,$$

establishing (S_2). By Theorem 3.3, (S_3) also implies (S_1). □

As a consequence of the above results, we have derived an LMI description (3.1) for the data (A, B, C, D), where the following quadratic matrix function inequality holds:

$$C + X^T B + B^T X + X^T AX \succeq 0 \text{ for all } I - X^T DX \succeq 0.$$

If $D \succeq 0$, then the same LMI description (3.1) applies to the nonnegativity condition

$$C + X^T B + B^T X + X^T AX \succeq 0 \text{ for all } \mathrm{Tr}(D(XX^T)) \leq 1.$$

Below we shall further extend the results in Theorem 3.3 to a setting where a matrix quadratic fraction is present.

Consider data matrices (A, B, C, D, F, G, H) satisfying the following robust fractional QMI:

$$\text{(3.3)} \qquad \begin{cases} H \succeq 0, \\ C + X^T B + B^T X + X^T AX \succeq 0, \\ H - (F + GX)(C + X^T B + B^T X + X^T AX)^+ (F + GX)^T \succeq 0, \end{cases}$$

whenever $I - X^T DX \succeq 0$, where M^+ stands for the pseudo-inverse of $M \succeq 0$. We remark that (A, B, C, D) satisfies (S_1) if and only if $(A, B, C, D, 0, 0, 0)$ satisfies (3.3).

THEOREM 3.5. *The data matrices (A, B, C, D, F, G, H) satisfy the robust fractional QMI* (3.3) *if and only if there is $t \geq 0$ such that*

$$\begin{bmatrix} H & F & G \\ F^T & C & B^T \\ G^T & B & A \end{bmatrix} - t \begin{bmatrix} 0 & 0 & 0 \\ 0 & I & 0 \\ 0 & 0 & -D \end{bmatrix} \succeq 0.$$

Proof. Consider the QMI

$$\text{(3.4)} \qquad \begin{bmatrix} H & F + GX \\ (F + GX)^T & C + X^T B + B^T X + X^T AX \end{bmatrix} \succeq 0 \text{ for all } I - X^T DX \succeq 0.$$

By taking Schur complements, it is clear that (3.3) and (3.4) are equivalent. Unfortunately, the above QMI is not in the form of (S_1); Theorem 3.3 is therefore not applicable. Nevertheless, we can use a similar argument as in the proof of Theorem 3.3.

We shall show that the QMI (3.4) is equivalent to the following robust quadratic inequality (3.5):

$$\text{(3.5)} \qquad \xi^T H\xi + 2\xi^T F\eta + 2\xi^T G\gamma + \eta^T C\eta + \gamma^T B\eta + \eta^T B^T\gamma + \gamma^T A\gamma \geq 0$$

for all $\eta^T\eta - \gamma^T D\gamma \geq 0$. Suppose first that (3.5) holds, and fix an X satisfying $I - X^T DX \succeq 0$. Let ξ and η be arbitrary vectors, and let $\gamma := X\eta$. By construction, we have $\eta^T\eta - \gamma^T D\gamma \geq 0$ so that (3.5) implies

$$\begin{aligned} 0 &\leq \xi^T H\xi + 2\xi^T F\eta + 2\xi^T G\gamma + \eta^T C\eta + \gamma^T B\eta + \eta^T B^T\gamma + \gamma^T A\gamma \\ &= \begin{bmatrix} \xi \\ \eta \end{bmatrix}^T \begin{bmatrix} H & F + GX \\ (F + GX)^T & C + X^T B + B^T X + X^T AX \end{bmatrix} \begin{bmatrix} \xi \\ \eta \end{bmatrix}, \end{aligned}$$

establishing (3.4). Conversely, suppose that (3.4) holds, and let ξ, η, and γ be such that

$$\eta^T\eta - \gamma^T D\gamma \geq 0. \tag{3.6}$$

Consider first the case where $\eta = 0$, and let $X(u) = \gamma u^T/u^T u$ for $u \neq 0$. Due to (3.6), we have

$$X(u)^T DX(u) = \frac{\gamma^T D\gamma}{(u^T u)^2} uu^T \preceq 0 \prec I \text{ for all } u \neq 0.$$

It thus follows from (3.4) that

$$\begin{aligned} 0 &\leq \begin{bmatrix} \xi \\ u \end{bmatrix}^T \begin{bmatrix} H & F + GX(u) \\ (F + GX(u))^T & C + X(u)^T B + B^T X(u) + X(u)^T AX(u) \end{bmatrix} \begin{bmatrix} \xi \\ u \end{bmatrix} \\ &= \begin{bmatrix} \xi \\ \gamma \end{bmatrix}^T \begin{bmatrix} H & G \\ G^T & A \end{bmatrix} \begin{bmatrix} \xi \\ \gamma \end{bmatrix} + o(\|u\|). \end{aligned}$$

This establishes (3.5) for the case where $\eta = 0$. If $\eta \neq 0$, we let $X = \gamma\eta/\eta^T\eta$. Due to (3.6), we have $X^T DX \preceq I$. Then, by (3.4) we have

$$\begin{aligned} 0 &\leq \begin{bmatrix} \xi \\ \eta \end{bmatrix}^T \begin{bmatrix} H & F + GX(u) \\ (F + GX(u))^T & C + X(u)^T B + B^T X(u) + X(u)^T AX(u) \end{bmatrix} \begin{bmatrix} \xi \\ \eta \end{bmatrix} \\ &= \xi^T H\xi + 2\xi^T F\eta + 2\xi^T G\gamma + \eta^T C\eta + \gamma^T B\eta + \eta^T B^T\gamma + \gamma^T A\gamma, \end{aligned}$$

establishing (3.5). We have proved the equivalence between (3.3) and (3.5). The theorem now follows by applying Proposition 3.1 to (3.5). ∎

Analogous to Proposition 3.4, we have the following equivalence result.

PROPOSITION 3.6. *If $D \succeq 0$, then* (3.3) *is equivalent to the following robust fractional QMI:*

$$\mathrm{Tr}(DXX^T) \leq 1 \Longrightarrow \begin{cases} H \succeq 0, \\ C + X^T B + B^T X + X^T AX \succeq 0, \\ H - (F + GX)(C + X^T B + B^T X + X^T AX)^+ (F + GX)^T \succeq 0. \end{cases}$$

It is interesting to note a related, but somewhat surprising, result which we formulate in the following theorem; see also [1].

THEOREM 3.7. *The data matrices (A, B, C, F, G, H) satisfy*

$$\begin{bmatrix} H & F + GX \\ (F + GX)^T & C + X^T B + B^T X + X^T AX \end{bmatrix} \succeq 0 \text{ for all } X^T X = I \tag{3.7}$$

if and only if

$$\begin{bmatrix} H & F & G \\ F^T & C & B^T \\ G^T & B & A \end{bmatrix} - t \begin{bmatrix} 0 & 0 & 0 \\ 0 & I & 0 \\ 0 & 0 & -I \end{bmatrix} \succeq 0 \text{ for some } t \in \Re.$$

Proof. Just as in the proof of Theorem 3.5, the robust QMI (3.7) is equivalent to the following robust quadratic inequality:

$$\xi^T H\xi + 2\xi^T F\eta + 2\xi^T G\gamma + \eta^T C\eta + \gamma^T B\eta + \eta^T B^T\gamma + \gamma^T A\gamma \geq 0 \text{ for all } \eta^T\eta - \gamma^T\gamma = 0.$$

Applying Proposition 3.2 to the above relation, the theorem follows. □

Theorem 3.7 allows us to model the robust QMI over the orthonormal matrix constraints as a linear matrix inequality.

Matrix orthogonality constrained quadratic optimization problems were studied in [2, 18], where it was shown that if the objective function is homogeneous, either purely linear or quadratic, then by adding some seemingly redundant constraints one achieves strong duality with its Lagrangian dual problem.

4. General robust QMIs. Section 3 shows how we can transform some special type of robust QMIs into an LMI. In this section, we consider general robust QMIs.

We remark that the matrix inequality $Z \succeq 0$ is equivalent to the fact that $x^T Zx \geq 0$ for all $x \in \Re^n$. Thus, the LMI itself is nothing but a special type of robust quadratic inequality. The same is true for the copositive matrix cone (2.1). From this viewpoint, we may formulate the general robust QMIs as an ordinary robust inequality involving polynomials of order no more than 4.

Consider a domain $D \subseteq \Re^n$ and a domain $\Delta \subseteq \Re^m$. In the same spirit as (2.1), let us define

$$\mathcal{C}_+(D,\Delta) := \left\{ Z \in \mathcal{L}_{n,m} \;\middle|\; \sum_{i=1}^n \sum_{j=1}^n x_i x_j y^T Z_{ij} y \geq 0 \text{ for all } x \in D,\, y \in \Delta \right\}, \tag{4.1}$$

where $\mathcal{L}_{n,m}$ represents the $mn(m+1)(n+1)/4$-dimensional linear space of biquadratic forms. More precisely, $\mathcal{L}_{n,m}$ is defined as follows:

$$\mathcal{L}_{n,m} := \left\{ \begin{bmatrix} G_{11} & G_{12} & \cdots & G_{1n} \\ G_{21} & G_{22} & \cdots & G_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ G_{n1} & G_{n2} & \cdots & G_{nn} \end{bmatrix} \in \mathcal{S}^{n\times m} \;\middle|\; G_{ij}^T = G_{ij} \in \mathcal{S}^m,\ i,j = 1,2,\ldots,n \right\}.$$

Notice that $\mathcal{C}_+(D) = \mathcal{C}_+(D, \Re_+) = \mathcal{C}_+(D, \Re)$.

Certainly, $\mathcal{C}_+(\Delta)$ is a well-defined closed convex cone. It is easy to see that $\mathcal{C}_+(D, \Delta)$ can be equivalently viewed as a robust QMI over D in the conic order defined by $\mathcal{C}_+(\Delta)$, i.e.,

$$\mathcal{C}_+(D, \Delta) = \left\{ Z \in \mathcal{L}_{n,m} \;\middle|\; \sum_{i=1}^{n}\sum_{j=1}^{n} x_i x_j Z_{ij} \in \mathcal{C}_+(\Delta) \text{ for all } x \in D \right\}.$$

Given a quadratic function $q : \Re^n \to \mathcal{S}^m$,

$$q(x) = C + 2\sum_{j=1}^{n} x_j B_j + \sum_{i=1}^{n}\sum_{j=1}^{n} x_i x_j A_{ij}, \tag{4.2}$$

we let $M(q(\cdot)) \in \mathcal{L}_{n+1,m}$ denote the matrix representation of $q(\cdot)$, i.e.,

$$M(q(\cdot)) = \begin{bmatrix} C & B_1 & \cdots & B_n \\ B_1 & A_{11} & \cdots & A_{1n} \\ \vdots & \vdots & \vdots & \vdots \\ B_n & A_{n1} & \cdots & A_{nn} \end{bmatrix}.$$

The cone of $\mathcal{C}_+(\Delta)$-nonnegative quadratic functions over D is now conveniently defined as

$$\mathcal{FC}_+(D, \Delta) = \{M(q(\cdot)) \mid q(x) \in \mathcal{C}_+(\Delta) \text{ for all } x \in D\}.$$

Clearly, $\mathcal{FC}_+(D) = \mathcal{FC}_+(D, \Re)$. Furthermore, it can be shown [17] that

$$\mathcal{FC}_+(D, \Delta) = \mathcal{C}_+(\mathcal{H}(D), \Delta).$$

This implies that we need only concentrate on the homogeneous form.

Similar to Lemma 2.2, we have the following representation.

LEMMA 4.1. *Let $D \subseteq \Re^n$ and $\Delta \subseteq \Re^m$. In the linear space $\mathcal{L}_{n,m}$ it holds that*

$$\mathcal{C}_+^*(D, \Delta) = \text{cone } \{(xx^T) \otimes (yy^T) \mid x \in D,\, y \in \Delta\} \tag{4.3}$$

$$= \text{cone } \{(xx^T) \otimes Y \mid x \in D,\, Y \in \mathcal{C}_+(\Delta)^*\} \tag{4.4}$$

$$= \text{cone } \{X \otimes Y \mid X \in \mathcal{C}_+(D)^*,\, Y \in \mathcal{C}_+(\Delta)^*\}. \tag{4.5}$$

Proof. It can be shown [17, Lemma 1] that cone $\{(xx^T)\otimes(yy^T) \mid x \in D,\, y \in \Delta\}$ is convex. Using also the bipolar theorem, an equivalent statement of (4.3) is therefore

$$\mathcal{C}_+(D, \Delta) = \text{cone } \{(xx^T) \otimes (yy^T) \mid x \in D,\, y \in \Delta\}^*. \tag{4.6}$$

If $Z \in \mathcal{C}_+^*(D, \Delta)$, then for all $x \in D$ and $y \in \Delta$, we have

$$0 \le y^T \left(\sum_{i=1}^{n}\sum_{j=1}^{n} x_i x_j Z_{ij} \right) y = (x \otimes y)^T Z (x \otimes y) = Z \bullet ((xx^T) \otimes (yy^T)).$$

This shows that

$$\mathcal{C}_+^*(D, \Delta) \subseteq \text{cone } \{(xx^T) \otimes (yy^T) \mid x \in D,\, y \in \Delta\}.$$

In order to establish the converse relation, suppose by contradiction that there exists

$$Z \in \text{cone}\,\{(xx^T) \otimes (yy^T) \mid x \in D,\, y \in \Delta\}^* \setminus \mathcal{C}_+(D,\Delta). \tag{4.7}$$

Since $Z \notin \mathcal{C}_+(D,\Delta)$, there must exist $x \in D$ and $y \in \Delta$ such that

$$0 > y^T \left(\sum_{i=1}^n \sum_{j=1}^n x_i x_j Z_{ij} \right) y = Z \bullet ((xx^T) \otimes (yy^T)) \geq 0,$$

where the latter inequality follows from (4.7). This impossible inequality completes the proof of (4.6), and hence (4.3). The equivalence between (4.3) and (4.4) and (4.5) follows from Lemma 2.2. □

It can be seen that

$$\mathcal{L}_{n,m} = \text{span}\,\{X \otimes Y \mid X \in \mathcal{S}^n,\, Y \in \mathcal{S}^m\}.$$

To verify this relation, we first notice that the right-hand-side linear subspace is contained in $\mathcal{L}_{n,m}$, since each matrix of the form $X \otimes Y$ is in $\mathcal{L}_{n,m}$. Then we check that the dimensions of the two linear subspaces are actually equal. This establishes the above equality.

There is a one-to-one correspondence between $\mathcal{L}_{n,m}$ and $\mathcal{L}_{m,n}$ by means of a permutation operator. In particular, we implicitly define the permutation matrix $N_{m,n}$ by

$$N_{m,n}\,\text{vec}(X) = \text{vec}(X^T) \text{ for all } X \in \Re^{m\times n}. \tag{4.8}$$

We are now in a position to list some standard results on the Kronecker product.

PROPOSITION 4.2. *Let $A \in \Re^{p\times m}$, $B \in \Re^{m\times n}$, and $C \in \Re^{n\times q}$. Then*

$$\text{vec}(ABC) = (C^T \otimes A)\,\text{vec}(B), \tag{4.9}$$

$$N_{m,n}^{-1} = N_{m,n}^T = N_{n,m}, \tag{4.10}$$

$$N_{p,q}(C^T \otimes A)N_{n,m} = A \otimes C^T. \tag{4.11}$$

Proof. We prove only (4.11), since the other two results are straightforward. We have

$$\begin{aligned} N_{p,q}(C^T \otimes A)N_{n,m}\,\text{vec}(B^T) &\overset{(4.8)}{=} N_{p,q}(C^T \otimes A)\,\text{vec}(B) \\ &\overset{(4.9)}{=} N_{p,q}\,\text{vec}(ABC) \\ &\overset{(4.8)}{=} \text{vec}(C^T B^T A^T) \\ &\overset{(4.9)}{=} (A \otimes C^T)\,\text{vec}(B^T) \end{aligned}$$

for arbitrary B. Hence we have (4.11). □

Notice that, in particular, from (4.11) we have

$$N_{m,n}(X \otimes Y)N_{n,m} = Y \otimes X \text{ for all } X \in \mathcal{S}^n,\, Y \in \mathcal{S}^m,$$

so that

$$\mathcal{L}_{m,n} = \{N_{m,n} Z N_{n,m} \mid Z \in \mathcal{L}_{n,m}\}.$$

THEOREM 4.3. *Let $D \subseteq \Re^n$ and $\Delta \subseteq \Re^m$. Consider the cones $\mathcal{C}_+(D,\Delta)$, $\mathcal{C}_+(\Delta,D)$ and their duals in $\mathcal{L}_{n,m}$ and $\mathcal{L}_{m,n}$ respectively. It holds that*

$$\mathcal{C}_+(\Delta,D) = \{N_{m,n} X N_{n,m} \in \mathcal{L}_{m,n} \mid X \in \mathcal{C}_+(D,\Delta)\} \tag{4.12}$$

and

$$\mathcal{C}_+^*(\Delta,D) = \{N_{m,n} Z N_{n,m} \in \mathcal{L}_{m,n} \mid Z \in \mathcal{C}_+^*(D,\Delta)\}. \tag{4.13}$$

Proof. Relation (4.13) follows from applying (4.11) to Lemma 4.1. Relation (4.12) follows by dualization. ∎

We shall first consider the cone $\mathcal{C}_+(\Re^n,\Re^m)$ residing in $\mathcal{L}_{n,m}$. The following theorem provides an LMI characterization if either $n = 2$ or $m = 2$.

THEOREM 4.4. *Consider the cone $\mathcal{C}_+(\Re^2,\Re^m)$ and its dual in $\mathcal{L}_{2,m}$. It holds that*

$$\begin{aligned}\mathcal{C}_+(\Re^2,\Re^m) &= \mathcal{FC}_+(\Re,\Re^m) = (\mathcal{L}_{2,m} \cap \mathcal{S}_+^{2m})^* \\ &= \left\{ \begin{bmatrix} A & B \\ B & C \end{bmatrix} \in \mathcal{L}_{n,m} \;\middle|\; \begin{bmatrix} A & B+\tilde{B} \\ B-\tilde{B} & C \end{bmatrix} \in \mathcal{S}_+^{2m} \text{ for some } \tilde{B} = -\tilde{B}^T \right\}.\end{aligned}$$

The cone

$$\mathcal{C}_+(\Re^n,\Re^2) = \mathcal{FC}_+(\Re^{n-1},\Re^2) = (\mathcal{L}_{n,2} \cap \mathcal{S}_+^{2n})^*$$

has a similar LMI characterization, which is due to Theorem 4.3.

Proof. The relation $\mathcal{FC}_+(\Re,\Re^m) = (\mathcal{L}_{2,m} \cap \mathcal{S}_+^{2m})^*$ is a special case of Theorem 4.2 in Genin et al. [8] on matrix polynomials. The second part of the lemma follows from Theorem 4.3. For completeness, we provide a direct proof below.

By Lemma 4.1, we have

$$\mathcal{C}_+^*(\Re^2,\Re^m) \subseteq \mathcal{S}_+^{2m} \cap \mathcal{L}_{2,m}.$$

We now show the inclusion in the reverse direction. For this purpose, we take any

$$G = \begin{bmatrix} G_{11} & G_{12} \\ G_{12} & G_{22} \end{bmatrix} \in \mathcal{S}_+^{2m} \cap \mathcal{L}_{2,m}$$

and prove that $G \in \mathcal{C}_+^*(\Re^2,\Re^m)$. We will use the obvious invariance relation

$$\mathcal{C}_+^*(\Re^2,\Re^m) = \left\{ \begin{bmatrix} P & 0 \\ 0 & P \end{bmatrix} Z \begin{bmatrix} P^T & 0 \\ 0 & P^T \end{bmatrix} \;\middle|\; Z \in \mathcal{C}_+^*(\Re^2,\Re^m) \right\},$$

where P is any nonsingular real matrix.

Let $G(\epsilon) = G + \epsilon I \succ 0$, where $\epsilon > 0$ is an arbitrarily small (but fixed) positive number. Since $G_{22}(\epsilon) \succ 0$, and $G_{12}(\epsilon) = G_{12}$ is symmetric, there exists a nonsingular matrix P_ϵ such that

$$P_\epsilon G_{22}(\epsilon) P_\epsilon^T = I \text{ and } P_\epsilon G_{12}(\epsilon) P_\epsilon^T = \Lambda_\epsilon,$$

where $\Lambda_\epsilon = \operatorname{diag}(\lambda_1(\epsilon),\ldots,\lambda_m(\epsilon))$ is a diagonal matrix. In fact, $P_\epsilon = G_{22}(\epsilon)^{-1/2} Q_\epsilon$ for some orthogonal matrix Q_ϵ. To show $G(\epsilon) \in \mathcal{C}_+^*(\Re^2,\Re^m)$, we need only prove

$$\begin{bmatrix} P_\epsilon & 0 \\ 0 & P_\epsilon \end{bmatrix} G(\epsilon) \begin{bmatrix} P_\epsilon^T & 0 \\ 0 & P_\epsilon^T \end{bmatrix} = \begin{bmatrix} P_\epsilon G_{11}(\epsilon) P_\epsilon^T & \Lambda_\epsilon \\ \Lambda_\epsilon & I \end{bmatrix} \in \mathcal{C}_+^*(\Re^2,\Re^m).$$

By the well-known Schur complement lemma, we have

$$P_\epsilon G_{11}(\epsilon) P_\epsilon^T - \Lambda_\epsilon^2 \succ 0.$$

Therefore, we obtain the following representation:

$$\begin{aligned}\left[\begin{array}{cc} P_\epsilon G_{11}(\epsilon) P_\epsilon^T & \Lambda_\epsilon \\ \Lambda_\epsilon & I \end{array}\right] &= \left[\begin{array}{cc} P_\epsilon G_{11}(\epsilon) P_\epsilon^T - \Lambda_\epsilon^2 & 0 \\ 0 & 0 \end{array}\right] + \left[\begin{array}{cc} \Lambda_\epsilon^2 & \Lambda_\epsilon \\ \Lambda_\epsilon & I \end{array}\right] \\ &= \left[\begin{array}{cc} 1 & 0 \\ 0 & 0 \end{array}\right] \otimes (P_\epsilon G_{11}(\epsilon) P_\epsilon^T - \Lambda_\epsilon^2) \\ &\quad + \sum_{i=1}^m \left[\begin{array}{cc} \lambda_i^2 & \lambda_i \\ \lambda_i & 1 \end{array}\right] \otimes (e_i e_i^T), \end{aligned} \tag{4.14}$$

where $e_i \in \Re^m$ is the ith column of the $m \times m$ identity matrix. By Lemma 4.1, the above representation shows that the matrix

$$\left[\begin{array}{cc} P_\epsilon G_{11}(\epsilon) P_\epsilon^T & \Lambda_\epsilon \\ \Lambda_\epsilon & I \end{array}\right]$$

lies in $\mathcal{C}_+^*(\Re^2, \Re^m)$. Consequently, $G(\epsilon) \in \mathcal{C}_+^*(\Re^2, \Re^m)$. Since $\mathcal{C}_+^*(\Re^2, \Re^m)$ is a closed cone, we have $G = \lim_{\epsilon \to 0} G(\epsilon) \in \mathcal{C}_+^*(\Re^2, \Re^m)$. This proves the first part of the theorem. The characterization of the primal cone $\mathcal{C}_+(\Re^2, \Re^m)$ follows by dualization, namely,

$$\begin{aligned}\mathcal{C}_+(\Re^2, \Re^m) &= \mathrm{cl}(\mathcal{C}_+(\Re^2, \Re^m)) = \mathcal{C}_+^{**}(\Re^2, \Re^m) \\ &= \mathrm{cl}(\mathcal{L}_{n,m}^\perp + \mathcal{S}_+^{2m}) \cap \mathcal{L}_{n,m} \\ &= \left\{ \left[\begin{array}{cc} A & B \\ B & C \end{array}\right] \in \mathcal{L}_{n,m} \;\middle|\; \left[\begin{array}{cc} A & B + \tilde{B} \\ B - \tilde{B} & C \end{array}\right] \in \mathcal{S}_+^{2m} \text{ for some } \tilde{B} = -\tilde{B}^T \right\}. \end{aligned}$$

This completes the proof. □

We remark that $\mathcal{C}_+(\Re^2, \Re^m)$ (or equivalently, $\mathcal{C}_+(\Re^m, \Re^2)$) is not self-dual. For instance, we have

$$\left[\begin{array}{cccc} 1 & 0 & 0 & 1/2 \\ 0 & 0 & 1/2 & 0 \\ 0 & 1/2 & 0 & 0 \\ 1/2 & 0 & 0 & 1 \end{array}\right] \in \mathcal{C}_+(\Re^2, \Re^2) \setminus \mathcal{C}_+^*(\Re^2, \Re^2).$$

The membership problem of $\mathcal{C}_+(\Re^n, \Re^m)$ for general n and m is a hard problem; see Corollary 4.10 later in this paper.

We study now the mixed copositive/positive semidefinite biquadratic forms, i.e., $\mathcal{C}_+(\Re_+^n, \Re^m)$ and $\mathcal{C}_+(\Re^n, \Re_+^m)$. For $m = 2$, we arrive at a special case of nonnegative polynomial matrices on the positive real half-line (see [14] for the scalar case).

THEOREM 4.5. *There holds*

$$\mathcal{C}_+^*(\Re_+^2, \Re^m) = \left\{ \left[\begin{array}{cc} G_{11} & G_{12} \\ G_{12} & G_{22} \end{array}\right] \in \mathcal{L}_{2,m} \cap \mathcal{S}_+^{2m} \;\middle|\; G_{12} \succeq 0 \right\}. \tag{4.15}$$

Consequently, the primal cone $\mathcal{C}_+(\Re_+^2, \Re^m)$ can be characterized as

$$\begin{aligned}&\mathcal{C}_+(\Re_+^2, \Re^m) = \mathcal{FC}_+(\Re_+, \Re^m) \\ &= \left\{ \left[\begin{array}{cc} C & B \\ B & A \end{array}\right] \in \mathcal{L}_{2,m} \;\middle|\; \left[\begin{array}{cc} C & B \\ B & A \end{array}\right] - \left[\begin{array}{cc} 0 & E \\ E^T & 0 \end{array}\right] \succeq 0,\ E + E^T \succeq 0 \text{ for some } E \right\}. \end{aligned}$$

Proof. First, it follows from Lemma 4.1 that

$$\mathcal{C}_+^*(\Re_+^2, \Re^m) \subseteq \left\{ \left[\begin{array}{cc} G_{11} & G_{12} \\ G_{12} & G_{22} \end{array} \right] \in \mathcal{S}_+^{2m} \cap \mathcal{L}_{2,m} \;\middle|\; G_{12} \succeq 0 \right\}.$$

It remains to argue the inclusion in the reverse direction. To this end, let

$$G \in \left\{ \left[\begin{array}{cc} G_{11} & G_{12} \\ G_{12} & G_{22} \end{array} \right] \in \mathcal{S}_+^{2m} \cap \mathcal{L}_{2,m} \;\middle|\; G_{12} \succeq 0 \right\}$$

be arbitrary. We follow the same proof technique for Theorem 4.4, and we use $G(\epsilon)$, P_ϵ, and Λ_ϵ defined there. The only difference is that $\Lambda_\epsilon \succeq 0$, due to the fact that $G_{12} \succeq 0$. Relation (4.14) states that

$$\left[\begin{array}{cc} P_\epsilon G_{11}(\epsilon) P_\epsilon^T & \Lambda_\epsilon \\ \Lambda_\epsilon & I \end{array} \right] = \left[\begin{array}{cc} 1 & 0 \\ 0 & 0 \end{array} \right] \otimes (P_\epsilon G_{11}(\epsilon) P_\epsilon^T - \Lambda_\epsilon^2) + \sum_{i=1}^m \left[\begin{array}{cc} \lambda_i^2 & \lambda_i \\ \lambda_i & 1 \end{array} \right] \otimes (e_i e_i^T).$$

Obviously, $\left[\begin{smallmatrix} 1 \\ 0 \end{smallmatrix} \right], \left[\begin{smallmatrix} \lambda_i \\ 1 \end{smallmatrix} \right] \in \Re_+^2$. By Lemma 4.1, the above representation thus shows that the matrix

$$\left[\begin{array}{cc} P_\epsilon G_{11}(\epsilon) P_\epsilon^T & \Lambda_\epsilon \\ \Lambda_\epsilon & I \end{array} \right]$$

lies in $\mathcal{C}_+^*(\Re_+^2, \Re^m)$. Consequently, $G(\epsilon) \in \mathcal{C}_+^*(\Re_+^2, \Re^m)$, and by continuity $G \in \mathcal{C}_+^*(\Re_+^2, \Re^m)$.

The remaining claim about the characterization of the primal cone $\mathcal{C}_+(\Re_+^2, \Re^m)$ can be easily verified by taking the dual on both sides of (4.15). ∎

As a special case of the above theorem, we see that $\mathcal{C}_+(\Re_+^2) = \mathcal{S}_+^2 + \Re_+^{2\times 2}$, which is a well-known characterization of the 2×2 copositive cone. However, for $n > 2$ one merely has $\mathcal{S}_+^n + \Re_+^{n\times n} \subset \mathcal{C}_+(\Re_+^n)$. In fact, the membership problem of the copositive cone is co-NP-complete [13]. Hence the membership problem of $\mathcal{C}_+(\Re_+^n, \Re^m)$ also is co-NP-complete.

Consider the case of $D = [0, 1]$. This is a special case of nonnegative polynomial matrices on an interval (see [14] for the scalar case).

THEOREM 4.6. *Let $D = [0, 1]$. Then, we have*

$$\mathcal{C}_+^*(\mathcal{H}([0,1]), \Re^m) = \left\{ \left[\begin{array}{cc} G_{11} & G_{12} \\ G_{12} & G_{22} \end{array} \right] \in \mathcal{L}_{2,m} \cap \mathcal{S}_+^{2m} \;\middle|\; G_{12} - G_{22} \succeq 0 \right\}. \tag{4.16}$$

As a result, the primal cone $\mathcal{C}_+(\mathcal{H}([0,1]), \Re^m) = \mathcal{FC}_+([0,1], \Re^m)$ can be characterized as

$$\begin{aligned} &\mathcal{FC}_+([0,1], \Re^m) \\ &= \left\{ \left[\begin{array}{cc} C & B \\ B & A \end{array} \right] \in \mathcal{L}_{2,m} \;\middle|\; \left[\begin{array}{cc} C & B - E \\ B - E^T & A + E + E^T \end{array} \right] \succeq 0,\ E + E^T \succeq 0 \text{ for some } E \right\}. \end{aligned}$$

Proof. First, it follows from Lemma 4.1 that

$$\mathcal{C}_+^*(\mathcal{H}([0,1]), \Re^m) \subseteq \left\{ \left[\begin{array}{cc} G_{11} & G_{12} \\ G_{12} & G_{22} \end{array} \right] \in \mathcal{S}_+^{2m} \cap \mathcal{L}_{2,m} \;\middle|\; G_{12} \succeq G_{22} \right\}.$$

(Of course, one also has $G_{11} \succeq G_{12}$, but this relation is implied by $G \succeq 0$ and $G_{12} \succeq G_{22}$.)

It remains to argue the inclusion in the reverse direction. To this end, let

$$G \in \left\{ \left[\begin{array}{cc} G_{11} & G_{12} \\ G_{12} & G_{22} \end{array} \right] \in \mathcal{S}_+^{2m} \cap \mathcal{L}_{2,m} \;\middle|\; G_{12} \succeq G_{22} \right\}$$

be arbitrary. We follow the same proof technique for Theorem 4.4, and we use $G(\epsilon)$, P_ϵ, and Λ_ϵ defined there. The only difference is that $\Lambda_\epsilon \succeq I$, due to the fact that $G_{12} \succeq G_{22}$. Relation (4.14) states that

$$\left[\begin{array}{cc} P_\epsilon G_{11}(\epsilon) P_\epsilon^T & \Lambda_\epsilon \\ \Lambda_\epsilon & I \end{array} \right] = \left[\begin{array}{cc} 1 & 0 \\ 0 & 0 \end{array} \right] \otimes (P_\epsilon G_{11}(\epsilon) P_\epsilon^T - \Lambda_\epsilon^2) + \sum_{i=1}^m \left[\begin{array}{cc} \lambda_i^2 & \lambda_i \\ \lambda_i & 1 \end{array} \right] \otimes (e_i e_i^T).$$

Obviously, $\left[\begin{smallmatrix} 1 \\ 0 \end{smallmatrix} \right], \left[\begin{smallmatrix} \lambda_i \\ 1 \end{smallmatrix} \right] = \lambda_i \left[\begin{smallmatrix} 1 \\ 1/\lambda_i \end{smallmatrix} \right] \in \mathcal{H}([0,1])$ because $1/\lambda_i \in [0,1]$ for all i. By Lemma 4.1, the above representation thus shows that the matrix

$$\left[\begin{array}{cc} P_\epsilon G_{11}(\epsilon) P_\epsilon^T & \Lambda_\epsilon \\ \Lambda_\epsilon & I \end{array} \right]$$

lies in $\mathcal{C}_+^*(\Re_+^2, \Re^m)$. Consequently, $G(\epsilon) \in \mathcal{C}_+^*(\Re_+^2, \Re^m)$, and by continuity $G \in \mathcal{C}_+^*(\Re_+^2, \Re^m)$.

The characterization of the primal cone $\mathcal{FC}_+([0,1], \Re^m)$ can be easily established by taking the dual on both sides of (4.16). □

Quadratic programming over a box $[0,1]^n$ is well known to be NP-complete for general n; see [13]. Hence, the membership problems of $\mathcal{FC}_+([0,1]^n)$ and $\mathcal{FC}_+([0,1]^n, \Re^m)$ with general n also are co-NP-complete.

Recall from (2.3) that

$$\mathcal{C}_+(\mathrm{SOC}(n)) = \mathcal{FC}_+(\{x \in \Re^n \mid x^T x \le 1\}) = \{Z \in \mathcal{S}_+^n \mid J \bullet Z \ge 0\}^*,$$

where $J := 2e_1 e_1^T - I$. Using Lemma 4.1, we have

$$\begin{aligned} \mathcal{C}_+^*(\mathrm{SOC}(n), \Re^m) &= \mathrm{cone}\,\{X \otimes Y \mid X \in \mathcal{C}_+(\mathrm{SOC}(n))^*,\, Y \in \mathcal{C}_+(\Re^m)^*\} \\ &= \mathrm{cone}\,\{X \otimes Y \mid X \in \mathcal{S}_+^n,\, Y \in \mathcal{S}_+^m,\, J \bullet X \ge 0\}. \end{aligned} \tag{4.17}$$

Furthermore, we know from Theorem 4.3 that this cone is isomorphic to (i.e., in one-to-one correspondence with)

$$\mathcal{C}_+^*(\Re^m, \mathrm{SOC}(n)) = \mathrm{cone}\,\{X \otimes Y \mid X \in \mathcal{S}_+^m,\, Y \in \mathcal{S}_+^n,\, J \bullet Y \ge 0\}. \tag{4.18}$$

From this relation, it is clear that

$$\mathcal{C}_+^*(\Re^2, \mathrm{SOC}(m)) \subseteq \left\{ \left[\begin{array}{cc} Z_{11} & Z_{12} \\ Z_{21} & Z_{22} \end{array} \right] \in \mathcal{L}_{2,m} \cap \mathcal{S}_+^{2m} \;\middle|\; \left[\begin{array}{cc} J \bullet Z_{11} & J \bullet Z_{12} \\ J \bullet Z_{21} & J \bullet Z_{22} \end{array} \right] \succeq 0 \right\}. \tag{4.19}$$

A natural conjecture is that (4.19) might be an equality. Unfortunately, this conjecture turns out to be false.

Counterexample. Let $p = [\ 2 \ \ 1 \ \ 0\ ]^T$ and

$$Z_{11} = pp^T + 2e_3 e_3^T, \quad Z_{12} = Z_{21} = pp^T, \quad Z_{22} = pp^T + 6e_1 e_1^T,$$

that is,

$$Z = \left[\begin{array}{c} p \\ p \end{array} \right] \left[\begin{array}{c} p \\ p \end{array} \right]^T + 2 \left[\begin{array}{c} e_3 \\ 0 \end{array} \right] \left[\begin{array}{c} e_3 \\ 0 \end{array} \right]^T + 6 \left[\begin{array}{c} 0 \\ e_1 \end{array} \right] \left[\begin{array}{c} 0 \\ e_1 \end{array} \right]^T,$$

where $e_1 = [\ 1\ \ 0\ \ 0\ ]^T$ and $e_3 = [\ 0\ \ 0\ \ 1\ ]^T$. Clearly, $Z \in \mathcal{L}_{2,3}$ and $Z \succeq 0$. Moreover, we find that

$$\begin{bmatrix} J \bullet Z_{11} & J \bullet Z_{12} \\ J \bullet Z_{21} & J \bullet Z_{22} \end{bmatrix} = \begin{bmatrix} 1 \\ 3 \end{bmatrix} \begin{bmatrix} 1 \\ 3 \end{bmatrix}^T. \tag{4.20}$$

So Z lies in the cone defined by the right-hand side of (4.19). We claim that $Z \notin \mathcal{C}_+^*(\Re^2, \mathrm{SOC}(3))$. Suppose to the contrary that $Z \in \mathcal{C}_+^*(\Re^2, \mathrm{SOC}(3))$. Then $Z = \sum_{i=1}^k (x_i x_i^T) \otimes (y_i y_i^T)$, where $x_i \in \Re^2$, $y_i \in \mathrm{SOC}(3)$. Notice that

$$\begin{bmatrix} J \bullet Z_{11} & J \bullet Z_{12} \\ J \bullet Z_{21} & J \bullet Z_{22} \end{bmatrix} = \sum_{i=1}^k (J \bullet (y_i y_i^T)) x_i x_i^T = \begin{bmatrix} 1 \\ 3 \end{bmatrix} \begin{bmatrix} 1 \\ 3 \end{bmatrix}^T,$$

where the last step follows from (4.20). Since $y_i \in \mathrm{SOC}(3)$, it follows that $J \bullet (y_i y_i^T) \geq 0$ for all i. Therefore, the above relation implies that each x_i must be a constant multiple of the vector $[\ 1\ \ 3\ ]^T$. By a renormalization, if necessary, we can assume $x_i = [\ 1\ \ 3\ ]^T$ for all i. As a result, we obtain

$$Z = \left(\begin{bmatrix} 1 \\ 3 \end{bmatrix} \begin{bmatrix} 1 \\ 3 \end{bmatrix}^T \right) \otimes Y = \begin{bmatrix} 1 & 3 \\ 3 & 9 \end{bmatrix} \otimes Y,$$

where $Y = \sum_i y_i y_i^T$. This implies that $Z_{22} = 3Z_{12} = 9Z_{11}$. This clearly contradicts the definitions of Z_{11}, Z_{12}, and Z_{22}. We therefore have proved that $Z \notin \mathcal{C}_+^*(\Re^2, \mathrm{SOC}(3))$.

It remains an open question as to whether the cone $\mathcal{C}_+^*(\Re^2, \mathrm{SOC}(m))$ is representable by LMIs. Below, we shall characterize the cone

$$\left\{ (A, C) \in \mathcal{L}_{n,m} \times \mathcal{S}^m \;\middle|\; \begin{bmatrix} C & 0 \\ 0 & A \end{bmatrix} \in \mathcal{C}_+(\mathrm{SOC}(n), \Delta) \right\} \tag{4.21}$$

for given $\Delta \subseteq \Re^m$. In other words, we consider quadratic functions $q : \{x \in \Re^n \mid x^T x \leq 1\} \to \mathcal{C}_+(\Delta)$, where the B_j's in (4.2) are all zero. Our result is the following.

THEOREM 4.7. *Let $r > 0$ be a given scalar quantity and let $\emptyset \neq \Delta \subseteq \Re^m$ be a given domain. It holds that $A \in \mathcal{L}_{n,m}$, $C \in \mathcal{S}^m$ satisfy*

$$y^T C y + \sum_{i=1}^n \sum_{j=1}^n x_i x_j\, y^T A_{ij} y \geq 0 \text{ for all } x^T x \leq r,\ y \in \Delta \tag{4.22}$$

if and only if

$$C \in \mathcal{C}_+(\Delta) \text{ and } rA + I \otimes C \in \mathcal{C}_+(\Re^n, \Delta). \tag{4.23}$$

Proof. We shall use the Rayleigh–Ritz characterization of the smallest eigenvalue of a symmetric matrix $Z = Z^T$. The smallest eigenvalue, denoted $\lambda_{\min}(Z)$, is characterized as follows:

$$\lambda_{\min}(Z) = \min\{u^T Z u \mid u^T u = 1\}. \tag{4.24}$$

Suppose now that (4.22) holds. Setting $x = 0$, we obtain that $C \in \mathcal{C}_+(\Delta)$. It also follows immediately from (4.22) that for arbitrary $y \in \Delta$,

$$\begin{aligned} 0 &\le y^T C y + \min_x \left\{ \sum_{i=1}^n \sum_{j=1}^n x_i x_j \, y^T A_{ij} y \;\middle|\; x^T x = r \right\} \\ &= y^T C y + r \min_\xi \{ \xi^T (I \otimes y)^T A (I \otimes y) \xi \mid \xi^T \xi = 1 \} \\ &= y^T C y + r \lambda_{\min} \left((I \otimes y)^T A (I \otimes y) \right), \end{aligned}$$

where we used (4.24). It follows that

$$r(I \otimes y)^T A (I \otimes y) \succeq -(y^T C y) I \text{ for all } y \in \Delta.$$

By pre- and postmultiplying both sides with an arbitrary $\xi \in \Re^n$, we obtain that

$$\begin{aligned} 0 &\le r(\xi \otimes y)^T A (\xi \otimes y) + (y^T C y) \xi^T \xi \\ &= (\xi \otimes y)^T (rA + I \otimes C)(\xi \otimes y) \\ &= (rA + I \otimes C) \bullet \left((\xi \xi^T) \otimes (y y^T) \right) \end{aligned}$$

for all $\xi \in \Re^n$, $y \in \Delta$. From Lemma 4.1, this in turn is equivalent to

$$rA + I \otimes C \in \mathcal{C}_+(\Re^n, \Delta).$$

We have shown that (4.22) implies (4.23). Conversely, suppose that (A, C) satisfies (4.23), and let $x \in \Re^n$, $y \in \Delta$, $x^T x \le r$ be arbitrary. We have

$$y^T C y + \sum_{i=1}^n \sum_{j=1}^n x_i x_j \, y^T A_{ij} y = \frac{(x \otimes y)^T (rA + I \otimes C)(x \otimes y) + (r - x^T x) y^T C y}{r} \ge 0,$$

where the inequality follows immediately from (4.23) and the nonnegativity of $r - x^T x$. This completes the proof. □

By the same argument, the following theorem is readily proven.

THEOREM 4.8. *Let $\emptyset \ne \Delta \subseteq \Re^m$ and let $r > 0$ be a given scalar quantity. It holds that $A \in \mathcal{L}_{n,m}$, $C \in \mathcal{S}^m$ satisfy*

$$y^T C y + \sum_{i=1}^n \sum_{j=1}^n x_i x_j \, y^T A_{ij} y \ge 0 \text{ for all } x^T x = r, \, y \in \Delta$$

if and only if

$$rA + I \otimes C \in \mathcal{C}_+(\Re^n, \Delta).$$

Recall from Theorems 4.4–4.6 that $\mathcal{C}_+(\Re^n, \Re^2)$, $\mathcal{C}_+(\Re^n, \Re^2_+)$, and $\mathcal{FC}_+(\Re^n, [0,1])$ are efficiently LMI representable. Theorems 4.7–4.8 therefore provide an efficient LMI characterization for the class of 2×2 robust multivariate QMIs whose entries are co-centered (e.g., centered at the origin) over the unit ball. Stated more clearly, we have obtained an efficient LMI representation for the following robust QMI:

$$\begin{bmatrix} x^T C x + c & x^T B x + b \\ x^T B x + b & x^T A x + a \end{bmatrix} \in \mathcal{FC}_+(\Delta) \text{ for all } x \in D,$$

where D is either $\{x \in \Re^n \mid x^T x \le r\}$ or $\{x \in \Re^n \mid x^T x = r\}$, and Δ is either $\Re$, $\Re_+$, or $[0, 1]$. In particular, we have the following equivalences:

1. For symmetric matrices A, B, and C, the robust QMI
$$\begin{bmatrix} x^TCx+c & x^TBx+b \\ x^TBx+b & x^TAx+a \end{bmatrix} \succeq 0 \text{ for all } \|x\|^2 \le r$$
holds if and only if
$$\begin{bmatrix} C & B \\ B & A \end{bmatrix} \in \mathcal{L}_{2,m}, \quad \begin{bmatrix} rC+cI & rB+bI-E \\ rB+bI+E & rA+aI \end{bmatrix} \succeq 0, \quad \begin{bmatrix} c & b \\ b & a \end{bmatrix} \succeq 0$$
for some E with $E+E^T=0$.
2. For symmetric matrices A, B, and C, the condition
$$y^T \begin{bmatrix} x^TCx+c & x^TBx+b \\ x^TBx+b & x^TAx+a \end{bmatrix} y \ge 0 \text{ for all } \|x\|^2 \le r \text{ and for all } y \in \Re^2_+$$
holds if and only if
$$\begin{bmatrix} C & B \\ B & A \end{bmatrix} \in \mathcal{L}_{2,m}, \quad \begin{bmatrix} rC+cI & rB+bI-E \\ rB+bI-E^T & rA+aI \end{bmatrix} \succeq 0, \quad \begin{bmatrix} c & b-e \\ b-e & a \end{bmatrix} \succeq 0$$
for some $e \ge 0$ and some E with $E+E^T \succeq 0$.
3. For symmetric matrices A, B, and C, the condition
$$y^T \begin{bmatrix} x^TCx+c & x^TBx+b \\ x^TBx+b & x^TAx+a \end{bmatrix} y \ge 0 \text{ for all } \|x\|^2 \le r \text{ and for all } y \in \Re^2_+ \text{ with } y_1 \ge y_2$$
holds if and only if
$$\begin{bmatrix} C & B \\ B & A \end{bmatrix} \in \mathcal{L}_{2,m}, \quad \begin{bmatrix} rC+cI & rB+bI-E \\ rB+bI-E^T & rA+aI+E+E^T \end{bmatrix} \succeq 0, \quad \begin{bmatrix} c & b-e \\ b-e & a+2e \end{bmatrix} \succeq 0$$
for some $e \ge 0$ and some E with $E+E^T \succeq 0$.

Similar equivalence relations hold for the case where $\|x\|^2 \le r$ is replaced with $\|x\|^2 = r$. In this case, we need only remove from the above equivalence relations the nonnegative parameter e and the respective conditions on the 2×2 matrix involving a, b, c, d.

It remains an open question whether one can obtain an LMI description for the general 2×2 robust QMIs over the unit ball without the cocenteredness condition.

For $\Delta = \Re^m$ with general m, however, checking the membership problem (4.21) is a hard problem.

THEOREM 4.9. *For general n and m, the (ϵ-approximate) membership problem*
$$\begin{bmatrix} C & 0 \\ 0 & A \end{bmatrix} \in \mathcal{C}_+(\mathrm{SOC}(n), \Re^m) \tag{4.25}$$
with data $(A, C) \in \mathcal{L}_{n,m} \times \mathcal{S}^m$ is co-NP-complete.

Proof. We choose to use the following well-known NP-complete partition problem for the purpose of reduction:

> Decide whether or not one can partition a given set of integers $a_1, \ldots, a_n$ such that the two subsets will have the same subset sum.

The above decision problem can be further reduced to the following decision problem:

> Given $a \in Z^n$ (the n-dimensional integer lattice) and a scalar $t \geq 0$, decide whether or not $p(x;t) \geq 0$ for all $\|x\|_2^2 = n$, where $p(x;t) = (t + (a^T x)^2)^2 - n^2 + \sum_{i=1}^n x_i^4$.

To see why it is so, we notice that for $t \geq 0$ and x with $\|x\|^2 = n$,

$$p(x;t) = (t + (a^T x)^2)^2 - n^2 + \sum_{i=1}^n x_i^4 \geq t^2 - n^2 + \sum_{i=1}^n x_i^4 \geq t^2 - n^2 + \frac{(\sum_{i=1}^n x_i^2)^2}{n} = t^2 - n(n-1),$$

where the second inequality is based on the Cauchy–Schwarz inequality.

The lower bound is attained, i.e., $p(x;t) = t^2 - n(n-1)$, if and only if $x_i^2 = 1$ for all $i = 1, \ldots, n$, and $a^T x = 0$, which is equivalent to the existence of a partition. Thus, a partition does *not* exist if and only if for $t = \sqrt{n(n-1)} - 1$ there holds $p(x;t) \geq 0$ for all $\|x\|_2^2 = n$.

Next we notice that

$$\|x\|_2^4 = \left(\sum_{i=1}^n x_i^2 \right)^2 = \sum_{i=1}^n x_i^4 + \sum_{i \neq j} x_i^2 x_j^2,$$

so that

$$p(x;t) = (t + (a^T x)^2)^2 + (\|x\|_2^4 - n^2) - \sum_{i \neq j} x_i^2 x_j^2,$$

where the second term vanishes if $\|x\|_2^2 = n$. It follows that $p(x;t) \geq 0$ for all $\|x\|_2^2 = n$ if and only if

$$t + (a^T x)^2 \geq \|\{x_i x_j\}_{i \neq j}\|_2 \text{ for all } x^T x = n. \tag{4.26}$$

The above robust quadratic SOC-constraint can be transformed into an equivalent robust QMI in the familiar way, namely,

$$L(t + (a^T x)^2, \{x_i x_j\}_{i \neq j}) \in \mathcal{S}_+^{1+n(n-1)} \text{ for all } x^T x = n, \tag{4.27}$$

where $L(\cdot, \cdot)$ denotes the so-called arrow-hat (or Jordan product representation) matrix

$$L(s, y) = \begin{bmatrix} s & y^T \\ y & sI \end{bmatrix}.$$

We have reduced the partitioning problem to the robust QMI (4.27), which is of form (4.25). □

COROLLARY 4.10. *The membership problem $X \in \mathcal{C}_+(\Re^n, \Re^m)$ is co-NP-complete.*

Proof. Due to Theorem 4.8, the co-NP-complete problem in Theorem 4.9 can be reduced to the membership problem for $\mathcal{C}_+(\Re^n, \Re^m)$, which must therefore also be co-NP-complete. □

To close the section, we remark that it is NP-hard in general to check whether a fourth order polynomial is nonnegative over the unit sphere (or over the whole space). The same partition problem as in the proof of Theorem 4.9 can also be used to reduce to the unconstrained minimization of the following fourth order polynomial:

$$\sum_{i=1}^{n}(x_i^2 - 1)^2 + (a^T x)^2.$$

In particular, a partition exists if and only if the polynomial attains zero. Now we pose as an open question the complexity of deciding whether a third order polynomial is nonnegative over the unit sphere. If this can be done in polynomial time, then the next question will be: Can we describe the set of the coefficients of such nonnegative third order polynomials (over the unit sphere) by (L)MIs?

5. Applications in robust linear programming. Robust optimization models in mathematical programming have received much attention recently; see, e.g., [3, 4, 9]. In this section we will discuss some of these models using the techniques developed in the previous sections.

Consider the following formulation of a robust linear program:

$$\begin{array}{ll} \text{Minimize} & \max_{\|\Delta x\| \le \delta, \|\Delta c\| \le \epsilon_0} (c + \Delta c)^T (x + \Delta x) \\ \text{subject to} & (a_i + \Delta a_i)^T (x + \Delta x) \ge (b_i + \Delta b_i) \\ & \text{for all } \|(\Delta a_i, \Delta b_i)\| \le \epsilon_i,\ i = 1, 2, \ldots, m,\ \|\Delta x\| \le \delta. \end{array} \tag{5.1}$$

Here two types of perturbation are considered. First, the problem data $(\{a_i\}, \{b_i\}, c)$ might be affected by unpredictable perturbation (e.g., measurement error). Second, the optimal solution x^{opt} is subject to implementation errors caused by the finite precision arithmetic of digital hardware. That is, we have $x^{actual} := x^{opt} + \Delta x$, where x^{actual} is the actually implemented solution. To ensure x^{actual} remains feasible and delivers a performance comparable to that of x^{opt}, we need to make sure x^{opt} is robust against both types of perturbation. This is essentially the motivation of the above robust linear programming model. Notice that our model is more general than the ones proposed by Ben-Tal and Nemirovskii [4] in that the latter only considers perturbation error in the data $(\{a_i\}, \{b_i\}, c)$.

The above model of robust linear programming arises naturally from the design of a linear phase FIR (finite impulse response) filter for digital signal processing. In particular, for a linear phase FIR filter $h = (h_1, \ldots, h_n) \in \Re^n$, the frequency response is

$$H(e^{j\omega}) = e^{-jn\omega}(h_1 + h_2 \cos\omega + \cdots + h_n \cos(n\omega)) = e^{-jn\omega}(\mathbf{cos}\ \omega)^T h,$$

where $\mathbf{cos}\ \omega = (1, \cos\omega, \ldots, \cos(n\omega))^T$. The FIR filter usually must satisfy a given spectral envelope constraint (typically specified by design requirement or industry standards); see Figure 1 for an example.

This gives

$$L(e^{-j\omega}) \le (\mathbf{cos}\ \omega)^T h \le U(e^{-j\omega}) \text{ for all } \omega \in [0, \pi]. \tag{5.2}$$

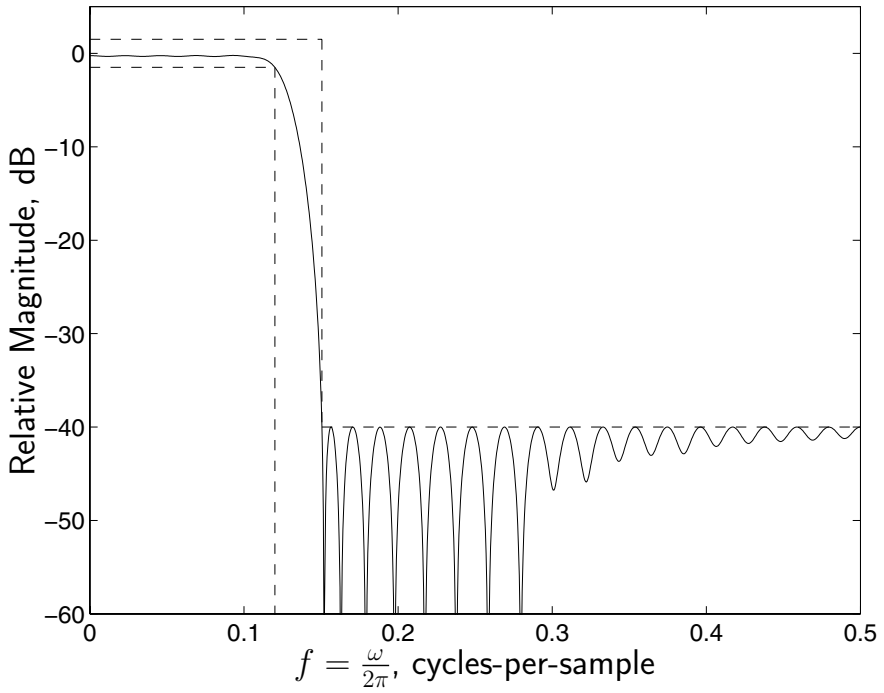


FIG. 1. *An example of spectral mask constraint.*

Finding a discrete h (say, a 4-bit integer) satisfying (5.2) is NP-hard. Ignoring the discrete structure of h, we can find an h satisfying (5.2) in polynomial time [6]. However, rounding such a solution to the nearest discrete h may degrade performance significantly. Our design strategy is then to first discretize the frequency $[0, \pi]$, then find a solution that is robust against discretization and rounding errors. This leads to the following notion of a robustly feasible solution:

$$(5.3)\quad L(e^{-j\omega_i}) \le (\mathbf{cos}\ \omega_i + \Delta_i)^T (h + \Delta h) \le U(e^{-j\omega_i}) \text{ for all } \|\Delta_i\| \le \epsilon,\ \|\Delta h\| \le \delta,$$

where Δ_i accounts for the discretization error, while Δh models the rounding errors.

We now reformulate the robust linear program (5.1) as a semidefinite program. We say the solution x is *robustly feasible* if, for all $i = 1, 2, \ldots, m$,

$$(a_i + \Delta a_i)^T (x + \Delta x) \ge (b_i + \Delta b_i) \text{ for all } \|(\Delta a_i, \Delta b_i)\| \le \epsilon_i,\ \|\Delta x\| \le \delta,\ i = 1, 2, \ldots, m.$$

It can be shown [4] that x is robustly feasible if and only if

$$(5.4)\quad a_i^T (x + \Delta x) - b_i - \epsilon_i \sqrt{\|x + \Delta x\|^2 + 1} \ge 0 \text{ for all } \|\Delta x\| \le \delta,\ i = 1, 2, \ldots, m.$$

Constraint (5.4) can be formulated as

$$(5.5)\quad \begin{bmatrix} (a_i^T (x + \Delta x) - b_i) I & \epsilon_i \begin{bmatrix} x + \Delta x \\ 1 \end{bmatrix} \\ \epsilon_i \begin{bmatrix} (x + \Delta x)^T & 1 \end{bmatrix} & a_i^T (x + \Delta x) - b_i \end{bmatrix} \succeq 0 \text{ for all } \|\Delta x\| \le \delta,\ i = 1, 2, \ldots, m.$$

Now the objective function can also be modeled by introducing an additional variable t to be minimized. At the same time we set as a constraint $t-(c+\Delta c)^T(x+\Delta x) \geq 0$ for all $\|\Delta c\| \leq \epsilon_0$ and $\|\Delta x\| \leq \delta$. Then the objective can be modeled by $t - c^T(x+\Delta x) \geq \epsilon_0 \|x+\Delta x\|$ for all $\|\Delta x\| \leq \delta$, which is equivalent to

$$\begin{bmatrix} (t-c^T(x+\Delta x))I & \epsilon_0(x+\Delta x) \\ \epsilon_0(x+\Delta x)^T & t-c^T(x+\Delta x) \end{bmatrix} \succeq 0 \text{ for all } \|\Delta x\| \leq \delta. \tag{5.6}$$

Using Proposition 3.3, we can show that (5.5) is equivalent to the following: There exists a $\mu_i \geq 0$ such that

$$\begin{bmatrix} (a_i^T x - b_i)I & \epsilon_i \begin{bmatrix} x \\ 1 \end{bmatrix} & \epsilon_i \begin{bmatrix} I \\ 0 \end{bmatrix} \\ \epsilon_i \begin{bmatrix} x^T & 1 \end{bmatrix} & a_i^T x - b_i & \frac{1}{2}a_i^T \\ \epsilon_i \begin{bmatrix} I \\ 0 \end{bmatrix} & \frac{1}{2}a_i & 0 \end{bmatrix} - \mu_i \begin{bmatrix} 0 & 0 & 0 \\ 0 & \delta^2 & 0 \\ 0 & 0 & -I \end{bmatrix} \succeq 0. \tag{5.7}$$

Similarly, (5.6) holds for all $\|\Delta x\| \leq \delta$ if and only if there is a $\mu_0 \geq 0$ such that

$$\begin{bmatrix} (t-c^T x)I & \epsilon_0 x & \epsilon_0 I \\ \epsilon_0 x^T & t-c^T x & -\frac{1}{2}c^T \\ \epsilon_0 I & -\frac{1}{2}c & 0 \end{bmatrix} - \mu_0 \begin{bmatrix} 0 & 0 & 0 \\ 0 & \delta^2 & 0 \\ 0 & 0 & -I \end{bmatrix} \succeq 0. \tag{5.8}$$

Therefore, the robust linear programming model becomes a semidefinite program: minimize t subject to (5.7) and (5.8).

Some computational results demonstrating the effectiveness of the robust linear programming formulation (5.1) have been reported recently in [12]. In particular, it was shown that for the robust magnitude filter design problem (5.3), the quantized versions of the robust filter still satisfy the spectral mask constraints, while the quantized nonrobust filters violate both the passband and the stopband spectral mask specifications. Thus, it is important to consider rounding errors in the robust filter design problem.

REFERENCES

[1] K. M. Anstreicher, X. Chen, H. Wolkowicz, and Y. Yuan, *Strong duality for a trust-region type relaxation of the quadratic assignment problem*, Linear Algebra Appl., 301 (1999), pp. 121–136.

[2] K. Anstreicher and H. Wolkowicz, *On Lagrangian relaxation of quadratic matrix constraints*, SIAM J. Matrix Anal. Appl., 22 (2000), pp. 41–55.

[3] A. Ben-Tal, L. El Ghaoui, and A. Nemirovskii, *Robustness*, in Handbook of Semidefinite Programming, H. Wolkowicz, R. Saigal, and L. Vandenberghe, eds., Kluwer Academic Publishers, Boston, MA, 2000, pp. 139–162.

[4] A. Ben-Tal and A. Nemirovski, *Robust convex optimization*, Math. Oper. Res., 23 (1998), pp. 769–805.

[5] S. Boyd, L. El Ghaoui, E. Feron, and V. Balakrishnan, *Linear Matrix Inequalities in System and Control Theory*, SIAM Stud. Appl. Math. 15, SIAM, Philadelphia, 1994.

[6] T. N. Davidson, Z.-Q. Luo, and J. Sturm, *Linear matrix inequality formulation of spectral mask constraints*, IEEE Trans. Signal Process., 50 (2002), pp. 2702–2715.

[7] Y. Genin, Yu. Nesterov, and P. Van Dooren, *Positive transfer functions and convex optimization*, in Proceedings of the 5th European Control Conference (ECC '99), Karlsruhe, 1999, Paper F-143. Also available online at http://nyquist.us.es/conferencias/ECC99/papers/F0143.pdf

[8] Y. Genin, Y. Hachez, Yu. Nesterov, and P. Van Dooren, *Optimization problems over positive pseudopolynomial matrices*, SIAM J. Matrix Anal. Appl., 25 (2003), pp. 57–79.
[9] L. El Ghaoui, F. Oustry, and H. Lebret, *Robust solutions to uncertain semidefinite programs*, SIAM J. Optim., 9 (1998), pp. 33–52.
[10] L. El Ghaoui and H. Lebret, *Robust solutions to least-squares problems with uncertain data*, SIAM J. Matrix Anal. Appl., 18 (1997), pp. 1035–1064.
[11] M. Grötschel, L. Lovász, and A. Schrijver, *Geometric Algorithms and Combinatorial Optimization*, Springer-Verlag, Berlin, New York, 1988.
[12] Z.-Q. Luo, *Applications of convex optimization in signal processing and digital communication*, Math. Program. Ser. B, 97 (2003), pp. 177–207.
[13] K. G. Murty and S. N. Kabadi, *Some NP-complete problems in quadratic and nonlinear programming*, Math. Program., 39 (1987), pp. 117–129.
[14] Yu. Nesterov, *Squared functional systems and optimization problems*, in High Performance Optimization, H. Frenk, K. Roos, T. Terlaky, and S. Zhang, eds., Kluwer Academic Publishers, Dordrecht, 2000, pp. 405–440.
[15] Yu. Nesterov and A. Nemirovsky, *Interior-Point Polynomial Methods in Convex Programming*, SIAM Stud. Appl. Math. 13, SIAM, Philadelphia, 1994.
[16] B. T. Polyak, *Convexity of quadratic transformations and its use in control and optimization*, J. Optim. Theory Appl., 99 (1998), pp. 553–583.
[17] J. F. Sturm and S. Zhang, *On cones of nonnegative quadratic functions*, Math. Oper. Res., 28 (2003), pp. 246–267.
[18] H. Wolkowicz, *A note on lack of strong duality for quadratic problems with orthogonal constraints*, European J. Oper. Res., 143 (2002), pp. 356–364.
[19] V. A. Yakubovich, *S-procedure in nonlinear control theory*, Vest. Leningr., Univ. 4 (1977), pp. 73–93.

SIAM J. OPTIM.
Vol. 14, No. 4, pp. 1163–1186

A ROBUST PRIMAL-DUAL INTERIOR-POINT ALGORITHM FOR NONLINEAR PROGRAMS*

XINWEI LIU† AND JIE SUN†

Abstract. We present a primal-dual interior-point algorithm for solving optimization problems with nonlinear inequality constraints. The algorithm has some of the theoretical properties of trust region methods, but works entirely by line search. Global convergence properties are derived without assuming regularity conditions. The penalty parameter ρ in the merit function is updated adaptively and plays two roles in the algorithm. First, it guarantees that the search directions are descent directions of the updated merit function. Second, it helps to determine a suitable search direction in a decomposed SQP step. It is shown that if ρ is bounded for each barrier parameter μ, then every limit point of the sequence generated by the algorithm is a Karush–Kuhn–Tucker point, whereas if ρ is unbounded for some μ, then the sequence has a limit point which is either a Fritz–John point or a stationary point of a function measuring the violation of the constraints. Numerical results confirm that the algorithm produces the correct results for some hard problems, including the example provided by Wächter and Biegler, for which many of the existing line search–based interior-point methods have failed to find the right answers.

Key words. nonlinear optimization, interior-point method, global convergence, regularity conditions

AMS subject classifications. 49M30, 49M37, 65K10, 90C22, 90C26, 90C30, 90C51

DOI. 10.1137/S1052623402400641

1. Introduction.

Applying an interior-point approach to nonlinear programming has been the subject of intensive studies in recent years; see [1, 4, 5, 11, 12, 15, 16, 18, 23, 24, 25, 27, 28, 29]. For simplicity of presentation, we concentrate in this paper on inequality constrained nonlinear programs

$$\text{minimize } f(x) \quad \text{subject to } c(x) \le 0, \tag{1.1}$$

where $c(x) = (c_1(x), \ldots, c_m(x))^\top$, $f : \Re^n \to \Re$, and $c : \Re^n \to \Re^m$. We do not assume any convexity on f and c. However, we suppose that f and c are twice continuously differentiable throughout this paper.

The interior-point approach solves, as $\mu \downarrow 0$, the barrier problems

$$\text{minimize } f(x) - \mu \sum_{i=1}^{m} \ln y_i \quad \text{subject to } c(x) + y = 0. \tag{1.2}$$

The direction-finding Newton equations then include

$$c(x) + y + \nabla c(x)^\top d_x + d_y = 0. \tag{1.3}$$

Note that (1.3) is always feasible even if the linearized inequality

$$c(x) + \nabla c(x)^\top d_x \le 0 \tag{1.4}$$

*Received by the editors January 5, 2002; accepted for publication (in revised form) December 1, 2003; published electronically July 20, 2004. This research was partially supported by the Singapore-MIT Alliance and grant R-314-000-042/057-112 of the National University of Singapore.
http://www.siam.org/journals/siopt/14-4/40064.html
†Department of Decision Sciences and Singapore-MIT Alliance, National University of Singapore, Singapore (smaliuxw@nus.edu.sg, jsun@nus.edu.sg).

may be inconsistent, which presents difficulties in convergence of interior-point–based methods. The examples discussed by Byrd, Marazzi, and Nocedal [7] and Wächter and Biegler [26] show that the interior-point methods using (1.3) may not find a feasible point of the original problem or a point with stationary properties. We also notice that the global convergence analysis of most existing interior-point methods requires rather strong assumptions on regularity at all iterates. Wächter and Biegler [26] indicate that these assumptions may not hold even though the local minima have very good regularity properties.

A remedy to these problems is to apply sequential quadratic programming (SQP) techniques to the barrier problems and to use a trust region strategy to ensure the robustness of the algorithm. Such algorithms have recently been proposed by Byrd, Gilbert, and Nocedal [4] and Tseng [24], for example. The numerical experiments in [5] show that the trust region–type algorithm is very promising.

We provide a different approach in this paper. Instead of introducing additional trust region constraints, we use refined line search rules to generate a new iterate in a decomposed SQP framework. The search direction is determined by either a Newton-type step or a Cauchy-type step with the choice being made with reference to a penalty parameter in the merit function. In addition, we adjust the penalty parameter of the merit function adaptively. As a result, we have been able to analyze convergence without regularity conditions and to avoid the convergence problems mentioned above. However, unlike the trust region methods, the algorithm does not have the flexibility to allow the direct use of indefinite second order derivatives.

The convergence properties of the algorithm can be summarized as follows. Let ρ_k be the value of the penalty parameter of the merit function at iterate k. If $\{\rho_k\}_{k=0}^{\infty}$ is bounded independent of the barrier parameter μ, then every convergent subsequence produced by the algorithm converges to a Karush–Kuhn–Tucker (KKT) point of the problem. If $\rho_k \to \infty$ for some μ, then the sequence has a limit point that is either feasible with linearly dependent gradients of the active constraints (i.e., a Fritz–John point) or infeasible but stationary with respect to the function $\|\max[0, c(x)]\|$, which is obviously a measure of the violation of the constraints (ℓ_2-infeasibility for short).

Besides, we show that, if the penalty parameters are bounded, then the algorithm generates the identical search directions with the original primal-dual methods such as LOQO (see Shanno and Vanderbei [23, 25]) after a finite number of iterations. Thus, superlinear convergence may be derived by existing works, such as [6, 29], under suitable conditions; while in the unbounded case, the algorithm may have linear convergence. For brevity, we mainly consider global convergence in this paper. By the same token, practical implementation techniques are not discussed. The interested reader is referred to the related literature, such as [6, 8, 11, 15, 16, 23, 25, 28, 29], for details.

Our numerical results show that the proposed algorithm can find solutions of the examples in [7, 26] and the least ℓ_2-infeasibility solution for an infeasible example in [3], among others.

The paper is organized as follows. In section 2, we present a two-step decomposition scheme of SQP and specify the requirement for an approximate solution to the resulting unconstrained penalty subproblems. In section 3, this scheme is applied to the barrier problem 1.2 and we present a modified primal-dual system of equations that is used in the algorithm for the barrier problem. The global convergence of the algorithm is analyzed in section 4. In section 5 we present the overall algorithm for problem (1.1) and its global convergence results. We provide some computational for-

mulae for the approximate solutions of the unconstrained penalty subproblems and report our preliminary numerical results in section 6.

We use standard notation from the literature of interior-point methods and nonlinear programming. For example, a letter with superscript k is related to the kth iteration; the subscript i is the ith component for a vector or the ith column for a matrix. The norm $\|\cdot\|$ is the Euclidean norm. We also use simplified notation, such as $f_k = f(x^k)$, $g^k = \nabla f(x^k)$, $c^k = c(x^k)$, and $A_k = \nabla c(x^k)$. For vector y, $Y = \mathrm{diag}(y)$ is the diagonal matrix whose ith diagonal element is y_i. All vector inequalities are understood componentwise. For two symmetric matrices A and B, $A \succ (\succeq) B$ means that $A - B$ is positive definite (semidefinite).

2. A decomposition scheme of SQP.

2.1. The basic idea. The barrier problem

$$\text{minimize } f(x) - \mu \sum_{i=1}^{m} \ln y_i \quad \text{subject to } c(x) + y = 0$$

is simply expressed as

$$\text{minimize } \psi_\mu(z) \tag{2.1}$$

$$\text{subject to } h(z) = 0, \tag{2.2}$$

where $z = (x, y)$, $h(z) = c(x) + y$, and $\psi_\mu(z) = f(x) - \mu \sum_{i=1}^{m} \ln y_i$. It is obvious that $\psi_\mu(z)$ is a continuously differentiable function for $y > 0$. At the current iteration point z, the SQP approach for (2.1)–(2.2) generates the search direction d_z by solving the quadratic programming problems

$$\text{minimize } \nabla\psi(z)^\top d + \frac{1}{2} d^\top Q d \tag{2.3}$$

$$\text{subject to } h(z) + \nabla h(z)^\top d = 0, \tag{2.4}$$

where Q is a positive definite approximation to the Lagrangian Hessian at z. Then the new iteration point z^+ is derived by a line search procedure,

$$z^+ = z + \alpha d_z, \tag{2.5}$$

where $\alpha \in (0, 1]$ is the steplength along d_z. This general framework requires regularity assumptions on $h(z)$ at all iterates. Otherwise, some of the slack variables may tend to zero too quickly and the algorithm may fail to find the right solution [26].

Our idea is rooted in the work of Fletcher [13, 14], Liu [19], and Yuan and Liu [20], although in the original works [19, 20] the authors need to *exactly* solve all the subproblems, including a nonsmooth unconstrained optimization problem. For the barrier problem, we first approximately solve the penalty optimization problem

$$\text{minimize}_{d \in \Re^n} \ \frac{1}{2} d^\top Q d + \rho \|h(z) + \nabla h(z)^\top d\|, \tag{2.6}$$

where $\rho > 0$ is the penalty parameter in the merit function

$$\phi(z; \rho) = \psi_\mu(z) + \rho \|h(z)\|. \tag{2.7}$$

Let $\tilde{d}_z$ be an approximate solution to (2.6). Then we generate the search direction d_z by solving the subproblem

$$\text{minimize } \nabla\psi(z)^\top d + \frac{1}{2}d^\top Q d \tag{2.8}$$

$$\text{subject to } \nabla h(z)^\top d = \nabla h(z)^\top \tilde{d}_z. \tag{2.9}$$

We consider subproblem (2.8)–(2.9) since it can provide us with the estimates of the multipliers, which are needed in the primal-dual interior-point approach. It can be proved (see Proposition 3.1) that, for sufficiently large ρ, the solution d_z to (2.8)–(2.9) is a descent direction of the merit function.

The idea is similar to the trust region interior-point method, in which the auxiliary step $\tilde{d}_z$ is generated by minimizing $\|h(z) + \nabla h(z)^\top d\|$ on a trust region; see [4, 9, 10, 21, 22]. Here, by adding a quadratic term, we remove the trust region constraint in deriving the auxiliary step for the modified system of primal-dual equations.

2.2. The approximate solution to subproblem (2.6). In this subsection we describe how to generate the approximate solution to subproblem (2.6). Subproblem (2.6) can be simply written as

$$\text{minimize } \ q(d) = \frac{1}{2}d^\top Q d + \rho\|r + R^\top d\|, \tag{2.10}$$

where $\rho > 0$, Q is any positive definite matrix, r is a vector, and R is a matrix with full column rank. It is easy to note that the exact solution is $d = 0$ if $r = 0$. Thus, in the following discussion, we assume that $r \neq 0$.

We generate the approximate solution $\tilde{d}_z$ to problem (2.10) by the following procedure.

PROCEDURE 2.1.

(1) *Compute the Q-weighted Newton step for minimizing $\|r + R^\top d\|$:*

$$\tilde{d}_z^N = -Q^{-1}R(R^\top Q^{-1}R)^{-1}r. \tag{2.11}$$

If $q(\tilde{d}_z^N) \le \nu q(0)$ ($\nu \in (0,1)$ is a fixed constant), then $\tilde{d}_z = \tilde{d}_z^N$; else go to (2).

(2) *Calculate the Q-weighted steepest descent step (Cauchy step)*

$$\tilde{d}_z^C = -Q^{-1}Rr. \tag{2.12}$$

Find $\tilde{d}_z$ in the subspace spanned by $\tilde{d}_z^N$ and $\tilde{d}_z^C$ (see details in section 6.1*) such that*

$$q(\tilde{d}_z) \le \max\{\nu q(0), q(\alpha^C \tilde{d}_z^C)\}, \tag{2.13}$$

where $\alpha^C = \operatorname{argmin}_{\alpha\in[0,1]} q(\alpha\tilde{d}_z^C)$.

Let us point out that, when our algorithm produces a sequence converging to a KKT point of the barrier problem, the Q-weighted Newton step will eventually be accepted under suitable conditions, so the direction-finding process (2.6)–(2.9) will generate an identical direction with the original primal-dual interior-point methods (see section 3). Intuitively, the Newton step can be rejected only if $q(\tilde{d}_z^N) > \nu q(0)$, namely,

$$\frac{1}{2}r^\top(R^\top Q^{-1}R)^{-1}r > \rho\nu\|r\|. \tag{2.14}$$

With a moderate value of ρ, if $R^\top Q^{-1}R$ is nonsingular, the above relationship indicates that $\|r\|$ is large, or at least is of the order of ρ. This cannot happen for an iterate close to a KKT point x^* since this iterate must be nearly feasible, i.e., $\|r\|$ must be small. Later, we will present more detailed analysis on this point (see Propositions 3.2 and 3.3).

We next provide a technical result on the decrement of the Cauchy step for later reference.

PROPOSITION 2.2. *There holds*

$$q(\alpha^C \tilde{d}_z^C) - q(0) \le \frac{1}{2}\left\{1 - \rho \min\left[\frac{1}{\|r\|}, \frac{\eta}{\|r\|}\right]\right\} r^\top (R^\top Q^{-1}R) r, \tag{2.15}$$

where $\eta = [r^\top (R^\top Q^{-1}R) r]/[r^\top (R^\top Q^{-1}R)^2 r]$.

Proof. Let $\chi(d) = \|r + R^\top d\|$. We have

$$\begin{aligned} \chi(0)^2 - \chi(\alpha \tilde{d}_z^C)^2 &= \|r\|^2 - \|(I - \alpha R^\top Q^{-1}R) r\|^2 \\ &= 2\alpha r^\top (R^\top Q^{-1}R) r - \alpha^2 r^\top (R^\top Q^{-1}R)^2 r. \end{aligned} \tag{2.16}$$

Suppose that $\tilde{\alpha} \in [0,1]$ minimizes $\chi(\alpha \tilde{d}_z^C)$. Then we have the following two cases:

(i) If $\eta \le 1$, then

$$\chi(0)^2 - \chi(\tilde{\alpha} \tilde{d}_z^C)^2 = \eta r^\top (R^\top Q^{-1}R) r, \tag{2.17}$$

which implies that

$$\chi(0) - \chi(\tilde{\alpha} \tilde{d}_z^C) \ge \frac{\eta}{2\|r\|} r^\top (R^\top Q^{-1}R) r. \tag{2.18}$$

(ii) If $\eta > 1$, then $\tilde{\alpha} = 1$ and $r^\top (R^\top Q^{-1}R) r > r^\top (R^\top Q^{-1}R)^2 r$; thus

$$\chi(0) - \chi(\tilde{\alpha} \tilde{d}_z^C) \ge \frac{1}{2\|r\|} r^\top (R^\top Q^{-1}R) r. \tag{2.19}$$

Then it follows from (2.18), (2.19), and $\tilde{\alpha} \le 1$ that

$$q(\tilde{\alpha} \tilde{d}_z^C) - q(0) \le \frac{1}{2}\left\{1 - \rho \min\left[\frac{1}{\|r\|}, \frac{\eta}{\|r\|}\right]\right\} r^\top (R^\top Q^{-1}R) r. \tag{2.20}$$

Since $q(\alpha^C \tilde{d}_z^C) \le q(\tilde{\alpha} \tilde{d}_z^C)$, we obtain (2.15). ∎

3. The algorithm for the barrier problem. We now specialize the formulae in the last section to the barrier problem (1.2) and present a modified primal-dual system of equations for generating the search directions. Later, based on this modification, we will propose our algorithm for the barrier problem.

By writing z as (x, y), $\psi_\mu(z)$ as $\psi_\mu(x, y)$, *and* $h(z)$ as $h(x, y)$, the barrier problem is

$$\text{minimize } \psi_\mu(x, y) = f(x) - \mu \sum_{i=1}^{m} \ln y_i \tag{3.1}$$

$$\text{subject to } h(x, y) = c(x) + y = 0, \tag{3.2}$$

where $y = (y_1, \ldots, y_m)^\top > 0$, and μ is a fixed positive scalar. The Lagrangian of problem (3.1)–(3.2) is

$$L(x, y, \lambda) = \psi_\mu(x, y) + \lambda^\top h(x, y), \tag{3.3}$$

and its Hessian is

$$\nabla^2 L(x,y,\lambda) = \begin{pmatrix} \nabla^2 \ell(x,\lambda) & \\ & \mu Y^{-2} \end{pmatrix}, \tag{3.4}$$

where $\lambda \in \Re^m$ is a multiplier vector associated with (3.2) and $\ell(x,\lambda) = f(x) + \lambda^\top c(x)$. The KKT conditions of program (3.1)–(3.2) can be written as

$$F_\mu(x,y,\lambda) = \begin{pmatrix} g(x) + A(x)\lambda \\ Y\Lambda e - \mu e \\ c(x) + y \end{pmatrix} = 0, \tag{3.5}$$

where $g(x) = \nabla f(x)$, $A(x) = \nabla c(x)$, $Y = \operatorname{diag}(y)$, $\Lambda = \operatorname{diag}(\lambda)$, and $e = (1,\ldots,1)^\top$.

Byrd, Marazzi, and Nocedal [7] showed that the algorithm using the norm of the residual function $\|F_\mu(x,y,\lambda)\|$ as the merit function may fail in converging to a stationary point of the problem. In this paper, as mentioned in (2.7), our merit function is

$$\phi_\mu(x,y;\rho) = \psi_\mu(x,y) + \rho\|h(x,y)\|, \tag{3.6}$$

where $\rho > 0$ is the penalty parameter and is updated automatically during the iterations. Then we have the following result.

PROPOSITION 3.1. *For any $\rho \geq 0$, $y > 0$, and $(d_x, d_y) \in \Re^{n+m}$, the directional derivative $\phi_\rho'((x,y);(d_x,d_y))$ of $\phi_\mu(x,y;\rho)$ along (d_x,d_y) exists, and*

$$\phi_\rho'((x,y);(d_x,d_y)) \leq \pi_\rho((x,y);(d_x,d_y)), \tag{3.7}$$

where

$$\begin{aligned} &\pi_\rho((x,y);(d_x,d_y)) \\ &\quad = g(x)^\top d_x - \mu e^\top Y^{-1} d_y + \rho(\|c(x) + y + A(x)^\top d_x + d_y\| - \|c(x) + y\|). \end{aligned} \tag{3.8}$$

Proof. The first term on the right-hand side of (3.6), ψ_μ, is continuously differentiable. Its directional derivative is

$$\psi_\mu'((x,y);(d_x,d_y)) = g(x)^\top d_x - \mu e^\top Y^{-1} d_y. \tag{3.9}$$

Let $\theta(x,y) = \|h(x,y)\|$. Its directional differentiability follows from its convexity. Since

$$\begin{aligned} &\theta'((x,y);(d_x,d_y)) \\ &\quad = \lim_{\alpha \downarrow 0} \frac{[\theta(x + \alpha d_x, y + \alpha d_y) - \theta(x,y)]}{\alpha} \\ &\quad = \lim_{\alpha \downarrow 0} \frac{\left[\|c(x) + \alpha A(x)^\top d_x + y + \alpha d_y + o(\alpha)\| - \|c(x) + y\|\right]}{\alpha} \\ &\quad \leq \lim_{\alpha \downarrow 0} \left[\frac{\|c(x) + y + \alpha(A(x)^\top d_x + d_y)\| - \|c(x) + y\|}{\alpha} + \frac{\|o(\alpha)\|}{\alpha} \right] \\ &\quad \leq \|c(x) + y + A(x)^\top d_x + d_y\| - \|c(x) + y\| + \lim_{\alpha \downarrow 0} \frac{o(\alpha)}{\alpha}, \end{aligned}$$

where the last two inequalities follow from the triangle inequality and the convexity of the norm, the result follows immediately. ∎

Suppose that (x^k, y^k) is the current iteration point and λ^k is the corresponding approximation of the multiplier vector. For problem (3.1)–(3.2), by substituting

$$Q = \begin{pmatrix} B_k & \\ & Y_k^{-1}\Lambda_k \end{pmatrix},\ R = \begin{pmatrix} A_k \\ I \end{pmatrix},\ d = \begin{pmatrix} d_x \\ d_y \end{pmatrix} \text{ and } r = (c^k + y^k) \tag{3.10}$$

into (2.10), our approach first approximately solves the problem

$$\text{minimize } q_k(d_x, d_y) = \frac{1}{2} d_x^\top B_k d_x + \frac{1}{2} d_y^\top S_k d_y + \rho_k \|c^k + y^k + A_k^\top d_x + d_y\|, \tag{3.11}$$

where $B_k \succ 0$ is an approximation to matrix $\nabla^2 \ell(x^k, \lambda^k)$, $S_k = Y_k^{-1}\Lambda_k$, $Y_k = \operatorname{diag}(y^k)$, $\Lambda_k = \operatorname{diag}(\lambda^k)$, $c^k = c(x^k)$, and $A_k = A(x^k)$, and ρ_k is the current value of the penalty parameter. The Q-weighted Newton step and the Q-weighted steepest descent step defined in Procedure 2.1 are, respectively,

$$(\tilde{d}_x^k)^N = -B_k^{-1} A_k (A_k^\top B_k^{-1} A_k + S_k^{-1})^{-1} (c^k + y^k), \tag{3.12}$$

$$(\tilde{d}_y^k)^N = -S_k^{-1} (A_k^\top B_k^{-1} A_k + S_k^{-1})^{-1} (c^k + y^k), \tag{3.13}$$

and

$$(\tilde{d}_x^k)^C = -B_k^{-1} A_k (c^k + y^k), \quad (\tilde{d}_y^k)^C = -S_k^{-1} (c^k + y^k). \tag{3.14}$$

Let $(\tilde{d}_x^k, \tilde{d}_y^k)$ be the approximate solution obtained through Procedure 2.1. We generate the search direction (d_x^k, d_y^k) for the new iterate by solving

$$\text{minimize } (g^k)^\top d_x - \mu e^\top Y_k^{-1} d_y + \frac{1}{2} d_x^\top B_k d_x + \frac{1}{2} d_y^\top S_k d_y \tag{3.15}$$

$$\text{subject to } A_k^\top d_x + d_y = A_k^\top \tilde{d}_x^k + \tilde{d}_y^k, \tag{3.16}$$

where $g^k = \nabla f(x^k)$. Since $(\tilde{d}_x^k, \tilde{d}_y^k)$ is a feasible solution to problem (3.15)–(3.16), by (3.8), we have the formula

$$\begin{aligned} &\pi_{\rho_k}(\ (x^k, y^k); (d_x^k, d_y^k)) + \frac{1}{2} (d_x^k)^\top B_k d_x^k + \frac{1}{2} (d_y^k)^\top S_k d_y^k \\ &\qquad \le \pi_{\rho_k}((x^k, y^k); (\tilde{d}_x^k, \tilde{d}_y^k)) + \frac{1}{2} (\tilde{d}_x^k)^\top B_k \tilde{d}_x^k + \frac{1}{2} (\tilde{d}_y^k)^\top S_k \tilde{d}_y^k, \end{aligned} \tag{3.17}$$

which plays an important role in our later global convergence analysis for the case $\rho_k \to \infty$.

The KKT conditions of problem (3.15)–(3.16) are

$$B_k d_x + A_k \tilde{\lambda} = -g^k, \tag{3.18}$$

$$S_k d_y + \quad \tilde{\lambda} = \mu Y_k^{-1} e, \tag{3.19}$$

$$A_k^\top d_x + \ \ d_y = A_k^\top \tilde{d}_x^k + \tilde{d}_y^k, \tag{3.20}$$

which, by letting $d_\lambda = \tilde{\lambda} - \lambda^k$, can be equivalently written as the modified primal-dual system of equations

$$B_k d_x + A_k d_\lambda = -(g^k + A_k \lambda^k), \tag{3.21}$$

$$\Lambda_k d_y + \ Y_k d_\lambda = -(Y_k \Lambda_k e - \mu e), \tag{3.22}$$

$$A_k^\top d_x + \ \ d_y = A_k^\top \tilde{d}_x^k + \tilde{d}_y^k. \tag{3.23}$$

It is well known that the original primal-dual interior-point approach generates the search direction by solving the system of equations

$$B_k d_x + A_k d_\lambda = -(g^k + A_k \lambda^k), \tag{3.24}$$

$$\Lambda_k d_y + Y_k d_\lambda = -(Y_k \Lambda_k e - \mu e), \tag{3.25}$$

$$A_k^\top d_x + d_y = -(c^k + y^k), \tag{3.26}$$

which follows from the Newton method applied to (3.5); for example, see [11, 16, 23, 25, 28]. Then we have the following results.

PROPOSITION 3.2. *The modified approach using* (3.21)–(3.23) *generates the same search directions as the original primal-dual interior-point methods using* (3.24)–(3.26) *if the weighted Newton step* (3.12)–(3.13) *is used.*

Proof. If $\tilde{d}_x^k = (\tilde{d}_x^k)^N$ and $\tilde{d}_y^k = (\tilde{d}_y^k)^N$, then $A_k^\top \tilde{d}_x^k + \tilde{d}_y^k = -(c^k + y^k)$. Thus the system (3.21)–(3.23) is the same as the system (3.24)–(3.26). ∎

PROPOSITION 3.3. *Suppose that the two sets* $\{(x^k, y^k)\}_{k=0}^\infty$ *and* $\{(A_k^\top B_k^{-1} A_k + S_k^{-1})^{-1}\}_{k=0}^\infty$ *are bounded. Then there exists a positive constant* $\hat{\rho}$ *(which is not dependent on* k*) such that for* $\rho_k \geq \hat{\rho}$*, the Newton step* $((\tilde{d}_x^k)^N, (\tilde{d}_y^k)^N)$ *defined in* (3.12)–(3.13) *will be accepted by Procedure* 2.1*.*

Proof. We have

$$\begin{aligned} & q_k((\tilde{d}_x^k)^N, (\tilde{d}_y^k)^N) - \nu q_k(0,0) \\ & \quad = \frac{1}{2}(c^k + y^k)^\top (A_k^\top B_k^{-1} A_k + S_k^{-1})^{-1}(c^k + y^k) - \nu \rho_k \|c^k + y^k\| \\ & \quad \leq \left[\frac{1}{2}\|(A_k^\top B_k^{-1} A_k + S_k^{-1})^{-1}(c^k + y^k)\| - \nu \rho_k\right] \|c^k + y^k\|. \end{aligned} \tag{3.27}$$

By the assumptions of the proposition, there exists a constant $\hat{\rho} > 0$ such that for all k we have

$$\|(A_k^\top B_k^{-1} A_k + S_k^{-1})^{-1}(c^k + y^k)\| \leq 2\nu\hat{\rho}. \tag{3.28}$$

Thus, for every $\rho_k \geq \hat{\rho}$, $q_k((\tilde{d}_x^k)^N, (\tilde{d}_y^k)^N) \leq \nu q_k(0,0)$. ∎

In the following, we describe our algorithm for the barrier problem (3.1)–(3.2), which solves the problem (3.11) and the system of equations (3.21)–(3.23) at each iteration.

ALGORITHM 3.4 (the algorithm for problem (3.1)–(3.2)).

Step 1. *Given* $(x^0, y^0, \lambda^0) \in \Re^n \times \Re^m_{++} \times \Re^m_{++}$, $0 \prec B_0 \in \Re^{n\times n}$, $0 < \beta_1 < 1 < \beta_2$, $\rho_0 > 0$, $0 < \delta < 1$, $0 < \sigma_0 < \frac{1}{2}$, $\epsilon_1 > 0$, $\epsilon_2 > \epsilon_3 > 0$. *Let* $k := 0$.

Step 2. *Compute an approximate solution* $(\tilde{d}_x^k, \tilde{d}_y^k)$ *of problem* (3.11) *by Procedure* 2.1 *(see section* 6.1 *on its implementation).*

Step 3. *Calculate the search direction* $(d_x^k, d_y^k, d_\lambda^k)$ *by solving the system of equations* (3.21)–(3.23).

Step 4 *(update* ρ_k*). If*

$$\pi_{\rho_k}((x^k, y^k); (d_x^k, d_y^k)) \leq -\frac{1}{2}(d_x^k)^\top B_k d_x^k - \frac{1}{2}(d_y^k)^\top S_k d_y^k, \tag{3.29}$$

then set $\rho_{k+1} = \rho_k$*; otherwise, we update* ρ_k *by*

$$\rho_{k+1} = \max\left\{\frac{\psi'_\mu((x^k, y^k); (d_x^k, d_y^k)) + \frac{1}{2}(d_x^k)^\top B_k d_x^k + \frac{1}{2}(d_y^k)^\top S_k d_y^k}{\Delta_k},\ 2\rho_k\right\}, \tag{3.30}$$

where

$$\pi_{\rho_k}((x^k,y^k);(d_x^k,d_y^k)) = (g^k)^\top d_x^k - \mu e^\top Y_k^{-1} d_y^k - \rho_k \Delta_k \tag{3.31}$$

and

$$\Delta_k = \|c^k + y^k\| - \|c^k + y^k + A_k^\top d_x^k + d_y^k\|. \tag{3.32}$$

Step 5 *(line search). Compute*

$$\hat{\alpha}_k = \frac{-0.995}{\min\{(y_i^k)^{-1}(d_y^k)_i, i = 1,\dots,m; -0.995\}}. \tag{3.33}$$

Select the least nonnegative integer l *such that*

$$\begin{aligned}&\phi_\mu(x^k + \delta^l\hat{\alpha}_k d_x^k, y^k + \delta^l\hat{\alpha}_k d_y^k; \rho_{k+1}) - \phi_\mu(x^k, y^k; \rho_{k+1})\\ &\quad\le \sigma_0\delta^l\hat{\alpha}_k\pi_{\rho_k}((x^k,y^k);(d_x^k,d_y^k)).\end{aligned} \tag{3.34}$$

Let $\alpha_k = \delta^l\hat{\alpha}_k$. *The new primal iterate is generated as*

$$x^{k+1} = x^k + \alpha_k d_x^k, \tag{3.35}$$

$$y^{k+1} = \max\{y^k + \alpha_k d_y^k, -c^{k+1}\}. \tag{3.36}$$

Step 6 *(update dual iterate). If there exists* $\gamma \in [0,1]$ *such that*

$$\beta_1\mu e \le Y_{k+1}(\Lambda_k + \gamma D_\lambda^k)e \le \beta_2\mu e, \tag{3.37}$$

where $D_\lambda^k = diag(d_\lambda^k)$, *then we select the maximum* $\gamma_k \in [0,1]$ *satisfying* (3.37) *and then update* λ^k *by*

$$\lambda^{k+1} = \lambda^k + \gamma_k d_\lambda^k; \tag{3.38}$$

otherwise, we increase l *by* 1 *successively such that* (3.37) *holds, and then update the primal and dual iterates in the same way as in* (3.35), (3.36), *and* (3.38).

Step 7 *(check the stopping criteria). We terminate the algorithm if one of the following conditions is satisfied:*

(i) $\|F_\mu(x^{k+1}, y^{k+1}, \lambda^{k+1})\| < \epsilon_1$;

(ii) $\|c^{k+1} + y^{k+1}\| \ge \epsilon_2$ *and* $\|\begin{pmatrix} A_{k+1} \\ Y_{k+1}\end{pmatrix}(c^{k+1} + y^{k+1})\| < \epsilon_3$;

(iii) $\|c^{k+1} + y^{k+1}\| < \epsilon_3$ *and* $\det(A_{\mathcal{I}_{k+1}}^\top A_{\mathcal{I}_{k+1}}) < \epsilon_3$,

where $\mathcal{I}_{k+1} = \{i | c_i^{k+1} \ge -\epsilon_3\}$, *and* $A_{\mathcal{I}_{k+1}}$ *is a submatrix of* A_{k+1} *consisting of all columns indexed by* $\mathcal{I}_{k+1}$. *Else update the approximate Hessian* B_k *by* B_{k+1}, *let* $k := k+1$, *and go to Step* 2.

We make the following remarks on the algorithm:

- The new primal and dual iterates are generated, respectively, by using different steplengths. Such a strategy has been used in [8, 28, 29]. We hope that $\gamma_k = 1$ can be accepted even if $\alpha_k < 1$.
- By (3.33), we have $y^k + \hat{\alpha}_k d_y^k \ge 0.005 y^k$. If $d_{yi}^k \ge 0$, we have $y_i^k + \alpha_k d_{yi}^k \ge y_i^k$; else $\alpha_k d_{yi}^k \ge \hat{\alpha}_k d_{yi}^k$ since $\alpha_k \le \hat{\alpha}_k$. Thus we always have $y^{k+1} \ge 0.005 y^k$ by (3.36).

- Formula (3.36) was first introduced in [4]; a similar, but more sophisticated, technique is also used in [24]. Since $y^{k+1} \geq y^k + \alpha_k d_y^k$ and $\|c^{k+1} + y^{k+1}\| \leq \|c^{k+1} + y^k + \alpha_k d_y^k\|$, we have

$$\begin{aligned} &\phi_\mu(x^{k+1}, y^{k+1}; \rho_{k+1}) - \phi_\mu(x^k, y^k; \rho_{k+1}) \\ &\quad \leq \phi_\mu(x^k + \alpha_k d_x^k, y^k + \alpha_k d_y^k; \rho_{k+1}) - \phi_\mu(x^k, y^k; \rho_{k+1}); \end{aligned} \tag{3.39}$$

  thus $\phi_\mu(x^{k+1}, y^{k+1}; \rho_{k+1}) \leq \phi_\mu(x^k, y^k; \rho_{k+1})$ for all $k \geq 0$.
- A way to implement (3.37) will be introduced in section 6.2. The well-definedness of this step is shown in Lemma 4.4.
- Since we do not assume any regularity on the constraints, the stopping condition (i) may never hold, in which case the algorithm will terminate at condition (ii) or (iii) of Step 7 by the convergence results in the next section.

4. The analysis of global convergence. The global convergence of Algorithm 3.4 is analyzed in this section. Suppose that in the algorithm the tolerance ϵ_2 is small, tolerances ϵ_1 and ϵ_3 are very small, and an infinite sequence $\{(x^k, y^k, \lambda^k)\}$ is generated.

We need the following blanket assumption for all analysis in what follows.

ASSUMPTION 4.1.

(1) *Functions f and c are twice continuously differentiable functions on $\Re^n$.*

(2) *The set $\{x^k\}_{k=0}^{\infty}$ is bounded.*

(3) *There exist positive constants ν_1 and ν_2 such that $\nu_1 I \preceq B_k \preceq \nu_2 I$ for all k, where I stands for the identity matrix.*

Assumptions (1) and (2) are used in the convergence analysis of most algorithms for nonlinear programming. Assumption (3) guarantees the existence of the solution of system (3.21)–(3.23). Similar assumptions are also used by most line search–based interior-point methods for nonlinear programming. An exception is [8], in which the global convergence results are derived by assuming B_k to be uniformly positive definite and bounded on the null space of the linear equality constraints.

By Algorithm 3.4, for each integer $k \geq 0$, we have either $\rho_{k+1} = \rho_k$ or $\rho_{k+1} \geq 2\rho_k$. Thus, the sequence $\{\rho_k\}$ is a monotonically nondecreasing sequence.

LEMMA 4.2. *If there exist a positive integer $\hat{k}$ and a positive constant $\hat{\rho}$ such that $\rho_k = \hat{\rho}$ for all $k \geq \hat{k}$, then we have that*

(i) *both $\{y^k\}$ and $\{\lambda^k\}$ are bounded above and componentwise bounded away from zero. The same is true for the diagonal of S_k.*

(ii) *$\{(d_x^k, d_y^k, d_\lambda^k)\}$ is bounded.*

Proof. Without loss of generality, we suppose that $\rho_k = \hat{\rho}$ for all $k \geq 0$. By (3.34) and (3.39), $\phi_\mu(x^k, y^k; \hat{\rho})$ is monotonically decreasing; thus $\phi_\mu(x^k, y^k; \hat{\rho}) \leq \phi_\mu(x^0, y^0; \hat{\rho})$ for all k. Now we prove that y^k is bounded above by contradiction. Suppose that $\max_i\{y_i^k\} \to \infty$. We have also that

$$f_k - \mu \sum_{i=1}^{m} \ln y_i^k + \hat{\rho}\|c^k + y^k\| \leq \phi_\mu(x^0, y^0; \hat{\rho}). \tag{4.1}$$

Dividing both sides of (4.1) by $\max_i\{y_i^k\}$ and taking the limit when $k \to \infty$, we have that $\hat{\rho} \leq 0$ since each term approaches zero except $\lim_{k\to\infty} \|c^k + y^k\| / \max_i\{y_i^k\} \geq 1$. This is a contradiction.

By the fact that x^k and y^k are bounded and that

$$-\mu \sum_{i=1}^{m} \ln y_i^k \leq -f_k - \hat{\rho}\|c^k + y^k\| + \phi_\mu(x^0, y^0; \hat{\rho}), \tag{4.2}$$

y^k is componentwise bounded away from zero. It follows from (3.37) that λ^k is bounded above and componentwise bounded away from zero; so is the diagonal of S_k since $S_k = Y_k^{-1}\Lambda_k$.

(ii) By Assumption 4.1(3), matrix $\hat{B}_k = B_k + A_k Y_k^{-1}\Lambda_k A_k^\top$ is invertible. By simple computation, the system (3.21)–(3.23) can be written as

$$\begin{pmatrix} B_k & A_k \\ A_k^\top & -\Lambda_k^{-1}Y_k \end{pmatrix}\begin{pmatrix} d_x^k \\ d_\lambda^k \end{pmatrix} = \begin{pmatrix} -(g^k + A_k\lambda^k) \\ (Y_k - \mu\Lambda_k^{-1})e + (A_k^\top \tilde{d}_x^k + \tilde{d}_y^k) \end{pmatrix} \tag{4.3}$$

and

$$d_y^k = (\mu\Lambda_k^{-1} - Y_k)e - \Lambda_k^{-1}Y_k d_\lambda^k. \tag{4.4}$$

Since

$$\begin{pmatrix} B_k & A_k \\ A_k^\top & -\Lambda_k^{-1}Y_k \end{pmatrix}^{-1} = \begin{pmatrix} \hat{B}_k^{-1} & \hat{B}_k^{-1}A_k Y_k^{-1}\Lambda_k \\ \Lambda_k Y_k^{-1}A_k^\top \hat{B}_k^{-1} & P_k \end{pmatrix}, \tag{4.5}$$

where $P_k = -Y_k^{-1}\Lambda_k + Y_k^{-1}\Lambda_k A_k^\top \hat{B}_k^{-1} A_k Y_k^{-1}\Lambda_k$, the boundedness of (d_x^k, d_λ^k) follows from (4.3). By (4.4), d_y^k is bounded. □

By Lemma 4.2, there exist constants $b_1 > 0$ and $b_2 > 0$ such that $y^k \geq b_1 e$ and $\|d_y^k\| \leq b_2$ for all k. If $\hat{\alpha}_1 = \min\{1, 0.995b_1/b_2\}$, then $y^k + \hat{\alpha}_1 d_y^k \geq 0.005y^k$. Thus, for all $\alpha \in [0, \hat{\alpha}_1]$,

$$y^k + \alpha d_y^k \geq 0.005y^k. \tag{4.6}$$

LEMMA 4.3. *If $\{\rho_k\}$ is bounded, then there is a constant $\hat{\alpha}_2 \in (0, \hat{\alpha}_1]$ such that, for every $\alpha \in (0, \hat{\alpha}_2]$ and for all $k \geq 0$, there holds that*

$$\phi_\mu(x^k + \alpha d_x^k, y^k + \alpha d_y^k; \rho_{k+1}) - \phi_\mu(x^k, y^k; \rho_{k+1}) \leq \alpha\sigma_0\pi_{\rho_{k+1}}((x^k, y^k); (d_x^k, d_y^k)). \tag{4.7}$$

Proof. Without loss of generality, we suppose that $\rho_k = \hat{\rho}$ for all $k \geq 0$. Then (3.29) holds at all iterates. For $\alpha \in (0, \hat{\alpha}_1]$, by (4.6), we have

$$(Y_k + \alpha D_y^k)^{-1} \preceq 200Y_k^{-1}, \tag{4.8}$$

where $D_y^k = \mathrm{diag}(d_y^k)$. Thus, for $\alpha \in (0, \hat{\alpha}_1]$,

$$\begin{aligned} &-\sum_{i=1}^m \ln[y_i^k + \alpha(d_y^k)_i] + \sum_{i=1}^m \ln y_i^k + \alpha e^\top {Y_k}^{-1} d_y^k \\ &\quad = e^\top \int_0^\alpha [Y_k^{-1} - (Y_k + tD_y^k)^{-1}] d_y^k dt \\ &\quad = e^\top \int_0^\alpha Y_k^{-1}(Y_k + tD_y^k)^{-1}(tD_y^k) d_y^k dt \leq 100\alpha^2 \|{Y_k}^{-1} d_y^k\|^2. \end{aligned} \tag{4.9}$$

Since f and c are twice continuously differentiable, there are positive constants b_3 and b_4 such that

$$f(x^k + \alpha d_x^k) - f(x^k) - \alpha g(x^k)^\top d_x^k \leq \frac{1}{2}\alpha^2 b_3 \|d_x^k\|^2 \tag{4.10}$$

and

$$\begin{aligned}&\|c(\ x^k+\alpha d_x^k)+y^k+\alpha d_y^k\|-\|c(x^k)+y^k+\alpha A(x_k)^\top d_x^k+\alpha d_y^k\|\\&\le\|c(x^k+\alpha d_x^k)-c(x^k)-\alpha A(x^k)^\top d_x^k\|\le\frac{1}{2}\alpha^2 b_4\|d_x^k\|^2.\end{aligned}\tag{4.11}$$

The constants b_3 and b_4 are the first order Lipschitzian constants of f and c, respectively.

Let $b_5=\max\{100\mu,\frac{1}{2}(b_3+\hat\rho b_4)\}$. Since

$$\begin{aligned}&\pi_{\hat\rho}((x^k,y^k);(\alpha d_x^k,\alpha d_y^k))\\&=\alpha\psi_\mu^{'}((x^k,y^k);(d_x^k,d_y^k))+\hat\rho(\|c^k+y^k+\alpha A_k^\top d_x^k+\alpha d_y^k\|-\|c^k+y^k\|)\end{aligned}\tag{4.12}$$

by (3.8), it follows from (4.9), (4.10), and (4.11) that

$$\begin{aligned}&\phi_\mu\ (x^k+\alpha d_x^k,y^k+\alpha d_y^k;\hat\rho)-\phi_\mu(x^k,y^k;\hat\rho)-\pi_{\hat\rho}((x^k,y^k);(\alpha d_x^k,\alpha d_y^k))\\&\le\alpha^2 b_5(\|d_x^k\|^2+\|Y_k^{-1}d_y^k\|^2).\end{aligned}\tag{4.13}$$

It is easy to note that $\pi_{\hat\rho}((x^k,y^k);(\alpha d_x^k,\alpha d_y^k))$ is a convex function on $\alpha\in[0,1]$. Thus, we have

$$\pi_{\hat\rho}((x^k,y^k);(\alpha d_x^k,\alpha d_y^k))-\alpha\pi_{\hat\rho}((x^k,y^k);(d_x^k,d_y^k))\le 0,\tag{4.14}$$

and as a result,

$$\begin{aligned}&\pi_{\hat\rho}((x^k,y^k);(\alpha d_x^k,\alpha d_y^k))-\alpha\sigma_0\pi_{\hat\rho}((x^k,y^k);(d_x^k,d_y^k))\\&\quad\le\alpha(1-\sigma_0)\pi_{\hat\rho}((x^k,y^k);(d_x^k,d_y^k))\\&\quad\le-\frac{1}{2}\alpha(1-\sigma_0)\hat\delta(\|d_x^k\|^2+\|Y_k^{-1}d_y^k\|^2),\end{aligned}\tag{4.15}$$

where $\hat\delta=\min\{\nu_1,\beta_1\mu\}$ and the last inequality follows from (3.29), Assumption 4.1(3), and (3.37).

Let $\hat\alpha_2=\min\{\hat\alpha_1,(1-\sigma_0)\hat\delta/(2b_5)\}$. Then, by (4.13) and (4.15), (4.7) holds for all $\alpha\in[0,\hat\alpha_2]$ and $k\ge 0$. □

LEMMA 4.4. *Under the assumption of* Lemma 4.2*, if* $\beta_1\mu e\le Y_k\Lambda_k e\le\beta_2\mu e$*, then there exists a constant* $\hat\alpha_3\in(0,1]$ *such that*

$$\beta_1\mu e\le(\Lambda_k+\alpha D_\lambda^k)\max\{y^k+\alpha d_y^k,-c(x^k+\alpha d_x^k)\}\le\beta_2\mu e\tag{4.16}$$

for all $\alpha\in[0,\hat\alpha_3]$ *and all* k.

Proof. At first, we prove that

$$\beta_1\mu e\le(Y_k+\alpha D_y^k)(\Lambda_k+\alpha D_\lambda^k)e\le\beta_2\mu e\tag{4.17}$$

for all $\alpha\in[0,\bar\alpha_3]$ and all k, where $\bar\alpha_3\in(0,1]$ is a constant.

By (3.22), we have $(Y_k+\alpha D_y^k)(\Lambda_k+\alpha D_\lambda^k)e=\alpha\mu e+(1-\alpha)Y_k\Lambda_k e+\alpha^2 D_y^k D_\lambda^k e$. Thus,

$$(Y_k+\alpha D_y^k)(\Lambda_k+\alpha D_\lambda^k)e\ge\beta_1\mu e+\alpha(1-\beta_1)\mu e+\alpha^2 D_y^k D_\lambda^k e,\tag{4.18}$$

$$(Y_k+\alpha D_y^k)(\Lambda_k+\alpha D_\lambda^k)e\le\beta_2\mu e-\alpha(\beta_2-1)\mu e+\alpha^2 D_y^k D_\lambda^k e.\tag{4.19}$$

Since (d_y^k, d_λ^k) is bounded and $0 < \beta_1 < 1 < \beta_2$, there exists a constant $\bar{\alpha}_3 \in (0, 1]$ such that (4.17) holds for all $\alpha \in [0, \bar{\alpha}_3]$ and all $k \geq 0$.

If $\max\{y^k + \alpha d_y^k, -c(x^k + \alpha d_x^k)\} = y^k + \alpha d_y^k$ for all $k \geq 0$ and all $\alpha \in [0, \bar{\alpha}_3]$, then the lemma follows from (4.17) directly. Now we suppose that, for some k and some constant $\bar{\alpha}_3^{'} \in (0, \bar{\alpha}_3]$, we have $y_i^k + \alpha d_{yi}^k < -c_i(x^k + \alpha d_x^k)$ for all $\alpha \in (0, \bar{\alpha}_3^{'}]$. We prove that there exists a constant $\tilde{\alpha}_3 \in (0, 1]$ not dependent on k such that, for all $\alpha \in [0, \tilde{\alpha}_3]$,

$$-(\lambda_i^k + \alpha d_{\lambda i}^k) c_i(x^k + \alpha d_x^k) \leq \beta_2 \mu. \tag{4.20}$$

For convenience of statement, we define $p_i(\alpha) = -(\lambda_i^k + \alpha d_\lambda^k) c_i(x^k + \alpha d_x^k)$. Then $p_i(0) = -c_i^k \lambda_i^k$. We show that there exists a positive constant $\bar{\epsilon}$ such that we have either $p_i(0) \leq \beta_2 \mu - \bar{\epsilon}$ or $p_i^{'}(0) \leq -\bar{\epsilon} < 0$. Then (4.20) follows from the continuity of function p_i and the boundedness of (d_x^k, d_λ^k).

By (3.36), we have $c^k + y^k \geq 0$ and $\lambda_k > 0$ for $k \geq 1$. Thus, $p_i(0) \leq y_i^k \lambda_i^k$. For any given small constant $\epsilon > 0$ satisfying $\beta_2 \mu - c\epsilon > \mu$ ($c > 1$ is a constant), if $c_i^k + y_i^k \geq \epsilon$, or $c_i^k + y_i^k < \epsilon$ and $y_i^k \lambda_i^k \leq \beta_2 \mu - \epsilon$, then $p_i(0) \leq \beta_2 \mu - \bar{\epsilon}$ for some constant $\bar{\epsilon} > 0$. Now suppose $c_i^k + y_i^k < \epsilon$ and $y_i^k \lambda_i^k > \beta_2 \mu - \epsilon$. Then, by Procedure 2.1 and Lemma 4.2, there exists a small positive constant $\epsilon^{'}$ dependent on ϵ such that $A_{ki}^\top d_x^k + d_{yi}^k \geq -\epsilon^{'}$. Thus, $p_i^{'}(0) = -\lambda_i^k A_{ki}^\top d_x^k - c_i^k d_{\lambda i}^k \leq \lambda_i^k d_{yi}^k + y_i^k d_{\lambda i}^k + \epsilon^{''}$ for some small positive constant $\epsilon^{''}$. By (3.22), we have $p_i^{'}(0) \leq \mu - y_i^k \lambda_i^k + \epsilon^{''} < \epsilon + \epsilon^{''} - (\beta_2 - 1)\mu < 0$ since $\beta_2 > 1$. ∎

Let $\hat{\alpha}_4 = \min\{\hat{\alpha}_2, \hat{\alpha}_3\}$, where $\hat{\alpha}_2$ and $\hat{\alpha}_3$ are defined as in Lemmas 4.3 and 4.4, respectively. Then $0 < \hat{\alpha}_4 \leq 1$. By Step 5 of Algorithm 3.4, $\alpha_k > \delta \hat{\alpha}_4$ for all k, which implies that our line search procedure is well defined.

LEMMA 4.5. *If $\rho_k = \hat{\rho}$ for all $k \geq \hat{k}$ and if $\{(x^k, y^k, \lambda^k)\}$ is an infinite sequence generated by* Algorithm 3.4*, then we have*

$$\lim_{k\to\infty} d_x^k = 0, \ \lim_{k\to\infty} d_y^k = 0, \tag{4.21}$$

$$\lim_{k\to\infty} \|c^{k+1} + y^{k+1}\| = 0, \tag{4.22}$$

$$\lim_{k\to\infty} Y_{k+1} \Lambda_{k+1} e = \mu e, \tag{4.23}$$

$$\lim_{k\to\infty} \|g^{k+1} + A_{k+1} \lambda^{k+1}\| = 0. \tag{4.24}$$

Proof. It follows from Lemma 4.2 that the sequence $\{\phi_\mu(x^k, y^k; \hat{\rho})\}$ is bounded. Combined with its monotonicity, the limit of $\{\phi_\mu(x^k, y^k; \hat{\rho})\}$ exists as $k \to \infty$. Since $\alpha_k > \delta \hat{\alpha}_4 > 0$ and $\pi_{\hat{\rho}}((x^k, y^k); (d_x^k, d_y^k)) \leq 0$ for all k, by taking the limit on the two sides of (3.34), we have $\lim_{k\to\infty} \pi_{\hat{\rho}}((x^k, y^k); (d_x^k, d_y^k)) = 0$, which implies that $\lim_{k\to\infty}(d_x^k, d_y^k) = 0$ by (3.29) and Lemma 4.2.

By (4.21) and (3.23), we have $A_k^\top \tilde{d}_x^k + \tilde{d}_y^k \to 0$ as $k \to \infty$. If $(\tilde{d}_x^k, \tilde{d}_y^k)$ satisfies $q_k(\tilde{d}_x^k, \tilde{d}_y^k) \leq \nu q_k(0, 0)$, then

$$\|c^k + y^k + A_k^\top \tilde{d}_x^k + \tilde{d}_y^k\| - \nu \|c^k + y^k\| \leq 0, \tag{4.25}$$

which implies that (4.22) holds. Otherwise, since $q_k(\tilde{d}_x^k, \tilde{d}_y^k) \leq q_k(0, 0)$, for $k \to \infty$ we have

$$0 \geq -\frac{1}{2\rho_k} \left(\tilde{d}_x^{k\top} B_k \tilde{d}_x^k + \tilde{d}_y^{k\top} S_k \tilde{d}_y^k \right) \geq \|c^k + y^k + A_k^\top \tilde{d}_x^k + \tilde{d}_y^k\| - \|c^k + y^k\| \to 0. \tag{4.26}$$

It follows that $(\tilde{d}_x^k, \tilde{d}_y^k) \to 0$ as $k \to \infty$. Thus, by Procedure 2.1, formulae (3.12)–(3.14), Lemma 4.2, and Assumption 4.1, we have $\lim_{k\to\infty} \|c^k + y^k\| = 0$. This proves (4.22) by (4.21).

It follows from (3.22) that $Y_k(\lambda^k + d_\lambda^k) = \mu e - \Lambda_k d_y^k$. Thus, by (4.21) and Lemma 4.2, $\lim_{k\to\infty} Y_{k+1}(\lambda^k + d_\lambda^k) = \lim_{k\to\infty} Y_k(\lambda^k + d_\lambda^k) = \mu e$. Then, by Step 6 of Algorithm 3.4, we have $\lambda^{k+1} = \lambda^k + d_\lambda^k$ for sufficiently large k; thus (4.23) holds. Moreover, for sufficiently large k, by (3.21), we have

$$g^k + A_k \lambda^{k+1} = -B_k d_x^k. \tag{4.27}$$

Thus, (4.24) follows immediately from Assumption 4.1 and (4.21). □

It follows from Lemmas 4.2 and 4.5 that the weighted Newton step will be accepted at last if $\{\rho_k\}_{k=0}^{\infty}$ is bounded, since (3.28) is satisfied after a finite number of iterations.

Now we consider the case of $\rho_k \to \infty$. For simplicity of statement, we give the following definitions.

DEFINITION 4.6.

(1) *$x^* \in \Re^n$ is called a singular stationary point of the problem* (1.1) *if $c(x^*) \le 0$ and $A_i(x^*)$, $i \in \mathcal{I}$, are linearly dependent, where $\mathcal{I} = \{i | c_i(x^*) = 0, i = 1, \ldots, m\}$;*

(2) *$x^* \in \Re^n$ is called an infeasible stationary point of the problem* (1.1)*, if x^* is an infeasible point of the problem* (1.1) *and $A(x^*)c(x^*)_+ = 0$, where $c(x^*)_+ = \max\{c(x^*), 0\}$.*

It is easy to see that both the singular stationary point and the infeasible stationary point have some first order stationary properties. Similar definitions are also used in [2, 20, 30]. A singular stationary point is also a Fritz–John point, where the linearly independent constraint qualification does not hold. An infeasible stationary point is also a stationary point for minimizing $\|c(x)_+\|$ because $A(x^*)c(x^*)_+ = 0$. Moreover, if all constraint functions are convex, then the infeasible stationary point is the "least infeasible solution" in ℓ_2 sense.

LEMMA 4.7. *If $\rho_k \to \infty$, then*

(i) *the sequence $\{y^k\}$ is bounded;*

(ii) *$\{y^k\}$ is not componentwise bounded away from zero.*

Proof. (i) By (3.34), we have $\phi_\mu(x^{k+1}, y^{k+1}; \rho_{k+1}) \le \phi_\mu(x^k, y^k; \rho_{k+1})$ for all $k \ge 0$. The boundedness of $\{x^k\}$ implies that there exists a constant $b_7 > 0$ such that $|f_k| < b_7$. Thus,

$$\begin{aligned} &\frac{1}{\rho_{k+1}}\phi_\mu(x^{k+1}, y^{k+1}; \rho_{k+1}) - \frac{1}{\rho_k}\phi_\mu(x^k, y^k; \rho_k) \\ &\quad \le \left(\frac{1}{\rho_k} - \frac{1}{\rho_{k+1}}\right)(-\psi_\mu(x^k, y^k)) \\ &\quad \le \left(\frac{1}{\rho_k} - \frac{1}{\rho_{k+1}}\right)(b_7 + \mu m \ln \|y^k\|). \end{aligned} \tag{4.28}$$

It follows from (4.28) that

$$\begin{aligned} &\frac{1}{\rho_{k+1}}\phi_\mu(x^{k+1}, y^{k+1}; \rho_{k+1}) \\ &\quad \le \frac{1}{\rho_0}\phi_\mu(x^0, y^0; \rho_0) + \left(\frac{1}{\rho_0} - \frac{1}{\rho_{k+1}}\right)\left(b_7 + \mu m \max_{0\le j\le k+1} \ln \|y^j\|\right). \end{aligned} \tag{4.29}$$

On the other hand, we have

$$\frac{1}{\rho_{k+1}}\phi_\mu(x^{k+1}, y^{k+1}; \rho_{k+1})$$

$$\geq -\frac{1}{\rho_{k+1}}\left(b_7+\mu m\max_{0\leq j\leq k+1}\ln\|y^j\|\right)+\|y^{k+1}\|-\|c^{k+1}\|. \tag{4.30}$$

Thus, by (4.29) and (4.30), there is a constant $b_8>0$ such that

$$b_8+\frac{\mu m}{\rho_0}\max_{0\leq j\leq k+1}\ln\|y^j\|\geq\|y^{k+1}\| \text{ for all } k\geq 0, \tag{4.31}$$

which implies that $\{y^k\}$ is bounded.

(ii) If $\{y^k\}$ is componentwise bounded away from zero, then, by (i) and (3.37), the sequence $\{\lambda^k\}$ is also bounded above and componentwise bounded away from zero. Thus, matrix S_k is uniformly bounded. Let $\mathcal{K}=\{k|\rho_k<\rho_{k+1}\}$. Then $\mathcal{K}$ is an infinite index set. It follows from Assumption 4.1 and Proposition 3.3 that there exists a positive constant $\hat{\rho}$ such that the weighted Newton step defined by (3.12) and (3.13) is accepted at iterate $k\in\mathcal{K}$ if $\rho_k>\hat{\rho}$. Thus, $\Delta_k=\|c^k+y^k\|$ by Proposition 3.2 and (3.32). Moreover, there exists a constant $b_9>0$ such that, for sufficiently large $k\in\mathcal{K}$,

$$\|\tilde{d}_x^k\|\leq b_9\|c^k+y^k\|,\ \ \|\tilde{d}_y^k\|\leq b_9\|c^k+y^k\|,\ \text{and}\ \|S_k\tilde{d}_y^k\|\leq b_9\|c^k+y^k\|. \tag{4.32}$$

Hence, by the boundedness of $\|c^k+y^k\|$ and Assumption 4.1(3), there exists a constant $b_{10}>0$ such that, for all sufficiently large $k\in\mathcal{K}$,

$$\begin{aligned}&\pi_{\rho_k}((x^k,y^k);(\tilde{d}_x^k,\tilde{d}_y^k))+\frac{1}{2}(\tilde{d}_x^k)^\top B_k\tilde{d}_x^k+\frac{1}{2}(\tilde{d}_y^k)^\top S_k\tilde{d}_y^k\\ &\qquad\leq b_{10}\|c^k+y^k\|-\rho_k\|c^k+y^k\|,\end{aligned} \tag{4.33}$$

which, by (3.17), implies that we have (3.29) for all iterates $k\in\mathcal{K}$ with $\rho_k\geq\max\{\hat{\rho},b_{10}\}$. This contradicts the fact that $\mathcal{K}$ is an infinite set. □

By Lemma 4.7 and (3.37), λ^k is componentwise bounded away from zero. Thus, both Λ_k^{-1} and S_k^{-1} are bounded above.

LEMMA 4.8. *Let $\mathcal{K}=\{k\mid\rho_k<\rho_{k+1}\}$. If $\rho_k\to\infty$ and if $\tilde{\mathcal{K}}$ is any subset of $\mathcal{K}$ such that $(x^k,y^k)\to(x^*,y^*)$ as $k\in\tilde{\mathcal{K}}$ and $k\to\infty$, then*

$$\det[(A_{\mathcal{J}}^*)^\top A_{\mathcal{J}}^*]=0, \tag{4.34}$$

where $\mathcal{J}=\{i|y_i^*=0,i=1,\ldots,m\}$.

Proof. We prove this lemma by contradiction. Suppose that there is a set $\tilde{\mathcal{K}}\subseteq\mathcal{K}$ such that, as $k\in\tilde{\mathcal{K}}$ and $k\to\infty$, $(x^k,y^k)\to(x^*,y^*)$ and $A_i(x^*)$, $i\in\mathcal{J}$, are linearly independent. Then, by Assumption 4.1 and (3.37), there exists a constant $b_{11}>0$ such that $A(x^*)^\top(B^*)^{-1}A(x^*)+G^*\succeq b_{11}I$, where I is the identity matrix, and for simplicity we assume that $B_k\to B^*$ and $S_k^{-1}\to G^*$ as $k\in\tilde{\mathcal{K}}$ and $k\to\infty$. Thus, by the continuity of $A(x)$, there exists a constant $b_{12}>0$ such that

$$\|(A_k^\top B_k^{-1}A_k+S_k^{-1})^{-1}\|\leq b_{12} \tag{4.35}$$

for all sufficiently large $k\in\tilde{\mathcal{K}}$. It follows from (3.27) that the weighted Newton step defined by (3.12) and (3.13) is accepted. Hence, we have the same results as (4.32) and (4.33), which result in a contradiction to the definition of $\mathcal{K}$. □

LEMMA 4.9. *If $\rho_k\to\infty$, then there must be a limit point which is either a singular stationary point or an infeasible stationary point.*

In order to prove Lemma 4.9, we need to prove three other lemmas first.

LEMMA 4.10. *If $\{(\tilde{d}_x^k,\tilde{d}_y^k)\}$ is a sequence such that $q_k(\tilde{d}_x^k,\tilde{d}_y^k)\leq\omega q_k(0,0)$ for $0<\omega\leq 1$, then $\|\tilde{d}_x^k\|/\sqrt{\rho_k}$ and $\|Y_k^{-1}\tilde{d}_y^k\|/\sqrt{\rho_k}$ are uniformly bounded above.*

Proof. Let $(\hat{d}_x^k, \hat{d}_y^k) = (\tilde{d}_x^k/\sqrt{\rho_k}, Y_k^{-1}\tilde{d}_y^k/\sqrt{\rho_k})$. Then by $q_k(\tilde{d}_x^k, \tilde{d}_y^k) \le \omega q_k(0,0)$, we have

$$\frac{1}{2}\hat{d}_x^{k\top} B_k \hat{d}_x^k + \frac{1}{2}\hat{d}_y^{k\top} Y_k \Lambda_k \hat{d}_y^k + \|c^k + y^k + \sqrt{\rho_k} A_k^\top \hat{d}_x^k + \sqrt{\rho_k} Y_k \hat{d}_y^k\| \le \omega \|c^k + y^k\|. \tag{4.36}$$

The boundedness of $(\hat{d}_x^k, \hat{d}_y^k)$ follows from the uniform lower boundedness of the quadratic terms by Assumption 4.1 and (3.37). ∎

LEMMA 4.11. *Suppose that $(\tilde{d}_x^k, \tilde{d}_y^k)$ is an approximate solution of program* (3.11) *such that $q_k(\tilde{d}_x^k, \tilde{d}_y^k) \le q_k(\alpha_k^C(\tilde{d}_x^k)^C, \alpha_k^C(\tilde{d}_y^k)^C)$, where $((\tilde{d}_x^k)^C, (\tilde{d}_y^k)^C)$ is the weighted steepest descent step (see Procedure* 2.1 *and* (3.10)*), $\alpha_k^C \in [0,1]$ minimizes the function $q_k(\alpha(\tilde{d}_x^k)^C, \alpha(\tilde{d}_y^k)^C)$. Then there exist positive constants $\tilde{\rho}$ and $\tilde{\omega}$ such that, for $\rho_k \ge \tilde{\rho}$, we have*

$$q_k(\tilde{d}_x^k, \tilde{d}_y^k) - q_k(0,0) \le -\tilde{\omega}\rho_k \left\| \begin{pmatrix} A_k \\ Y_k \end{pmatrix} (c^k + y^k) \right\|^2. \tag{4.37}$$

Proof. By (3.10), the value of η in Proposition 2.2 is

$$\eta_k = \|(A_k^\top B_k^{-1} A_k + S_k^{-1})^{1/2}(c^k + y^k)\|^2 / \|(A_k^\top B_k^{-1} A_k + S_k^{-1})(c^k + y^k)\|^2. \tag{4.38}$$

It follows from Assumption 4.1 and (3.37) that

$$\begin{aligned} &(c^k + y^k)^\top \begin{pmatrix} A_k \\ I \end{pmatrix}^\top \begin{pmatrix} B_k^{-1} & \\ & Y_k \Lambda_k^{-1} \end{pmatrix} \begin{pmatrix} A_k \\ I \end{pmatrix} (c^k + y^k) \\ &\ge \omega_1 \left\| \begin{pmatrix} A_k \\ Y_k \end{pmatrix} (c^k + y^k) \right\|^2, \end{aligned} \tag{4.39}$$

where $\omega_1 = \min\{\nu_2^{-1}, \beta_2^{-1}\mu^{-1}\}$. By Assumption 4.1 and Lemma 4.7(i), there is a constant $\omega_2 > 0$ such that $\|c^k + y^k\| \le \omega_2$. Let $\tilde{\rho}_1 = 2\omega_2$. Then, for $\rho_k \ge \tilde{\rho}_1$, we have $1 - (\rho_k/\|c^k + y^k\|) \le -\rho_k/(2\omega_2)$. If $\eta_k \ge 1$, by Proposition 2.2, we have (4.37) if $\tilde{\omega} \le \omega_1/(4\omega_2)$.

Now we suppose that $\eta_k < 1$. By Assumption 4.1, Lemma 4.7, and (3.37), there is a constant $\omega_3 > 0$ such that $\|(A_k^\top B_k^{-1} A_k + S_k^{-1})^{1/2}\|^2 \|c^k + y^k\| \le \omega_3$ for all k. Since $\eta_k \ge 1/\|(A_k^\top B_k^{-1} A_k + S_k^{-1})^{1/2}\|^2$ by (4.38), if we select $\tilde{\rho}_2 = 2\omega_3$, then, for $\rho_k \ge \tilde{\rho}_2$, we have $1 - (\rho_k \eta_k/\|c^k + y^k\|) \le -\rho_k/(2\omega_3)$. Thus, for $\rho_k \ge \tilde{\rho}_2$, it follows from Proposition 2.2 and (4.39) that (4.37) holds if $\tilde{\omega} \le \omega_1/(4\omega_3)$.

Let $\tilde{\omega} = \min\{\omega_1/(4\omega_2), \omega_1/(4\omega_3)\}$. Then the result follows by taking the constant $\tilde{\rho} = \max\{\tilde{\rho}_1, \tilde{\rho}_2\}$. ∎

LEMMA 4.12. *Let $\mathcal{K} = \{k \mid \rho_k < \rho_{k+1}\}$. If $\rho_k \to \infty$, then*

$$\left\| \begin{pmatrix} A_k \\ Y_k \end{pmatrix} (c^k + y^k) \right\| \to 0 \tag{4.40}$$

as $k \in \mathcal{K}$ and $k \to \infty$.

Proof. Suppose that (4.40) does not hold. Then there exist an infinite subset $\tilde{\mathcal{K}} \subseteq \mathcal{K}$, positive constants τ_1 and τ_2 such that

$$\left\| \begin{pmatrix} A_k \\ Y_k \end{pmatrix} (c^k + y^k) \right\| \ge \tau_1, \tag{4.41}$$

and $\|c^k + y^k\| \ge \tau_2$ for all $k \in \tilde{\mathcal{K}}$.

The approximate solution $(\tilde{d}_x^k, \tilde{d}_y^k)$ is generated such that either $q_k(\tilde{d}_x^k, \tilde{d}_y^k) \le \nu q_k(0,0)$ or $q_k(\tilde{d}_x^k, \tilde{d}_y^k) \le q_k(\alpha_k^C(\tilde{d}_x^k)^C, \alpha_k^C(\tilde{d}_y^k)^C)$ (which implies that $q_k(\tilde{d}_x^k, \tilde{d}_y^k) \le q_k(0,0)$). Then, by Lemma 4.10, there is a constant $\tau_3 > 0$ such that $\|\tilde{d}_x^k\| \le \tau_3\sqrt{\rho_k}$, $\|Y_k^{-1}\tilde{d}_y^k\| \le \tau_3\sqrt{\rho_k}$.

If $q_k(\tilde{d}_x^k, \tilde{d}_y^k) \le \nu q_k(0,0)$ for all $k \in \tilde{\mathcal{K}}$, then there exists a constant $\tau_4 > 0$ such that

$$\begin{aligned}\pi_{\rho_k}\,((x^k,y^k);(\tilde{d}_x^k,\tilde{d}_y^k)) &+ \frac{1}{2}(\tilde{d}_x^k)^\top B_k\tilde{d}_x^k + \frac{1}{2}(\tilde{d}_y^k)^\top S_k\tilde{d}_y^k\\ &\le (g^k)^\top\tilde{d}_x^k - \mu e^\top Y_k^{-1}\tilde{d}_y^k - (1-\nu)\rho_k\|c^k + y^k\| \\ &\le \tau_4\sqrt{\rho_k} - (1-\nu)\tau_2\rho_k.\end{aligned} \tag{4.42}$$

Thus, by (3.17), we can select a positive constant $\hat{\rho}$ such that (3.29) holds for all $\rho_k \ge \hat{\rho}$. This contradicts the definition of $\mathcal{K}$. Hence, there must exist an infinite subset $\hat{\mathcal{K}}$ of $\tilde{\mathcal{K}}$ such that $q_k(\tilde{d}_x^k, \tilde{d}_y^k) \le q_k(\alpha_k^C(\tilde{d}_x^k)^C, \alpha_k^C(\tilde{d}_y^k)^C)$ for all $k \in \hat{\mathcal{K}}$. It follows from Lemma 4.11 that (4.37) holds for all $k \in \hat{\mathcal{K}}$. Then, by (4.41), there is a positive constant b_{13} such that, for all $k \in \hat{\mathcal{K}}$,

$$q_k(\tilde{d}_x^k, \tilde{d}_y^k) - q_k(0,0) \le -b_{13}\tau_1^2\rho_k. \tag{4.43}$$

Thus, we have

$$\begin{aligned}\pi_{\rho_k}\,((x^k,y^k);(\tilde{d}_x^k,\tilde{d}_y^k)) &+ \frac{1}{2}(\tilde{d}_x^k)^\top B_k\tilde{d}_x^k + \frac{1}{2}(\tilde{d}_y^k)^\top S_k\tilde{d}_y^k\\ &\le (g^k)^\top\tilde{d}_x^k - \mu e^\top Y_k^{-1}\tilde{d}_y^k - b_{13}\tau_1^2\rho_k \\ &\le \tau_4\sqrt{\rho_k} - b_{13}\tau_1^2\rho_k\end{aligned} \tag{4.44}$$

for all sufficiently large $k \in \bar{\mathcal{K}}$, which implies a contradiction to the definition of $\mathcal{K}$. □

Proof of Lemma 4.9. Since (x^k, y^k) is bounded, without loss of generality, we suppose that $(A_k, c^k, x^k, y^k, Y_k) \to (A^*, c^*, x^*, y^*, Y^*)$ as $k \in \mathcal{K}$ and $k \to \infty$, where $\mathcal{K}$ is defined as in Lemma 4.12, $A^* = A(x^*)$, and $c^* = c(x^*)$. If the limit point (x^*, y^*) is such that $c^* + y^* = 0$, i.e., $c_i^* = 0$ if and only if $y_i^* = 0$, then this limit point is a singular stationary point by Lemma 4.8 since $\mathcal{I} = \mathcal{J}$, where $\mathcal{I}$ and $\mathcal{J}$ are defined as in Definition 4.6 and Lemma 4.8, respectively. Now we consider the case of $\|c^* + y^*\| \ne 0$. By Lemma 4.12,

$$\begin{pmatrix} A^* \\ Y^* \end{pmatrix}(c^* + y^*) = 0, \tag{4.45}$$

and so for any i,

$$y_i^* > 0 \;\Rightarrow\; c_i^* + y_i^* = 0 \;\Rightarrow\; c_i^* < 0. \tag{4.46}$$

Since $c^k + y^k \ge 0$ and $y^k \ge 0$ for all $k \ge 1$ by the algorithm, for each i such that $c_i^* + y_i^* \ne 0$, one has $y_i^* = 0$ by (4.45), and hence $c_i^* > 0$, implying that x^* is infeasible. Then $c^* + y^* = c_+^* = \max\{c^*, 0\}$. It follows from (4.45) that $A^* c_+^* = 0$. Therefore, x^* is an infeasible stationary point. The proof is finished. □

Now we can state our global convergence theorem on Algorithm 3.4.

THEOREM 4.13. *Suppose that $\{(x^k, y^k, \lambda^k)\}$ is an infinite sequence generated by applying Algorithm* 3.4 *to the barrier problem* (3.1)–(3.2)*, and suppose that Assumption* 4.1 *holds. The penalty parameter sequence $\{\rho_k\}$ is automatically updated and monotonically nondecreasing.*

(i) *If $\{\rho_k\}$ is bounded, then any cluster point of $\{(x^k, y^k, \lambda^k)\}$ is a KKT point of the barrier problem* (3.1)–(3.2). *In this case, $\{y^k\}$ is componentwise bounded away from zero, $\{x^k\}$ is asymptotically strictly feasible for the constraints* (1.1), *and $g^k + A_k\lambda^k \to 0$.*

(ii) *If $\rho_k \to \infty$, then $\{y^k\}$ is not componentwise bounded away from zero, and there is at least one cluster point of $\{(x^k, y^k, \lambda^k)\}$ which is either a singular stationary point or an infeasible stationary point. In the latter case, if (x^k, y^k) is asymptotically feasible for constraints* (3.2), *then $\{x^k\}$ is asymptotically feasible for and close to the boundary of constraints* (1.1). *At the limit the gradients of active constraints of* (1.1) *are linearly dependent. If (x^k, y^k) is not asymptotically feasible for constraints* (3.2), *then at the limit point x^* we have $A^* c_+^* = 0$.*

Proof. Part (i) follows from Lemma 4.5. Part (ii) can be derived directly by Lemma 4.9. ∎

5. The overall interior-point algorithm and its convergence. We denote by $\mathcal{F}$ the class of continuous functions $\theta : \Re_{++} \to \Re_{++}$ satisfying $\lim_{\mu\to 0} \theta(\mu) = 0$. Now we present our algorithm for nonlinearly constrained optimization (1.1).

ALGORITHM 5.1 (the line search–based interior-point algorithm for (1.1)).

Step 1. *Given initial point $(x^0, y^0, \lambda^0) \in \Re^n \times \Re^m_{++} \times \Re^m_{++}$, initial barrier parameter $\mu_0 > 0$, $\tau \in (0, 1)$, tolerance $\epsilon > 0$, and function $\theta \in \mathcal{F}$. Let $j := 0$.*

Step 2. *For the given barrier parameter μ_j, we apply Algorithm* 3.4 *to the barrier problem* (3.1)–(3.2). *If the iterate $(x^{k_j}, y^{k_j}, \lambda^{k_j})$ satisfies*

$$\|F_{\mu_j}(x^{k_j}, y^{k_j}, \lambda^{k_j})\| < \theta(\mu_j), \tag{5.1}$$

then let

$$(x^{j+1}, y^{j+1}, \lambda^{j+1}) = (x^{k_j}, y^{k_j}, \lambda^{k_j}) \tag{5.2}$$

and $\rho_{j+1} = \rho_{k_j}$, and go to Step 3*; if one of conditions* (ii) *and* (iii) *of Algorithm* 3.4 *holds, stop.*

Step 3. *If $\mu_j < \epsilon$ stop; otherwise, let $\mu_{j+1} = \tau\mu_j$, $j := j + 1$ and go to Step* 2.

Now we consider the convergence of Algorithm 5.1. The result closely depends on how Algorithm 3.4 behaves for each μ_j. For $\theta(\mu_j) > 0$, if condition (5.1) is satisfied, then Algorithm 5.1 will proceed with μ_{j+1}. The global convergence results of the algorithm are as follows.

THEOREM 5.2. *Suppose that $\theta \in \mathcal{F}$ and $\{(x^j, y^j, \lambda^j)\}$ is a sequence generated by Algorithm* 5.1. *If Assumption* 4.1 *holds for each barrier problem, and if $\{(x^k, y^k, \lambda^k)\}$ is a sequence generated by Algorithm* 3.4, *then, for sufficiently small ϵ, Algorithm* 5.1 *may terminate in finitely many steps in one of the following two cases:*

(i) *For some μ_j, Algorithm* 5.1 *terminates at Step* 2. *If the termination point is an approximately feasible point, then it is an approximately singular stationary point. Otherwise, it is an approximately infeasible stationary point.*

(ii) *For each μ_j, Algorithm* 3.4 *terminates at* (5.1). *Then Algorithm* 5.1 *terminates at Step* 3, *in which case an approximate KKT point of the original problem* (1.1) *is obtained.*

Proof. The results follow immediately from Theorem 4.13 and Algorithm 5.1. ∎

6. Numerical experiment.

6.1. Formulae used in Procedure 2.1. We present an implementation of Procedure 2.1 in this subsection.

Suppose that the full Q-weighted Newton step is not accepted. Then we compute the weighted Cauchy step $\tilde{d}_z^C$ and try to get an approximate solution $\tilde{d}_z$ to (2.10) along the Q-weighted Newton step, or the so-called dog-leg step, so that (2.13) holds and $q(\tilde{d}_z)$ has as much reduction as possible. If this is impossible, then we do a line search along the Q-weighted steepest descent step and take the approximate solution $\tilde{d}_z$ to be either the truncated Q-weighted Newton step or the truncated Q-weighted steepest descent step so that $q(\tilde{d}_z)$ has more reduction. Thus, (2.13) holds. The details are as follows.

We first compute the optimal steplength along the Q-weighted Newton step to derive as much reduction as possible in this direction. Thus, we solve the single-variable minimizing problem

$$\text{minimize}_{\alpha\in[0,1]}\ \ \hat{q}(\alpha) = \frac{1}{2}\alpha^2 \tilde{d}_z^{N\top} Q \tilde{d}_z^N + \rho\|r + \alpha R^\top \tilde{d}_z^N\|. \tag{6.1}$$

By direct computation, we have the solution

$$\tilde{\alpha}_1 = \min\left\{\frac{\rho\|r\|}{r^\top (R^\top Q^{-1} R)^{-1} r}, 1\right\}. \tag{6.2}$$

Set $d_z^1 = \tilde{\alpha}_1 \tilde{d}_z^N$. Then we have $\hat{q}(\tilde{\alpha}_1) \le \hat{q}(0)$. It is more convenient in the implementation to compute a dog-leg step in the line segment spanned by the Q-weighted Newton step $\tilde{d}_z^N$ and the following scaled Cauchy step (where η is defined as in Proposition 2.2):

$$\tilde{d}_z^C = -\min\{\eta, 1\} Q^{-1} R r. \tag{6.3}$$

It is apparent that this scaling on $\tilde{d}_z^C$ will not result in any change in our theoretical results. If $\eta \le 1$, then $\tilde{d}_z^C$ is the so-called Cauchy point in minimizing $\|r + R^\top d\|^2$ with starting point $d = 0$. Let $d_z(\alpha) = \alpha \tilde{d}_z^N + (1-\alpha)\tilde{d}_z^C$. Then we calculate $\tilde{\alpha}_2$ by

$$\text{minimize}_{\alpha\in[0,1]} \tilde{q}(\alpha) = \frac{1}{2} d_z(\alpha)^\top Q d_z(\alpha) + \rho\|r + R^\top d_z(\alpha)\|. \tag{6.4}$$

By setting $\tilde{q}'(\alpha) = 0$, we have

$$\alpha_2^* = \frac{\rho\|r + R^\top \tilde{d}_z^C\| - (\tilde{d}_z^N - \tilde{d}_z^C)^\top Q \tilde{d}_z^C}{(\tilde{d}_z^N - \tilde{d}_z^C)^\top Q (\tilde{d}_z^N - \tilde{d}_z^C)}. \tag{6.5}$$

If $\alpha_2^* \le 0$, then $\tilde{\alpha}_2 = 0$; else if $\alpha_2^* \ge 1$, then $\tilde{\alpha}_2 = 1$; else we have $\tilde{\alpha}_2 = \alpha_2^*$. If $\min\{\hat{q}(\tilde{\alpha}_1), \tilde{q}(\tilde{\alpha}_2)\} \le \nu q(0)$ (where ν is defined as in Procedure 2.1), we define $d_z^2 = d_z(\tilde{\alpha}_2)$, else we set $d_z^2 = \tilde{\alpha}_3 \tilde{d}_z^C$, where $\tilde{\alpha}_3 \in (0, 1]$ minimizes the function

$$\bar{q}(\alpha) = \frac{1}{2}\alpha^2 (\tilde{d}_z^C)^\top Q \tilde{d}_z^C + \rho\|r + \alpha R^\top \tilde{d}_z^C\|. \tag{6.6}$$

We select the approximate solution $\tilde{d}_z$ from d_z^1 or d_z^2, whichever gives a lower value of $q(\tilde{d}_z)$.

The process for solving (2.10) approximately is summarized into the following algorithm.

ALGORITHM 6.1 (the algorithm for solving problem (2.10) approximately).

Step 1. *Compute the Newton step* $\tilde{d}_z^N$ *by* (2.11). *If* $q(\tilde{d}_z^N) \le \nu q(0)$, *then* $\tilde{d}_z = \tilde{d}_z^N$. *Stop.*

Step 2. *Compute the steepest descent step* $\tilde{d}_z^C$ *by* (2.13).

Step 3. *Calculate* $d_z^1 = \tilde{\alpha}_1 \tilde{d}_z^N$ *by* (6.2) *and* $d_z^2 = \tilde{\alpha}_2 \tilde{d}_z^N + (1-\tilde{\alpha}_2)\tilde{d}_z^C$ *by* (6.4). *If* $\min\{\hat{q}(\tilde{\alpha}_1), \tilde{q}(\tilde{\alpha}_2)\} \le \nu q(0)$, *then go to Step* 5.

Step 4. *Calculate* $d_z^2 = \tilde{\alpha}_3 \tilde{d}_z^C$ *by* (6.6). *If* $\hat{q}(\tilde{\alpha}_1) \le \bar{q}(\tilde{\alpha}_3)$, *we have the approximate solution* $\tilde{d}_z = d_z^1$; *else we select* $\tilde{d}_z = d_z^2$. *Stop.*

Step 5. *If* $\hat{q}(\tilde{\alpha}_1) \le \tilde{q}(\tilde{\alpha}_2)$, *then* $\tilde{d}_z = d_z^1$; *else we have* $\tilde{d}_z = d_z^2$. *Stop.*

6.2. Numerical results. The algorithm is programmed in MATLAB 6.1 and is run on a personal computer under Windows 98. In order to obtain rapid convergence, it is also necessary to carefully control the rate at which the barrier parameter μ and the tolerance $\theta(\mu)$ are decreased. This question has been studied in [6, 11, 29].

It is restrictive to require that (3.37) holds for given β_1 and β_2 for all iterates of Algorithm 3.4 in practice. In our implementation, we update the dual iterate flexibly by selecting the maximal $\gamma_k \in [0,1]$ such that

$$\min\{Y_{k+1}\Lambda_k e, \bar{\beta}_1 \mu e\} \le Y_{k+1}\Lambda_{k+1} e \le \max\{Y_{k+1}\Lambda_k e, \bar{\beta}_2 \mu e\}, \tag{6.7}$$

where $0 < \bar{\beta}_1 < 1 < \bar{\beta}_2$, $\Lambda_{k+1} = \operatorname{diag}(\lambda^{k+1})$, and $\lambda^{k+1} = \lambda^k + \gamma_k d_\lambda^k$. If $\{\rho_k\}_{k=0}^\infty$ is bounded, then, by Lemma 4.2 and (6.7), there exist β_1 and β_2 such that (3.37) holds for all iterates. In the case of $\rho_k \to \infty$, suppose that Algorithm 3.4 is terminated within a given number of iterations (for example, 300 iterations). Then, by the fact that $y^{k+1} \ge 0.005 y^k$ and (6.7), $Y_{k+1}\Lambda_{k+1} e \ge \min\{0.005 Y_k \Lambda_k e, \bar{\beta}_1 \mu e\}$. Thus, $Y_k \Lambda_k e \ge \beta_1 \mu e$ if we select $\beta_1 = 0.005^{300} \min\{\mu^{-1} Y_0 \Lambda_0 e, 200 \bar{\beta}_1 e\}$. If $y_i^k \lambda_i^k \to \infty$ as $k \to \infty$ for some i, then, by (6.7), $\lambda_i^k \le \lambda_i^{k-1}$ and $\lambda_i^k \to \infty$ as $k \to \infty$ since $\{y^k\}$ is bounded. This is a contradiction. Thus, there exist a constant $\beta_2 > 0$ and an infinite index set $\mathcal{K}$ such that $Y_k \Lambda_k e \le \beta_2 \mu e$ for $k \in \mathcal{K}$. Hence, we have (3.37) for all $k \in \mathcal{K}$.

We select the initial parameters $\mu_0 = 0.01$, $\bar{\beta}_1 = 0.01$, $\bar{\beta}_2 = 10$, $\sigma_0 = 0.1$, $\delta = 0.8$, and the initial matrix B_0 to be the $n \times n$ identity matrix. The scalar in Algorithm 6.1 is $\nu = 0.98$. The choice of the initial penalty parameter ρ_0 is scale dependent and $\rho_0 = 1$ is chosen for our experiment. Simply, we select $\theta(\mu) = \mu$, $\tau = 0.01$, $\epsilon = 10^{-6}$. For conditions (ii) and (iii) of Step 7 of Algorithm 3.4, we select $\epsilon_2 = \epsilon$ and $\epsilon_3 = \epsilon^2$.

The approximate Lagrangian Hessian B_{k+1} is computed by the damped BFGS update formula

$$B_{k+1} = B_k - \frac{B_k s^k (s^k)^\top B_k}{(s^k)^\top B_k s^k} + \frac{w^k (w^k)^\top}{(s^k)^\top w^k}, \tag{6.8}$$

where

$$w^k = \begin{cases} \hat{w}^k & \text{if } (\hat{w}^k)^\top s^k \ge 0.2 (s^k)^\top B_k s^k, \\ \theta_k \hat{w}^k + (1-\theta_k) B_k s^k & \text{otherwise,} \end{cases} \tag{6.9}$$

and $\hat{w}^k = g^{k+1} - g^k + (A_{k+1} - A_k)\lambda^{k+1}$, $s^k = x^{k+1} - x^k$, $\theta_k = 0.8 (s^k)^\top B_k s^k / ((s^k)^\top B_k s^k - (s^k)^\top \hat{w}^k)$. For all test problems, we select the initial slack and dual variables as

$$y^0 = e, \quad \lambda^0 = e \tag{6.10}$$

if not specified.

TABLE 1. *Numerical results by Algorithm* 3.4 *when* $\mu = 0.01$.

IT	x_1	x_2	x_3	RC_1	RC_2	ρ	$\tilde{d}_x$
0	-4	1	1	14	-7	1	full-Newton
1	-3.6590	12.3880	0.0050	0	-5.6640	2	dog-leg
2	-2.2786	4.1919	0.0040	0	-4.2826	4	full-Newton
3	-1.3633	0.8586	0.0030	0	-3.3663	4	full-Newton
4	-1.0500	0.1025	0.0026	0	-3.0525	8	dog-leg
5	-0.8756	0.0005	0.0019	-0.2339	-2.8775	8	dog-leg
6	-0.4536	0.0015	0.0000	-0.7957	-2.4537	8	dog-leg
7	0.4972	0.0430e-03	0.5770e-03	-0.7528	-1.5033	8	dog-leg
8	1.4035	0.9697	0.0009	0	-0.5975	8	full-Newton
9	2.0008	3.0031	0.0008	0	-0.9324e-09	8	full-Newton
10	2.0017	3.0067	0.0017	0	0	8	

TABLE 2. *Numerical results by the ordinary approach with* y^{k+1} *generated by* (3.36) *when* $\mu = 0.01$.

IT	x_1	x_2	x_3	RC_1	RC_2	ρ
0	-4	1	1	14	-7	1
1	-3.6590	12.3880	0.0050	0	-5.6640	2
2	-1.9746	2.8990	0.0028	0	-3.9774	5.2958
3	-1.2442	0.5480	0.0018	0	-3.2460	11.9755
4	-1.0251	0.0508	0.0007	0	-3.0258	101.7079
5	-1.0004	0.8606e-03	0.1721e-03	0	-3.0006	4.4576e+03
6	-1.0000	0.0449e-04	0.1219e-04	0	-3.0000	1.1483e+06
7	-1.0000	0.0224e-06	0.1183e-06	0	-3.0000	7.7089e+08
8	-1.0000	0.1122e-09	0.5969e-09	0	-3.0000	9.1419e+12
9	-1.0000	0.0561e-11	0.2984e-11	0	-3.0000	3.0875e+17

First, we apply our algorithm to three simple examples. The first one is the example presented by Wächter and Biegler and further discussed by Byrd, Marazzi, and Nocedal [7, 26]:

$$\text{Minimize} \quad x_1 \tag{6.11}$$

$$\text{(TP1)} \quad \text{subject to } x_1^2 - x_2 - 1 = 0, \tag{6.12}$$

$$x_1 - x_3 - 2 = 0, \tag{6.13}$$

$$x_2 \geq 0, \;\; x_3 \geq 0. \tag{6.14}$$

Note that the initial point $(x_1^0, x_2^0, x_3^0) = (-4, 1, 1)$ satisfies the conditions of Theorem 1 of [26]. There is a unique stationary point for this problem, which is the global minimizer. Moreover, this problem is well-posed, since at the solution the second order sufficient optimality condition, strict complementarity, and nondegeneracy hold. However, it is proved by [26] that many existing interior-point methods using line search (let us call them the "ordinary" interior-point methods for convenience) fail to converge to the stationary point.

Algorithm 5.1 terminates at the approximate KKT point $(2, 3, 0)$ with the Lagrangian multiplier $(0, 1)$ in 16 iterations. The residuals, respectively, are $\|g^k + A_k\lambda^k\| = 6.3283e\text{-}14$, $\|Y_k\Lambda_k e - \mu_k e\| = 2.0000e\text{-}08$, and $\|c^k + y^k\| = 0.8232e\text{-}17$. The value of the penalty parameter is $\hat{\rho} = 8$. In order to see the performance clearly, we give the numerical results of Algorithm 3.4 when $\mu = 0.01$, which is listed in Table 1, where RC_1 and RC_2 are residual values of constraints, (6.12) and (6.13), respectively. The last column in Table 1 shows the performance of Algorithm 6.1, where

TABLE 3. *Numerical results by the ordinary approach with* $y^{k+1} = y^k + \alpha_k d_y^k$ *when* $\mu = 0.01$.

IT	x_1	x_2	x_3	RC_1	RC_2	ρ
0	-4	1	1	14	-7	1
1	-3.6590	0.9438	0.0050	11.4442	-5.6640	2
2	-3.4809	0.0047	0.0029	11.1118	-5.4838	11.9086
3	-3.4789	0.0236e-03	0.3727e-03	11.1028	-5.4793	5.4425e+03
4	-3.4788	0.0118e-05	0.8007e-05	11.1017	-5.4788	3.8388e+05
5	-3.4787	0.0059e-07	0.4240e-07	11.1017	-5.4787	8.9516e+08
6	-3.4787	0.0029e-09	0.2121e-09	11.1017	-5.4787	3.3359e+13

"full-Newton" means that the approximate solution to (3.11) is the full weighted Newton step, and "dog-leg" represents the dog-leg step. In order to observe how the ordinary interior-point approach using (3.24)–(3.26) behaves, we also solve this example by solving (3.24)–(3.26) with y^{k+1} generated by (3.36) and $y^{k+1} = y^k + \alpha_k d_y^k$, respectively; the results are presented in Tables 2 and 3.

It is easy to note from Table 1 that Algorithm 3.4 terminates at the approximate feasible point when $\mu = 0.01$. The approximate feasibility will be further improved when μ is decreased in Algorithm 5.1. However, the results in Tables 2 and 3 show that the ordinary interior-point approach using (3.24)–(3.26) terminates at the infeasible points as $\mu = 0.01$. The infeasibility cannot be improved by decreasing μ since x_2 and x_3 are close to the boundary of the feasible region.

The last column of Table 1 shows that the weighted Newton steps are accepted as the iterates are nearly feasible, which is important for the algorithm to have rapid convergence.

Our second test example is taken from [3], which minimizes any objective function on an obviously infeasible set defined by the constraints:

$$\text{(TP2)} \quad x^2 + 1 \le 0, \;\; x \le 0. \tag{6.15}$$

We select to minimize x as the objective. The initial point is $x^0 = 4$. For $\mu = 0.01$, Algorithm 3.4 terminates at the point $x^* =$ -6.0363e-07, and correspondingly the slack variables $y_1^* = 6.3712e$-13 and $y_2^* = 6.0363e$-07 after 38 iterations. It is easy to see that x^* is close to a point by which the norm $\|c(x)_+\|$ is minimized. Algorithm 6.1 takes four full weighted Newton steps at first and then uses the truncated weighted Newton steps in 34 later iterations. The value of the penalty parameter is $\hat{\rho} = 1.2767e$+10.

The third simple test problem is a standard one taken from [17, Problem 13]:

$$\text{Minimize} \;\; (x_1 - 2)^2 + x_2^2 \tag{6.16}$$

$$\text{(TP3)} \quad \text{subject to } (1 - x_1)^3 - x_2 \ge 0, \tag{6.17}$$

$$x_1 \ge 0, \;\; x_2 \ge 0. \tag{6.18}$$

The standard initial point $(-2, -2)$ is an infeasible point. The optimal solution $(1, 0)$ is not a KKT point but is a singular stationary point, at which the gradients of active constraints are linearly dependent. This problem has not been solved in [23, 25, 28], but has been solved in [5, 24].

Algorithm 5.1 applied to problem (TP3) terminates at the singular stationary point in 44 iterations and $\mu = 0.01$, $y^* = (0, 1, 0)$, $\lambda^* = (3.4923e$+10, 0.0, 3.4923e+10). The residuals, respectively, are $\|g^k + A_k\lambda^k\| = 1.2716$, $\|Y_k\Lambda_k e - \mu_k e\| = 0.0292$, and $\|c^k + y^k\| = 0.0$. The value of the penalty parameter is $\hat{\rho} = 2.6370e$+10.

TABLE 4. *Numerical results by Algorithm* 5.1.

Problem	Iter	RD	RP	RG	$\hat{\rho}$
TP001	25	2.5953e-11	0	1.0000e-08	1
TP002	22	3.8447e-12	0	1.0000e-08	2
TP003	16	1.9997e-09	0	1.0000e-08	1
TP004	10	1.7656e-13	8.8947e-17	2.0001e-08	4.9402
TP010	18	2.4976e-14	3.0564e-14	1.0000e-08	1
TP011	15	1.1383e-14	9.2021e-16	1.0000e-08	4
TP012	15	2.5011e-14	1.8881e-15	1.0000e-08	1
TP020	38	9.0994e-14	0.5983e-17	5.0000e-08	512
TP021	18	1.3468e-09	0	5.0000e-08	1
TP022	11	1.0991e-12	1.4037e-16	2.0000e-08	1
TP023	14	7.1677e-12	7.1056e-15	9.0000e-08	1
TP024	14	2.5103e-12	4.3581e-16	5.0000e-08	1
TP038	95	7.6785e-09	0	8.0000e-08	1
TP043	22	2.7486e-10	7.2071e-13	3.0000e-08	2
TP044	15	1.3328e-13	7.8580e-16	1.0000e-07	2
TP076	17	2.6222e-09	1.1974e-15	7.0000e-08	1

We also apply our algorithm to some other test problems taken from [17], which are numbered in the same way as that in [17]. For example, "TP022" is Problem 22 in the book. We use these test problems (but not all test problems) since they have only inequality constraints, and thus are suitable for testing the algorithm. The initial points are the same as in [17]. The numerical results are reported in Table 4, where "Iter" represents the number of iterations, $RD = \|g^k + A_k\lambda^k\|$, $RP = \|c^k + y^k\|$, $RG = \|Y_k\Lambda_k e - \mu_k e\|$, and $\hat{\rho}$ is the value of the penalty parameter when the algorithm terminates.

Acknowledgments. The authors are grateful to Professor Philippe L. Toint for reading the draft of this paper and giving insightful comments on it. We also appreciate the associate editor Professor Jorge Nocedal and the anonymous referees for their valuable suggestions and comments.

REFERENCES

[1] P. ARMAND, J. C. GILBERT, AND S. JAN-JÉGOU, *A feasible BFGS interior point algorithm for solving convex minimization problems,* SIAM J. Optim., 11 (2000), pp. 199–222.

[2] J. V. BURKE, *A sequential quadratic programming method for potentially infeasible mathematical programs,* J. Math. Anal. Appl., 139 (1989), pp. 319–351.

[3] J. V. BURKE AND S. P. HAN, *A robust sequential quadratic programming method,* Math. Program., 43 (1989), pp. 277–303.

[4] R. H. BYRD, J. C. GILBERT, AND J. NOCEDAL, *A trust region method based on interior point techniques for nonlinear programming,* Math. Program., 89 (2000), pp. 149–185.

[5] R. H. BYRD, M. E. HRIBAR, AND J. NOCEDAL, *An interior point algorithm for large-scale nonlinear programming,* SIAM J. Optim., 9 (1999), pp. 877–900.

[6] R. H. BYRD, G. LIU, AND J. NOCEDAL, *On the local behavior of an interior point algorithm for nonlinear programming,* in Numerical Analysis 1997, D. F. Griffiths and D. J. Higham, eds., Longman, Harlow, UK, 1998, pp. 37–56.

[7] R. H. BYRD, M. MARAZZI, AND J. NOCEDAL, *On the convergence of Newton iterations to non-stationary points,* Math. Program., 99 (2004), pp. 127–148.

[8] A. R. CONN, N. GOULD, AND PH. L. TOINT, *A primal-dual algorithm for minimizing a nonconvex function subject to bound and linear equality constraints,* in Nonlinear Optimization and Related Topics, G. Dipillo and F. Giannessi, eds., Kluwer Academic Publishers, Dordrecht, 2000, pp. 15–49.

[9] J. E. Dennis, Jr., M. El-Alem, and M. C. Maciel, *A global convergence theory for general trust-region-based algorithms for equality constrained optimization,* SIAM J. Optim., 7 (1997), pp. 177–207.

[10] J. E. Dennis and L. N. Vicente, *On the convergence theory of trust-region-based algorithms for equality-constrained optimization,* SIAM J. Optim., 7 (1997), pp. 927–950.

[11] A. S. El-Bakry, R. A. Tapia, T. Tsuchiya, and Y. Zhang, *On the formulation and theory of the Newton interior-point method for nonlinear programming*, J. Optim. Theory Appl., 89 (1996), pp. 507–541.

[12] A. V. Fiacco and G. P. McCormick, *Nonlinear Programming: Sequential Unconstrained Minimization Techniques,* John Wiley and Sons, New York, 1968; republished as Classics in Appl. Math. 4, SIAM, Philadelphia, 1990.

[13] R. Fletcher, *Practical Methods for Optimization. Vol.* 2*: Constrained Optimization,* John Wiley and Sons, Chichester, 1981.

[14] R. Fletcher, *A model algorithm for composite nondifferentiable optimization problems,* Math. Prog. Stud., 17 (1982), pp. 67–76.

[15] A. Forsgren and Ph. E. Gill, *Primal-dual interior methods for nonconvex nonlinear programming*, SIAM J. Optim., 8 (1998), pp. 1132–1152.

[16] D. M. Gay, M. L. Overton, and M. H. Wright, *A primal-dual interior method for nonconvex nonlinear programming*, in Advances in Nonlinear Programming, Y. X. Yuan, ed., Kluwer Academic Publishers, Dordrecht, 1998, pp. 31–56.

[17] W. Hock and K. Schittkowski, *Test Examples for Nonlinear Programming Codes,* Lecture Notes in Econom. and Math. Systems 187, Springer, New York, 1981.

[18] L. S. Lasdon, J. Plummer, and G. Yu, *Primal-dual and primal interior point algorithms for general nonlinear programs,* ORSA J. Comput., 7 (1995), pp. 321–332.

[19] X.-W. Liu, *A globally convergent, locally superlinearly convergent algorithm for equality constrained optimization,* in Numerical Linear Algebra and Optimization, Y. X. Yuan, ed., Science Press, Beijing, New York, 1999, pp. 131–144.

[20] X.-W. Liu and Y.-X. Yuan, *A robust algorithm for optimization with general equality and inequality constraints,* SIAM J. Sci. Comput., 22 (2000), pp. 517–534.

[21] E. O. Omojokun, *Trust Region Algorithms for Optimization with Nonlinear Equality and Inequality Constraints,* Ph. D. Dissertation, University of Colorado, Boulder, CO, 1991.

[22] M. J. D. Powell and Y. Yuan, *A trust region algorithm for equality constrained optimization,* Math. Program., 49 (1991), pp. 189–211.

[23] D. F. Shanno and R. J. Vanderbei, *Interior-point methods for nonconvex nonlinear programming: Orderings and higher-order methods,* Math. Program., 87 (2000), pp. 303–316.

[24] P. Tseng, *Convergent infeasible interior-point trust-region methods for constrained minimization,* SIAM J. Optim., 13 (2002), pp. 432–469.

[25] R. J. Vanderbei and D. F. Shanno, *An interior-point algorithm for nonconvex nonlinear programming,* Comput. Optim. Appl., 13 (1999), pp. 231–252.

[26] A. Wächter and L. T. Biegler, *Failure of global convergence for a class of interior point methods for nonlinear programming,* Math. Program., 88 (2000), pp. 565–574.

[27] M. H. Wright, *Why a pure primal Newton barrier step may be infeasible,* SIAM J. Optim., 5 (1995), pp. 1–12.

[28] H. Yamashita, *A globally convergent primal-dual interior point method for constrained optimization,* Optim. Methods Softw., 10 (1998), pp. 448–469.

[29] H. Yamashita and H. Yabe, *Superlinear and quadratic convergence of some primal-dual interior point methods for constrained optimization,* Math. Program., 75 (1996), pp. 377–397.

[30] Y. Yuan, *On the convergence of a new trust region algorithm,* Numer. Math., 70 (1995), pp. 515–539.

SIAM J. OPTIM.
Vol. 14, No. 4, pp. 1187–1215

REGIME SWITCHING STOCHASTIC APPROXIMATION ALGORITHMS WITH APPLICATION TO ADAPTIVE DISCRETE STOCHASTIC OPTIMIZATION*

G. YIN†, VIKRAM KRISHNAMURTHY‡, AND CRISTINA ION†

Abstract. This work is devoted to a class of stochastic approximation problems with regime switching modulated by a discrete-time Markov chain. Our motivation stems from using stochastic recursive algorithms for tracking Markovian parameters such as those in spreading code optimization in CDMA (code division multiple access) wireless communication. The algorithm uses constant step size to update the increments of a sequence of occupation measures. It is proved that least squares estimates of the tracking errors can be developed. Assume that the adaptation rate is of the same order of magnitude as that of the time-varying parameter, which is more difficult to deal with than that of slower parameter variations. Due to the time-varying characteristics and Markovian jumps, the usual stochastic approximation (SA) techniques cannot be carried over in the analysis. By a combined use of the SA method and two-time-scale Markov chains, asymptotic properties of the algorithm are obtained, which are distinct from the usual SA results. In this paper, it is shown for the first time that, under simple conditions, a continuous-time interpolation of the iterates converges weakly not to an ODE, as is widely known in the literature, but to a system of ODEs with regime switching, and that a suitably scaled sequence of the tracking errors converges not to a diffusion but to a system of switching diffusion. As an application of these results, the performance of an adaptive discrete stochastic optimization algorithm is analyzed.

Key words. stochastic approximation, Markovian parameter, time-varying parameter, regime switching model, tracking, regime switching ODEs, switching diffusion

AMS subject classifications. 60J10, 60J27, 62L20, 93C50, 93E10

DOI. 10.1137/S1052623403423709

1. Introduction.

In this paper, we consider a class of stochastic approximation (SA) algorithms for tracking the invariant distribution of a conditional Markov chain (conditioned on another Markov chain whose transition probability matrix is "near" identity). Here and henceforth, we refer to such a Markov chain with infrequent jumps as a slow Markov chain, for simplicity. It is well known that if the parameter changes too drastically, there is no chance one can track the time-varying properties using an SA algorithm. Such a phenomenon is known as tracking capability; see [4] for related discussions. Our objectives include evaluating the tracking capability of the SA algorithm in terms of mean squares tracking error, characterizing the dynamic behavior of the iterates, revealing the structure of a scaled sequence of tracking errors, and obtaining the asymptotic covariance of the associated limit process.

Motivation. While there are several papers that analyze tracking properties of SA algorithms when the underlying parameter varies according to a slow random walk [4, 19], fewer papers consider the case when the underlying parameter evolves according to a slow Markov chain. Yet such slow Markov chain models arise in several

*Received by the editors March 3, 2003; accepted for publication (in revised form) December 29, 2003; published electronically August 4, 2004.
http://www.siam.org/journals/siopt/14-4/42370.html

†Department of Mathematics, Wayne State University, Detroit, MI 48202 (gyin@math.wayne.edu, cion@math.wayne.edu). The research of the first author was supported in part by the National Science Foundation under DMS-0304928, and in part by Wayne State University Research Enhancement Program. The research of the third author was supported in part by Wayne State University.

‡Department of Electrical Engineering, University of British Columbia, Vancouver, BC V6T 1Z4, Canada, and Department of Electrical and Electronic Engineering, University of Melbourne (vikramk@ece.ubc.ca). The research of this author was supported in part by NSERC and ARC.

applications. The main motivation for our work stems from applications in discrete stochastic optimization. Such problems appeared in [21] and were subsequently considered in [2, 3, 10] among others; we refer the reader to [20] for a recent survey of several methods for discrete stochastic optimization including selection and multiple comparison methods, multi-armed bandits, the stochastic ruler, nested partition methods, and discrete stochastic optimization algorithms based on simulated annealing [1, 2, 3, 9].

The discrete stochastic optimization algorithms in [2, 3] can be thought of as random search procedures, in which there is a feasible set $\mathcal{S}$ that contains the minima together with other potential search candidates. One devises a strategy so that the optimal parameter (minimum) is estimated with minimal effort. An important variation of this is to devise and analyze the performance of an *adaptive* discrete stochastic optimization algorithm when the underlying parameter (minimum) is slowly time-varying. Such tracking problems lie at the heart of applications of SA algorithms. In such cases, because the parameter set is finite, it is often reasonable to assume that the underlying parameter (termed "hypermodel" in [4]) evolves according to a slow finite state Markov chain. As will be shown in section 6, the general tracking analysis presented in this paper for a slow Markov chain parameter readily applies to analyzing the tracking performance of such adaptive discrete stochastic optimization algorithms. To the best of our knowledge, this is the first time a tracking analysis has been presented for a discrete stochastic optimization algorithm.

Applications. Discrete stochastic optimization problems arise in emerging applications such as adaptive coding in wireless CDMA (code division multiple access) communication networks. In our recent work [11], we considered optimizing the spreading code of the CDMA system at the transmitter. This was formulated as a discrete stochastic optimization problem (since the spreading codes are finite-length and finite-state sequences), and the random-search–based discrete stochastic optimization algorithm of [2] was used to compute the optimal spreading code. In addition to the random-search–type algorithms, we also designed adaptive SA algorithms with both fixed step size and adaptive step sizes to track slowly time-varying optimal spreading codes caused by fading characteristics of the wireless channel. The numerical results in [11, 12] have shown remarkable improvement compared with that of several heuristic algorithms. Section 6 explicitly derives performance bounds in terms of error probabilities for the adaptive discrete stochastic optimization algorithm.

Outline. This paper considers an algorithm with constant step size and updates that are essentially of the form of occupation measures. We are interested in the analysis of tracking errors. First, using perturbed Lyapunov function methods [16], we derive mean squares–type error bounds. The argument is mainly based on stability analysis. Naturally, one then asks whether an associated limit ODE (ordinary differential equation) can be derived via ODE methods as in the usual analysis of SA and stochastic optimization–type algorithms. The standard ODE method cannot be carried over due to the fact that the system is now time-varying, and the adaptation rate is the same as that of the parameter variation. By a combined use of the updated treatment on SA [16] and two-time-scale Markov chains [22, 23], we demonstrate that a limit system can still be obtained. However, very different from the usual stochastic approximation methods in the existing literature, the limit system is no longer a single ODE, but a system of ODEs modulated by a continuous-time Markov chain. Thus, the limit is not deterministic but stochastic. Such systems are referred to as ODEs with regime switching. Based on the system of switching ODEs obtained, we further examine a sequence of suitably normalized errors aiming at understanding the

rate of variation (rate of convergence) of the scaled sequence of tracking errors. It is well known that for an SA algorithm, if the true parameter is a fixed constant, then a suitably scaled sequence of estimation errors has a Gaussian diffusion limit. In contrast, somewhat remarkably, the scaled tracking error sequence generated by the SA algorithm in this paper does not have a diffusion limit. Instead, the limit is a system of diffusions with regime switching. In the limit system, the diffusion coefficient depends on the modulating Markov chain, which reveals the distinctive time-varying nature of the underlying system and provides new insight on Markov modulated SA problems.

Context. The main weak convergence results in this paper in sections 4 and 5 assume that the dynamics of the true parameter (modeled as a slow Markov chain with transition probability matrix $I + \varepsilon Q$) evolves on the same time scale as the adaptive SA algorithm with step size μ, i.e., $\varepsilon = O(\mu)$. We note that the case $\varepsilon = O(\mu)$ addressed in this paper is much more difficult to handle than $\varepsilon = o(\mu)$ (e.g., $\varepsilon = O(\mu^2)$), which is widely used in the analysis of tracking algorithms [4]. The meaning of $\varepsilon = o(\mu)$ is that the true parameter evolves much more slowly than the adaptation speed of the stochastic optimization algorithm and is more restrictive than $\varepsilon = O(\mu)$. Furthermore, with $\varepsilon = o(\mu)$ one obtains a standard ODE and linear diffusion limit, whereas with $\varepsilon = O(\mu)$ we show for the first time in this paper that one obtains a randomly switching system of ODEs and switching diffusion limit. Finally, in several applications arising in wireless telecommunication network optimization, e.g., signature code optimization in spread spectrum systems over fading channels [11, 12], the optimal signature sequence (true parameter) changes as quickly as the adaptation of the algorithm, i.e., $\varepsilon = O(\mu)$.

The rest of the paper is organized as follows. Section 2 contains the formulation of the problem. Section 3 is devoted to obtaining mean squares error bounds. In section 4, we obtain a weak convergence result of an interpolated sequence of the iterates. Section 5 further examines a suitably scaled tracking error sequence of the iterates and derives a switching diffusion limit. Section 6 presents an example of an adaptive discrete stochastic optimization algorithm, which is motivated by [11], where such algorithms have been used to perform adaptive spreading code optimization in wireless CDMA systems. The analysis of section 3 and section 5 is used to derive bounds on the error probability of this adaptive discrete stochastic optimization algorithm.

Before proceeding, a bit of notation is in order. Throughout the paper, z' denotes the transpose of $z \in \mathbb{R}^{\ell \times r}$ for some $\ell, r \geq 1$; unless otherwise noted, all vectors are column vectors; $|z|$ denotes the norm of z; K denotes a generic positive constant whose values may vary for different usage (the conventions $K + K = K$ and $KK = K$ will be used without notice).

2. Formulation of the problem. We will use the following conditions throughout the paper. Condition (M) characterizes the time-varying underlying parameter as a Markov chain with infrequent transitions, while condition (S) characterizes the observed signal.

(M) Let $\{\theta_n\}$ be a discrete-time Markov chain with finite state space

$$\mathcal{M} = \{\overline{\theta}_1, \ldots, \overline{\theta}_{m_0}\} \tag{2.1}$$

and transition probability matrix

$$P^\varepsilon = I + \varepsilon Q, \tag{2.2}$$

where $\varepsilon > 0$ is a small parameter, I is an $m_0 \times m_0$ identity matrix, and $Q = (q_{ij}) \in \mathbb{R}^{m_0 \times m_0}$ is a generator of a continuous-time Markov chain (i.e., Q satisfies $q_{ij} \geq 0$ for $i \neq j$ and $\sum_{j=1}^{m_0} q_{ij} = 0$ for each $i = 1, \ldots, m_0$). For simplicity, suppose that the initial distribution $P(\theta_0 = \overline{\theta}_i) = p_{0,i}$ is independent of ε for each $i = 1, \ldots, m_0$, where $p_{0,i} \geq 0$ and $\sum_{i=1}^{m_0} p_{0,i} = 1$. Q is irreducible.

(S) Let $\{X_n\}$ be an S-state conditional Markov chain (conditioned on the parameter process). The state space of $\{X_n\}$ is $\mathcal{S} = \{e_1, \ldots, e_S\}$, where e_i for $i = 1, \ldots, S$ denotes the ith standard unit vectors, with the ith component being 1 and the rest of the components being 0. For each $\theta \in \mathcal{M}$, $A(\theta) = (a_{ij}(\theta)) \in \mathbb{R}^{S \times S}$, the transition probability matrix of X_n is defined by

$$a_{ij}(\theta) = P(X_{n+1} = e_j | X_n = e_i, \theta_n = \theta) = P(X_1 = e_j | X_0 = e_i, \theta_0 = \theta),$$

where $i, j \in \{1, \ldots, S\}$. For $\theta \in \mathcal{M}$, $A(\theta)$ is irreducible and aperiodic.

Remark 2.1. Note that the underlying Markov chain $\{\theta_n\}$ is in fact ε-dependent. We suppress the ε-dependence for notational simplicity. The small parameter ε in (2.2) ensures that the entries of the transition probability matrix are nonnegative, since $p_{ij}^{\varepsilon} = \delta_{ij} + \varepsilon q_{ij} \geq 0$ for $\varepsilon > 0$ small enough, where δ_{ij} denotes the Kronecker δ satisfying $\delta_{ij} = 1$ if $i = j$ and 0 otherwise. The use of the generator Q makes the row sum of the matrix P be one. The main idea is that, although the true parameter is time-varying, it is piecewise constant. Moreover, due to the dominating identity matrix in (2.2), $\{\theta_n\}$ varies slowly in time. The time-varying parameter takes a constant value $\overline{\theta}_i$ for a random duration and jumps to another state $\overline{\theta}_j$ with $j \neq i$ at a random time.

The assumptions on irreducibility and aperiodicity of $A(\theta)$ imply that for each $\theta \in \mathcal{M}$ there exists a unique stationary distribution $\pi(\theta) \in \mathbb{R}^{S \times 1}$ satisfying

$$\pi'(\theta) = \pi'(\theta) A(\theta) \quad \text{and} \quad \pi'(\theta) \mathbb{1}_S = 1,$$

where $\mathbb{1}_\ell \in \mathbb{R}^{\ell \times 1}$ with all entries being equal to 1. We aim to use an SA algorithm to track the time-varying distribution $\pi(\theta_n)$ that depends on the underlying Markov chain θ_n.

2.1. Adaptive algorithm. We use the following adaptive algorithm of least mean squares (LMS) type with constant step size in order to construct a sequence of estimates $\{\widehat{\pi}_n\}$ of the time-varying distribution $\pi(\theta_n)$,

$$\widehat{\pi}_{n+1} = \widehat{\pi}_n + \mu(X_{n+1} - \widehat{\pi}_n), \tag{2.3}$$

where μ denotes the step size. Define $\widetilde{\pi}_n = \widehat{\pi}_n - \mathbf{E}\pi(\theta_n)$. Then (2.3) can be rewritten as

$$\widetilde{\pi}_{n+1} = \widetilde{\pi}_n - \mu\widetilde{\pi}_n + \mu(X_{n+1} - \mathbf{E}\pi(\theta_n)) + \mathbf{E}(\pi(\theta_n) - \pi(\theta_{n+1})). \tag{2.4}$$

Note that $\widehat{\pi}_n$, $\pi(\theta_n)$, and hence $\widetilde{\pi}_n$ are column vectors (i.e., they take values in $\mathbb{R}^{S \times 1}$).

The underlying parameter θ_n is called a *hypermodel* in [4]. Note that while the dynamics of the hypermodel θ_n is used in our analysis, it does not explicitly enter the implementation of the LMS algorithm (2.3).

To accomplish our goal, we derive a mean squares error bound, proceed with the examination of an interpolated sequence of the iterates, and derive a limit result for a scaled sequence. These three steps are realized in the following three sections.

3. Mean square error. This section establishes a mean square estimate for $\mathbf{E}|\widetilde{\pi}_n|^2 = \mathbf{E}|\widehat{\pi}_n - \mathbf{E}\pi(\theta_n)|^2$. Analyzing SA algorithms often requires the use of Lyapunov-type functions for proving stability; see [7, 16]. In what follows, we obtain the desired estimate via a stability argument using the perturbed Lyapunov function method [16]. Use $\mathbf{E}_n$ to denote the conditional expectation with respect to $\mathcal{F}_n$, the σ-algebra generated by $\{X_k, \theta_k : k \le n\}$.

THEOREM 3.1. *Assume* (M) *and* (S). *In addition, suppose that* $\varepsilon^2 \ll \mu$. *Then for sufficiently large* n,

$$\mathbf{E}|\widetilde{\pi}_n|^2 = O\left(\mu + \varepsilon + \frac{\varepsilon^2}{\mu}\right). \tag{3.1}$$

Proof. Define $V(x) = (x'x)/2$. Direct calculations lead to

$$\begin{aligned}\mathbf{E}_n V(\widetilde{\pi}_{n+1}) - V(\widetilde{\pi}_n) &= \mathbf{E}_n\{\widetilde{\pi}_n'[-\mu\widetilde{\pi}_n + \mu(X_{n+1} - \mathbf{E}\pi(\theta_n)) + \mathbf{E}[\pi(\theta_n) - \pi(\theta_{n+1})]]\}\\ &\quad + \mathbf{E}_n|-\mu\widetilde{\pi}_n + \mu(X_{n+1} - \mathbf{E}\pi(\theta_n)) + \mathbf{E}[\pi(\theta_n) - \pi(\theta_{n+1})]|^2. \end{aligned}\tag{3.2}$$

In view of the Markovian assumption and the structure of the transition probability matrix given by (2.2),

$$\begin{aligned}\mathbf{E}_n[\pi(\theta_n) - \pi(\theta_{n+1})] &= \mathbf{E}[\pi(\theta_n) - \pi(\theta_{n+1})|\theta_n]\\ &= \sum_{i=1}^{m_0} \mathbf{E}[\pi(\overline{\theta}_i) - \pi(\theta_{n+1})|\theta_n = \overline{\theta}_i] I_{\{\theta_n = \overline{\theta}_i\}}\\ &= \sum_{i=1}^{m_0}\left[\pi(\overline{\theta}_i) - \sum_{j=1}^{m_0}\pi(\overline{\theta}_j)p_{ij}^{\varepsilon}\right] I_{\{\theta_n = \overline{\theta}_i\}}\\ &= -\varepsilon\sum_{i=1}^{m_0}\sum_{j=1}^{m_0}\pi(\overline{\theta}_j)q_{ij}I_{\{\theta_n = \overline{\theta}_i\}}\\ &= O(\varepsilon),\end{aligned}\tag{3.3}$$

and likewise, detailed computation also shows that

$$\mathbf{E}_n|\pi(\theta_n) - \pi(\theta_{n+1})|^2 = O(\varepsilon). \tag{3.4}$$

Owing to (2.2), the transition probability matrix P^{ε} is independent of time n. As a result, the k-step transition probability depends only on the time lags and can be denoted by $(P^{\varepsilon})^k$. By an elementary inequality, we have $|\widetilde{\pi}_n| = |\widetilde{\pi}_n|\cdot 1 \le (|\widetilde{\pi}_n|^2 + 1)/2$. Thus,

$$O(\varepsilon)|\widetilde{\pi}_n| \le O(\varepsilon)(V(\widetilde{\pi}_n) + 1).$$

Noting that the sequence of signals $\{X_n\}$ is bounded, the boundedness of $\{\widehat{\pi}_n\}$, and $O(\varepsilon\mu) = O(\mu^2 + \varepsilon^2)$ via the elementary inequality $ab \le (a^2 + b^2)/2$ for any real numbers a and b, the estimate (3.4) yields

$$\begin{aligned}&\mathbf{E}_n|-\mu\widetilde{\pi}_n + \mu(X_{n+1} - \mathbf{E}\pi(\theta_n)) + \mathbf{E}[\pi(\theta_n) - \pi(\theta_{n+1})]|^2\\ &\le K\mathbf{E}_n\Big[\mu^2|\widetilde{\pi}_n|^2 + \mu^2|X_{n+1} - \mathbf{E}\pi(\theta_n)|^2 + \mu^2|\widetilde{\pi}_n'\mathbf{E}(X_{n+1} - \mathbf{E}\pi(\theta_n))|\\ &\qquad + \mu|\widetilde{\pi}_n'\mathbf{E}(\pi(\theta_n) - \pi(\theta_{n+1}))| + \mu|(X_{n+1} - \mathbf{E}\pi(\theta_n))'\mathbf{E}(\pi(\theta_n) - \pi(\theta_{n+1}))|\Big]\\ &\quad + |\mathbf{E}(\pi(\theta_n) - \pi(\theta_{n+1}))|^2\\ &= O(\mu^2 + \varepsilon^2)(V(\widetilde{\pi}_n) + 1) + |\mathbf{E}(\pi(\theta_n) - \pi(\theta_{n+1}))|^2\end{aligned}\tag{3.5}$$

and

$$\begin{aligned}(3.6)\qquad &\mathbf{E}_n\{\widetilde{\pi}_n'[-\mu\widetilde{\pi}_n+\mu(X_{n+1}-\mathbf{E}\pi(\theta_n))+\mathbf{E}(\pi(\theta_n)-\pi(\theta_{n+1}))]\}\\ &\quad=-2\mu V(\widetilde{\pi}_n)+\mu\mathbf{E}_n\widetilde{\pi}_n'(X_{n+1}-\mathbf{E}\pi(\theta_n))+\mathbf{E}_n\widetilde{\pi}_n'\mathbf{E}(\pi(\theta_n)-\pi(\theta_{n+1})).\end{aligned}$$

Using (3.5) and (3.6) in (3.2) together with (3.3), we obtain

$$\begin{aligned}(3.7)\qquad &\mathbf{E}_nV(\widetilde{\pi}_{n+1})-V(\widetilde{\pi}_n)\\ &\quad=-2\mu V(\widetilde{\pi}_n)+\mu\mathbf{E}_n\widetilde{\pi}_n'(X_{n+1}-\mathbf{E}\pi(\theta_n))+\mathbf{E}_n\widetilde{\pi}_n'\mathbf{E}(\pi(\theta_n)-\pi(\theta_{n+1}))\\ &\qquad+O(\mu^2+\varepsilon^2)(V(\widetilde{\pi}_n)+1).\end{aligned}$$

To obtain the desired estimate, we need to "average out" the second to the fourth terms on the right-hand side of (3.7). To do so, for any $0<T<\infty$, we define the following perturbations:

$$\begin{aligned}(3.8)\qquad V_1^\varepsilon(\widetilde{\pi},n)&=\mu\sum_{j=n}^{T/\varepsilon}\widetilde{\pi}'\mathbf{E}_n(X_{j+1}-\mathbf{E}\pi(\theta_j)),\\ V_2^\varepsilon(\widetilde{\pi},n)&=\sum_{j=n}^{T/\varepsilon}\widetilde{\pi}'\mathbf{E}(\pi(\theta_j)-\pi(\theta_{j+1})).\end{aligned}$$

In the above and hereafter, T/ε is understood to be $\lfloor T/\varepsilon\rfloor$, i.e., the integer part of T/ε.

Throughout the rest of the paper, we often need to use the notion of fixed-θ processes. For example, by $X_j(\theta)$ for $n\le j\le O(1/\varepsilon)$, we mean a process in which $\theta_j=\theta$ is fixed for all j with $n\le j\le O(1/\varepsilon)$.

For $V_1^\varepsilon(\widetilde{\pi},n)$ defined in (3.8),

$$\begin{aligned}(3.9)\qquad \left|\sum_{j=n}^{T/\varepsilon}\mathbf{E}_n[X_{j+1}-\pi(\theta_j)]\right|&\le\left|\sum_{j=n}^{T/\varepsilon}\mathbf{E}_n[X_{j+1}-\mathbf{E}X_{j+1}]\right|\\ &\quad+\left|\sum_{j=n}^{T/\varepsilon}[\mathbf{E}X_{j+1}-\mathbf{E}\pi(\theta_j)]\right|.\end{aligned}$$

Using the ϕ-mixing property of $\{X_j\}$ (see [5, p. 166]),

$$(3.10)\qquad \left|\sum_{j=n}^{T/\varepsilon}\mathbf{E}_n[X_{j+1}-\mathbf{E}X_{j+1}]\right|\le K<\infty\qquad\text{uniformly in } n.$$

We can also show

$$(3.11)\qquad \left|\sum_{j=n}^{T/\varepsilon}[\mathbf{E}X_{j+1}-\mathbf{E}\pi(\theta_j)]\right|<\infty.$$

Thus, using (3.9)–(3.11), for each $\widetilde{\pi}$,

$$(3.12)\qquad |V_1^\varepsilon(\widetilde{\pi},n)|\le O(\mu)(V(\widetilde{\pi})+1).$$

By virtue of the definition of $V_2^\varepsilon(\cdot)$ and (2.2), it follows that there exists an N_ε for all $n\ge N_\varepsilon$ such that

$$\begin{aligned}(3.13)\qquad |V_2^\varepsilon(\widetilde{\pi},n)|&=\left|\sum_{j=n}^{T/\varepsilon}\widetilde{\pi}'[\mathbf{E}(\pi(\theta_j)-\pi(\theta_{j+1}))]\right|\\ &=\left|\widetilde{\pi}'\mathbf{E}[\pi(\theta_n)-\pi(\theta_{T/\varepsilon})]\right|\\ &\le|\widetilde{\pi}|O(\varepsilon)\\ &\le O(\varepsilon)(V(\widetilde{\pi})+1).\end{aligned}$$

We next show that they result in the desired cancellation in the error estimate. Note that

$$\begin{aligned}\mathbf{E}_n V_1^\varepsilon(\widetilde{\pi}_{n+1}, n+1) - V_1^\varepsilon(\widetilde{\pi}_n, n) \\ = \mathbf{E}_n[V_1^\varepsilon(\widetilde{\pi}_{n+1}, n+1) - V_1^\varepsilon(\widetilde{\pi}_n, n+1)] + \mathbf{E}_n V_1^\varepsilon(\widetilde{\pi}_n, n+1) - V_1^\varepsilon(\widetilde{\pi}_n, n).\end{aligned} \tag{3.14}$$

It can be seen that

$$\mathbf{E}_n V_1^\varepsilon(\widetilde{\pi}_n, n+1) - V_1^\varepsilon(\widetilde{\pi}_n, n) = -\mu \mathbf{E}_n \widetilde{\pi}_n'(X_{n+1} - \mathbf{E}\pi(\theta_n)) \tag{3.15}$$

and

$$\begin{aligned}&\mathbf{E}_n V_1^\varepsilon(\widetilde{\pi}_{n+1}, n+1) - \mathbf{E}_n V_1^\varepsilon(\widetilde{\pi}_n, n+1) \\ &= \mu \sum_{j=n+1}^{T/\varepsilon} \mathbf{E}_n \widetilde{\pi}_{n+1}' \mathbf{E}_{n+1}(X_{j+1} - \mathbf{E}\pi(\theta_j)) - \mu \sum_{j=n+1}^{T/\varepsilon} \mathbf{E}_n \widetilde{\pi}_n' \mathbf{E}_{n+1}(X_{j+1} - \mathbf{E}\pi(\theta_j)) \\ &= \mu \sum_{j=n+1}^{T/\varepsilon} \mathbf{E}_n (\widetilde{\pi}_{n+1} - \widetilde{\pi}_n)' \mathbf{E}_{n+1}(X_{j+1} - \mathbf{E}\pi(\theta_j)) \\ &= \mu \sum_{j=n+1}^{T/\varepsilon} \mathbf{E}_n [-\mu \widetilde{\pi}_n + \mu(X_{n+1} - \mathbf{E}\pi(\theta_n)) + \mathbf{E}(\pi(\theta_n) - \pi(\theta_{n+1}))]' \mathbf{E}_{n+1}[X_{j+1} - \mathbf{E}\pi(\theta_j)] \\ &= O(\mu^2)(V(\widetilde{\pi}_n) + 1) + O(\mu\varepsilon) = O(\mu^2)(V(\widetilde{\pi}_n) + 1) + O(\varepsilon^2).\end{aligned} \tag{3.16}$$

In the above, we have used $O(\mu\varepsilon) = O(\mu^2 + \varepsilon^2)$, (2.4), and (3.2) to obtain

$$\begin{aligned}|\mathbf{E}_n[\widetilde{\pi}_{n+1} - \widetilde{\pi}_n]| &\le \mu \mathbf{E}_n |\widetilde{\pi}_n| + \mu \mathbf{E}_n |X_{n+1} - \mathbf{E}\pi(\theta_n)| + O(\varepsilon) \\ &= O(\mu)(V(\widetilde{\pi}_n) + 1) + O(\varepsilon).\end{aligned} \tag{3.17}$$

Thus

$$\begin{aligned}\mathbf{E}_n V_1^\varepsilon(\widetilde{\pi}_{n+1}, n+1) - V_1^\varepsilon(\widetilde{\pi}_n, n) \\ = -\mu \mathbf{E}_n \widetilde{\pi}_n'(X_{n+1} - \mathbf{E}\pi(\theta_n)) + O(\mu^2)(V(\widetilde{\pi}_n) + 1) + O(\varepsilon^2).\end{aligned} \tag{3.18}$$

Analogous estimates yield that

$$\begin{aligned}&\mathbf{E}_n V_2^\varepsilon(\widetilde{\pi}_{n+1}, n+1) - \mathbf{E}_n V_2^\varepsilon(\widetilde{\pi}_n, n+1) \\ &\quad = \sum_{j=n+1}^{T/\varepsilon} \mathbf{E}_n (\widetilde{\pi}_{n+1} - \widetilde{\pi}_n)' \mathbf{E}(\pi(\theta_j) - \pi(\theta_{j+1})) \\ &\quad = O(\mu\varepsilon)(V(\widetilde{\pi}_n) + 1) + O(\varepsilon^2) = O(\varepsilon^2 + \mu^2)(V(\widetilde{\pi}_n) + 1),\end{aligned} \tag{3.19}$$

and that

$$\mathbf{E}_n V_2^\varepsilon(\widetilde{\pi}_n, n+1) - V_2^\varepsilon(\widetilde{\pi}_n, n) = -\widetilde{\pi}_n' \mathbf{E}(\pi(\theta_n) - \pi(\theta_{n+1})). \tag{3.20}$$

Thus,

$$\begin{aligned}&\mathbf{E}_n V_2^\varepsilon(\widetilde{\pi}_{n+1}, n+1) - V_2^\varepsilon(\widetilde{\pi}_n, n) \\ &= -\widetilde{\pi}_n' \mathbf{E}(\pi(\theta_n) - \pi(\theta_{n+1})) + O(\mu^2 + \varepsilon^2)(V(\widetilde{\pi}_n) + 1).\end{aligned} \tag{3.21}$$

Redefine V_1^ε and V_2^ε with T/ε replaced by ∞. Estimates (3.9)–(3.21) still hold.

Define

$$W(\widetilde{\pi}, n) = V(\widetilde{\pi}) + V_1^\varepsilon(\widetilde{\pi}, n) + V_2^\varepsilon(\widetilde{\pi}, n).$$

Then, using the above estimates, we have

$$\begin{aligned}
&\mathbf{E}_n W(\widetilde{\pi}_{n+1}, n+1) - W(\widetilde{\pi}_n, n)\\
&\quad = \mathbf{E}_n V(\widetilde{\pi}_{n+1}) - V(\widetilde{\pi}_n) + \mathbf{E}_n[V_1^\varepsilon(\widetilde{\pi}_{n+1}, n+1) - V_1^\varepsilon(\widetilde{\pi}_n, n)]\\
&\qquad + \mathbf{E}_n[V_2^\varepsilon(\widetilde{\pi}_{n+1}, n+1) - V_2^\varepsilon(\widetilde{\pi}_n, n)]\\
&\quad = -2\mu V(\widetilde{\pi}_n) + O(\mu^2 + \varepsilon^2)(V(\widetilde{\pi}_n) + 1).
\end{aligned} \tag{3.22}$$

This, together with (3.12) and (3.13) and T/ε replaced by ∞, implies

$$\begin{aligned}
&\mathbf{E}_n W(\widetilde{\pi}_{n+1}, n+1) - W(\widetilde{\pi}_n, n)\\
&\qquad \le -2\mu W(\widetilde{\pi}_n, n) + O(\mu^2 + \varepsilon^2)(W(\widetilde{\pi}_n, n) + 1).
\end{aligned} \tag{3.23}$$

Choose μ and ε small enough so that there is a $\lambda > 0$ satisfying

$$-2\mu + O(\varepsilon^2) + O(\mu^2) \le -\lambda\mu.$$

Then, we get

$$\mathbf{E}_n W(\widetilde{\pi}_{n+1}, n+1) \le (1 - \lambda\mu) W(\widetilde{\pi}_n, n) + O(\mu^2 + \varepsilon^2). \tag{3.24}$$

Taking the expectation and iterating on the resulting inequality yields

$$\begin{aligned}
\mathbf{E}W(\widetilde{\pi}_{n+1}, n+1) &\le (1 - \lambda\mu)^{n - N_\varepsilon} \mathbf{E}W(\widetilde{\pi}_0, 0) + \sum_{j=N_\varepsilon}^{n} (1 - \lambda\mu)^{j - N_\varepsilon} O(\mu^2 + \varepsilon^2)\\
&\le (1 - \lambda\mu)^{n - N_\varepsilon} \mathbf{E}W(\widetilde{\pi}_0, 0) + O\left(\mu + \frac{\varepsilon^2}{\mu}\right).
\end{aligned} \tag{3.25}$$

By taking n large enough, we can make $(1 - \lambda\mu)^{n - N_\varepsilon} = O(\mu)$. Then

$$\mathbf{E}W(\widetilde{\pi}_{n+1}, n+1) \le O\left(\mu + \frac{\varepsilon^2}{\mu}\right). \tag{3.26}$$

Finally, applying (3.12) and (3.13) again, replacing $W(\widetilde{\pi}, n)$ by $V(\widetilde{\pi})$ adds another $O(\varepsilon)$ term. Thus we obtain

$$\mathbf{E}V(\widetilde{\pi}_{n+1}) \le O\left(\mu + \varepsilon + \frac{\varepsilon^2}{\mu}\right). \tag{3.27}$$

This concludes the proof. □

Remark 3.2. In view of Theorem 3.1, in order that our adaptive algorithm can track the time-varying parameter, the ratio ε/μ must not be large. Given the order-of-magnitude estimate $O(\mu + \varepsilon + \varepsilon^2/\mu)$, to balance the two terms μ and ε^2/μ, we need to choose $\varepsilon = O(\mu)$. Therefore, we obtain the following result.

COROLLARY 3.3. *Under the conditions of Theorem* 3.1, *if* $\varepsilon = O(\mu)$, *then for sufficiently large* n, $\mathbf{E}|\widetilde{\pi}_n|^2 = O(\mu)$.

4. Limit system of regime switching ODEs. Our objective in this section is to derive a limit system for an interpolated sequence of the iterates. Different from the usual approach of stochastic approximation [4], where $\varepsilon = o(\mu)$, here and henceforth, we take $\varepsilon = O(\mu)$. For notational simplicity, however, we use $\varepsilon = \mu$. For $0 < T < \infty$, we construct a sequence of piecewise constant interpolation of the stochastic approximation iterates $\widehat{\pi}_n$ as

$$\widehat{\pi}^\mu(t) = \widehat{\pi}_n, \qquad t \in [\mu n, \mu n + \mu). \tag{4.1}$$

The process $\widehat{\pi}^\mu(\cdot)$ so defined is in $D([0,T];\mathbb{R}^S)$, which is the space of functions defined on $[0,T]$ taking values in $\mathbb{R}^S$ that are right continuous, have left limits, and are endowed with the Skorohod topology. We use weak convergence methods to carry out the analysis. The application of weak convergence ideas usually requires proof of tightness and the characterization of the limit processes. Different from the usual approach of stochastic approximation, the limit is not a deterministic ODE but rather a system of ODEs modulated by a continuous-time Markov chain.

LEMMA 4.1. *Under conditions* (M) *and* (S), $\{\pi^\mu(\cdot)\}$ *is tight in* $D([0,T];\mathbb{R}^S)$.

Proof. By using the tightness criteria [14, p. 47], it suffices to verify that for any $\delta > 0$ and $0 < s \le \delta$,

$$\lim_{\delta\to 0} \limsup_{\mu\to 0} \mathbf{E}|\widehat{\pi}^\mu(t+s) - \widehat{\pi}^\mu(t)|^2 = 0. \tag{4.2}$$

To begin, note that

$$\begin{aligned} \widehat{\pi}^\mu(t+s) - \widehat{\pi}^\mu(t) &= \widehat{\pi}_{(t+s)/\mu} - \widehat{\pi}_{t/\mu} \\ &= \mu \sum_{k=t/\mu}^{(t+s)/\mu-1} (X_{k+1} - \widehat{\pi}_k). \end{aligned} \tag{4.3}$$

Note also that both the iterates and the observations are bounded uniformly. Then the boundedness of $\{X_k\}$ and $\{\widehat{\pi}_k\}$ implies that

$$\begin{aligned} &\mathbf{E}|\widehat{\pi}^\mu(t+s) - \widehat{\pi}^\mu(t)|^2 \\ &\quad = \mathbf{E}\left[\mu \sum_{k=t/\mu}^{(t+s)/\mu-1} (X_{k+1} - \widehat{\pi}_k)'\right]\left[\mu \sum_{j=t/\mu}^{(t+s)/\mu-1} (X_{j+1} - \widehat{\pi}_j)\right] \\ &\quad = \mu^2 \sum_{k=t/\mu}^{(t+s)/\mu-1} \sum_{j=t/\mu}^{(t+s)/\mu-1} \mathbf{E}(X_{k+1} - \widehat{\pi}_k)'(X_{j+1} - \widehat{\pi}_j) \\ &\quad \le K\mu^2 \left(\frac{t+s}{\mu} - \frac{t}{\mu}\right)^2 \\ &\quad = K((t+s) - t)^2 = O(s^2). \end{aligned} \tag{4.4}$$

Taking $\limsup_{\mu\to 0}$ and then $\lim_{\delta\to 0}$ in (4.4), equation (4.2) is verified, and so the desired tightness follows. □

4.1. Limit of the modulating Markov chain. Consider the Markov chain θ_n. Regarding the probability vector and the n-step transition probability matrix, we have the following approximation results.

LEMMA 4.2. *Suppose that α_n^η is a Markov chain with a finite state space $\mathcal{M} = \mathcal{M}_1 \cup \mathcal{M}_2 \cup \cdots \cup \mathcal{M}_l$ and transition probability matrix*

$$P^\eta = \operatorname{diag}(P^1, \ldots, P^l) + \eta Q, \tag{4.5}$$

where for each i, P^i is a transition probability matrix that is irreducible and aperiodic, and Q is a generator of a continuous-time Markov chain. For simplicity, denote $\mathcal{M} = \{1, \ldots, m_0\}$, $p_n^\eta = (P(\alpha_n^\eta = 1), \ldots, P(\alpha_n^\eta = m_0))$ with $p_0^\eta = p_0$, and the stationary distribution of P^i by ν^i (a row vector) for $i = 1, \ldots, l$. Then for some $k_0 > 0$,

$$p_n^\eta = \operatorname{diag}(\nu^1, \ldots, \nu^l) z(t) + O\left(\eta + \exp\left(\frac{-k_0 t}{\eta}\right)\right), \tag{4.6}$$

where $z(t) \in \mathbb{R}^{1\times l}$ (with $t = \eta n$) satisfies

$$\frac{dz(t)}{dt} = z(t)\overline{Q},\ z(0) = p_0 \operatorname{diag}(\mathbb{1}_{m_1}, \ldots, \mathbb{1}_{m_l}),$$

with

$$\overline{Q} = \operatorname{diag}(\nu^1, \ldots, \nu^l) Q \operatorname{diag}(\mathbb{1}_{m_1}, \ldots, \mathbb{1}_{m_l}). \tag{4.7}$$

In addition, for $n \le O(1/\eta)$, the n-step transition probability matrix satisfies (with $t = \eta n$),

$$(P^\eta)^n = \Xi(t) + O\left(\eta + \exp\left(\frac{-k_0 t}{\eta}\right)\right), \tag{4.8}$$

where

$$\begin{aligned} &\Xi(t) = \operatorname{diag}(\mathbb{1}_{m_1}, \ldots, \mathbb{1}_{m_l})\Theta(t) \operatorname{diag}(\nu^1, \ldots, \nu^l), \\ &\frac{d\Theta(t)}{dt} = \Theta(t)\overline{Q},\ \Theta(0) = I. \end{aligned} \tag{4.9}$$

Proof. The proof is that of Theorems 3.5 and 4.3 of [23]. □

LEMMA 4.3. *Suppose that α_n^η is the Markov chain given in Lemma 4.2. Define an aggregated process $\overline{\alpha}_n^\eta = i$ if $\alpha_n^\eta \in \mathcal{M}_i$, and define an interpolated process $\overline{\alpha}^\eta(\cdot)$ by $\overline{\alpha}^\eta(t) = \overline{\alpha}_n^\eta$ if $t \in [n\eta, n\eta + \eta)$. Then $\overline{\alpha}^\eta(\cdot)$ converges weakly to $\overline{\alpha}(\cdot)$, which is a continuous-time Markov chain generated by $\overline{Q}$ given in* (4.7).

Proof. The proof of this result can be found in [24].

With the above two lemmas, we can now derive a result that will be used in the subsequent analysis. The proof is essentially an application of the above lemmas.

PROPOSITION 4.4. *Assume* (M). *Choose $\varepsilon = \mu$ and consider the Markov chain θ_n. Then the following assertions hold:*

- *Denote $p_n^\mu = (P(\theta_n = \overline{\theta}_1), \ldots, P(\theta_n = \overline{\theta}_{m_0}))$. Then*

$$\begin{aligned} &p_n^\mu = z(t) + O\left(\mu + \exp\left(\frac{-k_0 t}{\mu}\right)\right), \qquad z(t) \in \mathbb{R}^{1\times m_0}, \\ &\frac{dz(t)}{dt} = z(t)Q, \qquad z(0) = p_0, \\ &(P^\mu)^n = Z(t) + O\left(\mu + \exp\left(\frac{-k_0 t}{\mu}\right)\right), \\ &\frac{dZ(t)}{dt} = Z(t)Q, \qquad Z(0) = I. \end{aligned} \tag{4.10}$$

- *Define the continuous-time interpolation of θ_n^μ by $\theta^\mu(t) = \theta_n$ if $t \in [n\mu, n\mu + \mu)$. Then $\theta^\mu(\cdot)$ converges weakly to $\theta(\cdot)$, which is a continuous-time Markov chain generated by Q.*

Proof. Observe that the identity matrix in (2.2) can be written as

$$I = \operatorname{diag}(1, \ldots, 1) \in \mathbb{R}^{m_0 \times m_0}.$$

Each of the 1's can be thought of as a 1×1 "transition matrix." Note that under the conditions for the Markov chain θ_n, the $\operatorname{diag}(\nu^1, \ldots, \nu^l)$ defined in (4.7) becomes $I \in \mathbb{R}^{m_0 \times m_0}$, and $\operatorname{diag}(\mathbb{1}_{m_1}, \ldots, \mathbb{1}_{m_l})$ in (4.7) is also I. Moreover, the $\overline{Q}$ defined in (4.7) is now simply Q. Straightforward applications of Lemmas 4.2 and 4.3 then yield the desired results. □

4.2. Characterization of the limit. Consider the pair $(\widehat{\pi}^\mu(\cdot), \theta^\mu(\cdot))$. Then $\{\widehat{\pi}^\mu(\cdot), \theta^\mu(\cdot)\}$ is tight in $D([0,T]; \mathbb{R}^S \times \mathcal{M})$ for $T > 0$ by virtue of Proposition 4.4 and Lemma 4.1 together with the Cramér–Wold device [5, p. 48]. By virtue of Prohorov's theorem, we can extract convergent subsequences. Do that, and still index the subsequence by μ for notational simplicity. Denote the limit by $\widehat{\pi}(\cdot)$. By virtue of the Skorohod representation, $\widehat{\pi}^\mu(\cdot)$ converges to $\widehat{\pi}(\cdot)$ w.p.1, and the convergence is uniform on any compact set. We proceed to characterize the limit $\widehat{\pi}(\cdot)$. The result is stated in the following theorem.

THEOREM 4.5. *Under conditions* (M) *and* (S), $(\widehat{\pi}^\mu(\cdot), \theta^\mu(\cdot))$ *converges weakly to* $(\widehat{\pi}(\cdot), \theta(\cdot))$, *which is a solution of the following switching ODE:*

$$\frac{d}{dt}\widehat{\pi}(t) = \pi(\theta(t)) - \widehat{\pi}(t), \qquad \widehat{\pi}(0) = \widehat{\pi}_0. \tag{4.11}$$

Remark 4.6. The above switching ODE displays a very different behavior than the trajectories of systems derived from the classical ODE approach for SA. It involves a random element since $\theta(t)$ is a continuous-time Markov chain with generator Q. Because of the regime switching, the system is qualitatively different from the existing literature on SA methods. To analyze SA algorithms, the ODE methods (see [15, 16] and [17]) are now standard and widely used in various applications. The rationale is that the discrete iterations are compared with the continuous dynamics given by a limit ODE. The ODE is then used to analyze the asymptotic properties of the recursive algorithms. Dealing with tracking algorithms having time-varying features, sometimes, one may obtain a nonautonomous differential equation [16, section 8.2.6], but the systems are still purely deterministic. Unlike those mentioned above, the limit dynamic system in Theorem 4.5 is only piecewise deterministic due to the underlying Markov chain. In lieu of one ODE, we have a number of ODEs modulated by a continuous-time Markov chain. At any given instance, the Markov chain dictates which regime the system belongs to, and the corresponding system then follows one of the ODEs until the modulating Markov chain jumps into a new location, which explains the time-varying and regime switching nature of the systems under consideration.

Proof. To obtain the desired limit, we prove that the limit $(\widehat{\pi}(\cdot), \theta(\cdot))$ is the solution of the martingale problem with operator L_1 given by

$$L_1 f(x, \overline{\theta}_i) = \nabla f'(x, \overline{\theta}_i)(\pi(\overline{\theta}_i) - x) + Qf(x, \cdot)(\overline{\theta}_i) \quad \text{for each } \overline{\theta}_i \in \mathcal{M}, \tag{4.12}$$

where

$$Qf(x, \cdot)(\overline{\theta}_i) = \sum_{j \in \mathcal{M}} q_{ij} f(x, \overline{\theta}_j) = \sum_{j \neq i} q_{ij}[f(x, \overline{\theta}_j) - f(x, \overline{\theta}_i)] \quad \text{for each } \overline{\theta}_i \in \mathcal{M},$$

and for each $\overline{\theta}_i \in \mathcal{M}$, $f(\cdot, \overline{\theta}_i)$ is twice continuously differentiable with compact support. In the above, $\nabla f(x, \overline{\theta}_i)$ denotes the gradient of $f(x, \overline{\theta}_i)$ with respect to x. Using an argument as in [22, Lemma 7.18], it can be shown that the martingale problem associated with the operator L_1 has a unique solution. Thus, it remains to show that the limit $(\widehat{\pi}(\cdot), \theta(\cdot))$ is the solution of the martingale problem. To this end, we need only show that for any positive integer ℓ_0, any $t > 0$, $s > 0$, and $0 < t_j \le t$, and any bounded and continuous function $h_j(\cdot, \overline{\theta}_i)$ for each $\overline{\theta}_i \in \mathcal{M}$ with $j \le \ell_0$,

$$\begin{aligned} &\mathbf{E}\prod_{j=1}^{\ell_0} h_j(\widehat{\pi}(t_j), \theta(t_j)) \\ &\quad \times \left[f(\widehat{\pi}(t+s), \theta(t+s)) - f(\widehat{\pi}(t), \theta(t)) - \int_t^{t+s} L_1 f(\widehat{\pi}(u), \theta(u)) du \right] = 0. \end{aligned} \tag{4.13}$$

To verify (4.13), we work with the processes indexed by μ and prove that the above equation holds as $\mu \to 0$.

First by the weak convergence of $(\widehat{\pi}^\mu(\cdot), \theta^\mu(\cdot))$ to $(\widehat{\pi}(\cdot), \theta(\cdot))$ and the Skorohod representation,

$$\begin{aligned} &\lim_{\mu \to 0} \mathbf{E}\prod_{j=1}^{\ell_0} h_j(\widehat{\pi}^\mu(t_j), \theta^\mu(t_j)) \left[f(\widehat{\pi}^\mu(t+s), \theta^\mu(t+s)) - f(\widehat{\pi}^\mu(t), \theta^\mu(t)) \right] \\ &= \mathbf{E}\prod_{j=1}^{\ell_0} h_j(\widehat{\pi}(t_j), \theta(t_j)) \left[f(\widehat{\pi}(t+s), \theta(t+s)) - f(\widehat{\pi}(t), \theta(t)) \right]. \end{aligned} \tag{4.14}$$

On the other hand, choose a sequence n_μ such that $n_\mu \to \infty$ as $\mu \to 0$, but $\mu n_\mu \to 0$. Divide $[t, t+s]$ into intervals of width $\delta_\mu = \mu n_\mu$. We have

$$\begin{aligned} &\mathbf{E}\prod_{j=1}^{\ell_0} h_j(\widehat{\pi}^\mu(t_j), \theta^\mu(t_j)) \left[f(\widehat{\pi}^\mu(t+s), \theta^\mu(t+s)) - f(\widehat{\pi}^\mu(t), \theta^\mu(t)) \right] \\ &= \mathbf{E}\prod_{j=1}^{\ell_0} h_j(\widehat{\pi}^\mu(t_j), \theta^\mu(t_j)) \left[\sum_{l n_\mu = t/\mu}^{(t+s)/\mu - 1} [f(\widehat{\pi}_{l n_\mu + n_\mu}, \theta_{l n_\mu + n_\mu}) - f(\widehat{\pi}_{l n_\mu + n_\mu}, \theta_{l n_\mu})] \right. \\ &\qquad \left. + \sum_{l n_\mu = t/\mu}^{(t+s)/\mu - 1} [f(\widehat{\pi}_{l n_\mu + n_\mu}, \theta_{l n_\mu}) - f(\widehat{\pi}_{l n_\mu}, \theta_{l n_\mu})] \right]. \end{aligned} \tag{4.15}$$

By virtue of the smoothness and boundedness of $f(\cdot, \theta)$, it can be seen that

$$\begin{aligned} &\lim_{\mu \to 0} \mathbf{E}\prod_{j=1}^{\ell_0} h_j(\widehat{\pi}^\mu(t_j), \theta^\mu(t_j)) \left[\sum_{l n_\mu = t/\mu}^{(t+s)/\mu - 1} [f(\widehat{\pi}_{l n_\mu + n_\mu}, \theta_{l n_\mu + n_\mu}) - f(\widehat{\pi}_{l n_\mu + n_\mu}, \theta_{l n_\mu})] \right] \\ &= \lim_{\mu \to 0} \mathbf{E}\prod_{j=1}^{\ell_0} h_j(\widehat{\pi}^\mu(t_j), \theta^\mu(t_j)) \left[\sum_{l n_\mu = t/\mu}^{(t+s)/\mu - 1} [f(\widehat{\pi}_{l n_\mu}, \theta_{l n_\mu + n_\mu}) - f(\widehat{\pi}_{l n_\mu}, \theta_{l n_\mu})] \right]. \end{aligned} \tag{4.16}$$

Thus we need only work with the latter term. Moreover, letting $\mu \to 0$ and $l\delta_\mu = \mu l n_\mu \to u$ and using nested expectation, we can insert $\mathbf{E}_k$ and obtain

(4.17)
$$\begin{aligned}
&\mathbf{E}\prod_{j=1}^{\ell_0} h_j(\widehat{\pi}^\mu(t_j),\theta^\mu(t_j))\left[\sum_{ln_\mu=t/\mu}^{(t+s)/\mu-1}[f(\widehat{\pi}_{ln_\mu},\theta_{ln_\mu+n_\mu})-f(\widehat{\pi}_{ln_\mu},\theta_{ln_\mu})]\right]\\
&\quad=\mathbf{E}\prod_{j=1}^{\ell_0} h_j(\widehat{\pi}^\mu(t_j),\theta^\mu(t_j))\left[\sum_{ln_\mu=t/\mu}^{(t+s)/\mu-1}\sum_{j=1}^{m_0}\sum_{i=1}^{m_0}\sum_{k=ln_\mu}^{ln_\mu+n_\mu-1}[f(\widehat{\pi}_{ln_\mu},\overline{\theta}_i)\right.\\
&\qquad\left.\times P(\theta_{k+1}=\overline{\theta}_i\big|\theta_k=\overline{\theta}_j)-f(\widehat{\pi}_{ln_\mu},\overline{\theta}_j)]I_{\{\theta_k=\overline{\theta}_j)\}}\right]\\
&\quad=\mathbf{E}\prod_{j=1}^{\ell_0} h_j(\widehat{\pi}^\mu(t_j),\theta^\mu(t_j))\left[\sum_{ln_\mu=t/\mu}^{(t+s)/\mu-1}\left[\frac{\delta_\mu}{n_\mu}\sum_{j=1}^{m_0}\sum_{k=ln_\mu}^{ln_\mu+n_\mu-1}Qf(\widehat{\pi}_{ln_\mu},\cdot)(\theta_k)I_{\{\theta_k=\overline{\theta}_j\}}\right]\right]\\
&\quad\to\mathbf{E}\prod_{j=1}^{\ell_0} h_j(\widehat{\pi}(t_j),\theta(t_j))\left[\int_t^{t+s}Qf(\widehat{\pi}(u),\theta(u))du\right]\quad\text{as }\mu\to 0.
\end{aligned}$$

Since $\widehat{\pi}^\mu_{ln_\mu}$ and θ_{ln_μ} are $\mathcal{F}_{ln_\mu}$-measurable, by virtue of the continuity and boundedness of $\nabla f(\cdot,\theta)$,

$$\begin{aligned}
&\mathbf{E}\prod_{j=1}^{\ell_0} h_j(\widehat{\pi}^\mu(t_j),\theta^\mu(t_j))\sum_{ln_\mu=t/\mu}^{(t+s)/\mu-1}[f(\widehat{\pi}_{ln_\mu+n_\mu},\theta_{ln_\mu})-f(\widehat{\pi}_{ln_\mu},\theta_{ln_\mu})]\\
&\quad=\mathbf{E}\prod_{j=1}^{\ell_0} h_j(\widehat{\pi}^\mu(t_j),\theta^\mu(t_j))\sum_{ln_\mu=t/\mu}^{(t+s)/\mu-1}\left[\mu\nabla f'(\widehat{\pi}_{ln_\mu},\theta_{ln_\mu})\sum_{k=ln_\mu}^{ln_\mu+n_\mu-1}\mathbf{E}_{ln_\mu}(X_{k+1}-\widehat{\pi}_k)\right]\\
&\qquad+o(1),
\end{aligned}$$

where $o(1)\to 0$ as $\mu\to 0$. Next, consider the term

$$\lim_{\mu\to 0}\mathbf{E}\prod_{j=1}^{\ell_0} h_j(\widehat{\pi}^\mu(t_j),\theta^\mu(t_j))\left[\sum_{ln_\mu=t/\mu}^{(t+s)/\mu-1}\delta_\mu\left[\frac{1}{n_\mu}\sum_{k=ln_\mu}^{ln_\mu+n_\mu-1}\mathbf{E}_{ln_\mu}X_{k+1}\right]\right]. \tag{4.18}$$

Consider a fixed-θ process $X_k(\theta)$, which is a process with θ_k fixed at $\theta_k=\theta$ for $ln_\mu\le k\le O(1/\mu)$. Close scrutiny of the inner summation shows that

$$\frac{1}{n_\mu}\sum_{k=ln_\mu}^{ln_\mu+n_\mu-1}\mathbf{E}_{ln_\mu}X_{k+1}\ \text{ can be approximated by }\ \frac{1}{n_\mu}\sum_{k=ln_\mu}^{ln_\mu+n_\mu-1}\mathbf{E}_{ln_\mu}X_{k+1}(\theta) \tag{4.19}$$

with an approximation error going to 0, since, $E_{ln_\mu}[X_{k+1}-X_{k+1}(\theta)]=O(\varepsilon)=O(\mu)$

by use of the transition matrix (2.2). Thus we have

$$\begin{aligned}
&\frac{1}{n_\mu}\sum_{k=ln_\mu}^{ln_\mu+n_\mu-1}\mathbf{E}_{ln_\mu}X_{k+1}\\
&=\sum_{j=1}^{m_0}\frac{1}{n_\mu}\sum_{k=ln_\mu}^{ln_\mu+n_\mu-1}\mathbf{E}\left(X_{k+1}(\overline{\theta}_j)I_{\{\theta_{ln_\mu}=\overline{\theta}_j\}}\big|\theta_{ln_\mu}=\overline{\theta}_j\right)+o(1)\\
&=\sum_{j=1}^{m_0}\frac{1}{n_\mu}\sum_{k=ln_\mu}^{ln_\mu+n_\mu-1}\sum_{j_1=1}^{\mathcal{S}}e_{j_1}[A(\overline{\theta}_j)]^{k+1-ln_\mu}I_{\{\theta_{ln_\mu}=\overline{\theta}_j\}}+o(1),
\end{aligned}$$

where $o(1)\to 0$ in probability as $\mu\to 0$. Henceforth, we write $\mathbb{1}$ in lieu of $\mathbb{1}_S$. Note that for each $j=1,\ldots,S$, as $n_\mu\to\infty$ (recall that $\delta_\mu=\mu n_\mu$),

$$\frac{1}{n_\mu}\sum_{k=ln_\mu}^{ln_\mu+n_\mu-1}[A(\overline{\theta}_j)]^{k+1-ln_\mu}\to\mathbb{1}\pi'(\overline{\theta}_j).$$

Note that $I_{\{\theta_{ln_\mu}=\overline{\theta}_j\}}$ can be written as $I_{\{\theta^\mu(l\delta_\mu)=\overline{\theta}_j\}}$. As $\mu\to 0$ and $l\delta_\mu\to u$, by the weak convergence of $\theta^\mu(\cdot)$ to $\theta(\cdot)$ and the Skorohod representation, $I_{\{\theta^\mu(\mu l n_\mu)=\overline{\theta}_j\}}\to I_{\{\theta(u)=\overline{\theta}_j\}}$ w.p.1. Consequently, since $\mathbb{1}\pi'(\overline{\theta}_j)$ has identical rows,

$$\begin{aligned}
\frac{1}{n_\mu}\sum_{k=ln_\mu}^{ln_\mu+n_\mu-1}\mathbf{E}_{ln_\mu}X_{k+1}&\to\sum_{j=1}^{m_0}\pi(\overline{\theta}_j)I_{\{\theta(u)=\overline{\theta}_j\}}\\
&=\pi(\theta(u)).
\end{aligned}\tag{4.20}$$

That is, the limit does not depend on the value of initial state, a salient feature of Markov chains. As a result,

$$\begin{aligned}
&\lim_{\mu\to 0}\mathbf{E}\prod_{j=1}^{\ell_0}h_j(\widehat{\pi}^\mu(t_j),\theta^\mu(t_j))\left[\sum_{ln_\mu=t/\mu}^{(t+s)/\mu-1}\frac{1}{n_\mu}\sum_{k=ln_\mu}^{ln_\mu+n_\mu-1}\mathbf{E}_{ln_\mu}X_{k+1}\right]\\
&=\mathbf{E}\prod_{j=1}^{\ell_0}h_j(\widehat{\pi}(t_j),\theta(t_j))\left[\sum_{j=1}^{m_0}\int_t^{t+s}\pi(\overline{\theta}_j)I_{\{\theta(u)=\overline{\theta}_j\}}du\right]\\
&=\mathbf{E}\prod_{j=1}^{\ell_0}h_j(\widehat{\pi}(t_j),\theta(t_j))\left[\int_t^{t+s}\pi(\theta(u))du\right].
\end{aligned}\tag{4.21}$$

Likewise, it can be shown that, as $\mu\to 0$,

$$\begin{aligned}
&\lim_{\mu\to 0}\mathbf{E}\prod_{j=1}^{\ell_0}h_j(\widehat{\pi}^\mu(t_j),\theta^\mu(t_j))\left[\sum_{ln_\mu=t/\mu}^{(t+s)/\mu-1}\delta_\mu\frac{1}{n_\mu}\sum_{k=ln_\mu}^{ln_\mu+n_\mu-1}\widehat{\pi}_k\right]\\
&=\mathbf{E}\prod_{j=1}^{\ell_0}h_j(\widehat{\pi}(t_j),\theta(t_j))\left[\int_t^{t+s}\widehat{\pi}(u)du\right].
\end{aligned}\tag{4.22}$$

Combining (4.14), (4.17), (4.21), and (4.22), the desired result follows. □

5. Switching diffusion limit. By Theorem 3.1, $\left\{\frac{\hat{\pi}_n - \mathbf{E}\pi(\theta_n)}{\sqrt{\mu}}\right\}$ is tight for $n \geq n_0$, for some positive integer n_0. In an effort to evaluate the rate of variation of the tracking error sequence, we define a scaled sequence of the tracking errors $\{v_n\}$ and its continuous-time interpolation $v^\mu(\cdot)$ by

$$v_n = \frac{\widehat{\pi}_n - \mathbf{E}\{\pi(\theta_n)\}}{\sqrt{\mu}}, \quad n \geq n_0, \quad v^\mu(t) = v_n \ \text{ for } \ t \in [n\mu, n\mu + \mu). \tag{5.1}$$

We will derive a limit process for $v^\mu(\cdot)$ as $\mu \to 0$. Similarly to the rate of convergence study when θ is a fixed parameter (see [16, Chapter 10]), the scaling factor $\sqrt{\mu}$, together with the asymptotic covariance of the limit process, gives us a "rate of convergence" result.

Note that from Proposition 4.4

$$\mathbf{E}\{\pi(\theta_n)\} = \bar{\pi}(\mu n) + O(\mu + \exp(-k_0 n)), \quad \text{where } \bar{\pi}(\mu n) \stackrel{\text{def}}{=} \sum_{i=1}^{S} z^i(\mu n)\pi(\overline{\theta}_i), \tag{5.2}$$

where $z^i(t)$ is the ith component of $z(t)$ given in Proposition 4.4. By (M), $\{\theta_n\}$ is a Markov chain with stationary (time-invariant) transition probabilities, so in view of (2.3),

$$v_{n+1} = v_n - \mu v_n + \sqrt{\mu}(X_{n+1} - \mathbf{E}\{\pi(\theta_n)\}) + \frac{\mathbf{E}[\pi(\theta_n) - \pi(\theta_{n+1})]}{\sqrt{\mu}}. \tag{5.3}$$

Our task in what follows is to figure out the asymptotic properties of $v^\mu(\cdot)$. We aim to show that the limit is a switching diffusion using a martingale problem formulation.

5.1. Truncation and tightness. Owing to the definition (5.1), $\{v_n\}$ is not a priori bounded. A convenient way to circumvent this difficulty is to use a truncation device [16]. Let $N > 0$ be a fixed but otherwise arbitrary real number, $S_N(z) = \{z \in \mathbb{R}^S : |z| \leq N\}$ be the sphere with radius N, and $\tau^N(z)$ be a smooth function satisfying

$$\tau^N(z) = \begin{cases} 1 & \text{if } |z| \leq N, \\ 0 & \text{if } |z| \geq N + 1. \end{cases}$$

Note that $\tau^N(z)$ is "smoothly" connected between the sphere S_N and S_{N+1}. Now define

$$v_{n+1}^N = v_n^N - \mu v_n^N \tau^N(v_n^N) + \sqrt{\mu}(X_{n+1} - \mathbf{E}\pi(\theta_n)) + \frac{\mathbf{E}[\pi(\theta_n) - \pi(\theta_{n+1})]}{\sqrt{\mu}}, \tag{5.4}$$

and define $v^{\mu,N}(\cdot)$ to be the continuous-time interpolation of v_n^N. It then follows that

$$\lim_{k_0 \to \infty} \limsup_{\mu \to 0} P\left(\sup_{0 \leq t \leq T} |v^{\mu,N}(t)| \geq k_0\right) = 0 \ \text{ for each } \ T < \infty$$

and that $v^{\mu,N}(\cdot)$ is a process that is equal to $v^\mu(\cdot)$ up until the first exit from S_N, and hence an N-truncation process of $v^\mu(\cdot)$ [16, p. 284]. To proceed, we work with $\{v^{\mu,N}(\cdot)\}$ and derive its tightness and weak convergence first. Finally, we let $N \to \infty$ to conclude the proof.

LEMMA 5.1. *Under conditions* (M) *and* (S), $\{v^{\mu,N}(\cdot)\}$ *is tight in* $D(S[0,T];\mathbb{R}^S)$, *and the process* $\{v^{\mu,N}(\cdot),\theta^{\mu}(\cdot)\}$ *is tight in* $D([0,T];\mathbb{R}^S\times\mathcal{M})$.

Proof. In fact, only the first assertion needs to be verified. In view of (5.4), for any $\delta>0$ and $t,s\geq 0$ with $s\leq\delta$,

(5.5)
$$v^{\mu,N}(t+s)-v^{\mu,N}(t)=-\mu\sum_{k=t/\mu}^{(t+s)/\mu-1}v_k^N\tau^N(v_k^N)+\sqrt{\mu}\sum_{k=t/\mu}^{(t+s)/\mu-1}(X_{k+1}-\mathbf{E}\pi(\theta_k))$$
$$+\frac{1}{\sqrt{\mu}}\sum_{k=t/\mu}^{(t+s)/\mu-1}\mathbf{E}(\pi(\theta_k)-\pi(\theta_{k+1})).$$

Owing to the N-truncation used,

$$\left|\mu\sum_{k=t/\mu}^{(t+s)/\mu-1}v_k^N\tau^N(v_k^N)\right|\leq Ks,$$

and as a result,

$$\lim_{\delta\to 0}\limsup_{\mu\to 0}\mathbf{E}\left|\mu\sum_{k=t/\mu}^{(t+s)/\mu-1}v_k^N\tau^N(v_k^N)\right|^2=0. \tag{5.6}$$

Next, by virtue of (M), the irreducibility of the conditional Markov chain $\{X_n\}$ implies that it is ϕ-mixing with exponential mixing rate [5, p. 167], $\mathbf{E}\pi(\theta_k)-\mathbf{E}X_{k+1}\to 0$ exponentially fast, and consequently

$$\mathbf{E}\left|\mu\sum_{k=t/\mu}^{(t+s)/\mu-1}(X_{k+1}-\mathbf{E}\pi(\theta_k))\right|^2$$
$$=\mathbf{E}\left|\mu\sum_{k=t/\mu}^{(t+s)/\mu-1}[(X_{k+1}-\mathbf{E}X_{k+1})-(\mathbf{E}\pi(\theta_k)-\mathbf{E}X_{k+1})]\right|^2=O(s).$$

This yields that

$$\lim_{\delta\to 0}\limsup_{\mu\to 0}\mathbf{E}\left|\mu\sum_{k=t/\mu}^{(t+s)/\mu-1}(X_{k+1}-\mathbf{E}\pi(\theta_k))\right|^2=0. \tag{5.7}$$

In addition,

$$\frac{1}{\sqrt{\mu}}\sum_{k=t/\mu}^{(t+s)/\mu-1}\mathbf{E}(\pi(\theta_k)-\pi(\theta_{k+1}))=\frac{1}{\sqrt{\mu}}[\mathbf{E}\pi(\theta_{t/\mu})-\mathbf{E}\pi(\theta_{(t+s)/\mu})]=O(\sqrt{\mu}). \tag{5.8}$$

Combining (5.5)–(5.8), we have

$$\lim_{\delta\to 0}\limsup_{\mu\to 0}\mathbf{E}\left|v^{\mu,N}(t+s)-v^{\mu,N}(t)\right|^2=0,$$

and hence the criterion [14, p. 47] implies that $\{v^{\mu,N}(\cdot)\}$ is tight. □

5.2. Representation of covariance. The main results to follow, Lemma 5.4 and Corollary 5.5 for the diffusion limit in section 5.3, require representation of the covariance of the conditional Markov chain $\{X_k\}$. This is again worked out via the use of fixed-θ process $X_k(\theta)$ similar in spirit to (4.19). For any integer $m \geq 0$, for $m \leq k \leq O(1/\mu)$, with θ_k fixed at θ, $X_{k+1}(\theta)$ is a finite-state Markov chain with 1-step irreducible transition matrix $A(\theta)$ and stationary distribution $\pi(\theta)$. Thus [5, p. 167] implies that $\{X_{k+1}(\theta) - \mathbf{E}X_{k+1}(\theta)\}$ is a ϕ-mixing sequence with zero mean and exponential mixing rate, and hence it is strongly ergodic. Similarly to (4.19), $X_{k+1} - \mathbf{E}X_{k+1}$ can be approximated by a fixed θ process $X_{k+1}(\theta) - \mathbf{E}X_{k+1}(\theta)$. Taking $n = n_\mu \leq O(1/\mu)$ as $\mu \to 0$, $n \to \infty$, and

(5.9)
$$\lim_{\mu\to 0} \frac{1}{n} \sum_{k_1=m}^{n+m-1} \sum_{k=m}^{n+m-1} (X_{k+1}(\theta) - \mathbf{E}X_{k+1}(\theta))(X_{k_1+1}(\theta) - \mathbf{E}X_{k_1+1}(\theta))' = \Sigma(\theta) \quad \text{w.p.1},$$

where $\Sigma(\theta)$ is an $S \times S$ deterministic matrix and

(5.10)
$$\lim_{\mu\to 0} \frac{1}{n} \sum_{k_1=m}^{n+m-1} \sum_{k=m}^{n+m-1} \mathbf{E}\{(X_{k+1}(\theta) - \mathbf{E}X_{k+1}(\theta))(X_{k_1+1}(\theta) - \mathbf{E}X_{k_1+1}(\theta))'\} = \Sigma(\theta).$$

Note that (5.9) is a consequence of ϕ-mixing and strong ergodicity, and (5.10) follows from (5.9) by means of the dominated convergence theorem. Clearly, $\Sigma(\theta)$ is symmetric and nonnegative definite. The following lemma gives an explicit formula for $\Sigma(\theta)$ in terms of $\pi(\theta)$ and $A(\theta)$ and is useful for computational purposes.

LEMMA 5.2. *The covariance matrix* $\Sigma(\theta)$ *in* (5.10) *can be explicitly computed as*

$$\Sigma(\theta) = Z'(\theta)D(\theta) + D(\theta)Z(\theta) - D(\theta) - \pi(\theta)\pi'(\theta), \tag{5.11}$$

where $D(\theta) = \operatorname{diag}(\pi_1(\theta), \ldots, \pi_{m_0}(\theta))$ *and* $Z(\theta)$ *is given by*

$$Z(\theta) = (I - A(\theta) + \mathbb{1}\pi'(\theta))^{-1}.$$

Remark 5.3. The $Z(\theta)$ is termed the "fundamental" matrix [6, p. 226]. As shown in the aforementioned reference, because $A(\theta)$ is irreducible, $Z(\theta)$ is nonsingular.

Proof. Note that $\Sigma(\theta) = \lim_{\mu\to 0} \Sigma^\mu(\theta)$, where $\Sigma^\mu(\theta)$ can be expressed in terms of $\pi(\theta)$ as

$$\begin{aligned} \Sigma^\mu(\theta) &= \mathbf{E}\xi_0(\theta)\xi_0'(\theta) + \sum_{k=-\lfloor 1/\mu \rfloor}^{-1} \mathbf{E}\xi_k(\theta)\xi_0'(\theta) + \sum_{k=1}^{\lfloor 1/\mu \rfloor} \mathbf{E}\xi_k(\theta)\xi_0'(\theta), \\ \xi_k(\theta) &\stackrel{\text{def}}{=} X_k(\theta) - \pi(\theta), \end{aligned} \tag{5.12}$$

and $\{X_k(\theta)\}$ is a fixed-θ Markov chain with $\theta_{-\lfloor 1/\mu \rfloor} = \theta$ and $\theta_k = \theta$ for all integer $k \leq O(1/\mu)$. Consider the terms in the above equation. For $0 < k \leq O(1/\mu)$,

$$\mathbf{E}\xi_k(\theta)\xi_0'(\theta) = \mathbf{E}X_k(\theta)X_0'(\theta) - \pi(\theta)\pi'(\theta) = (A^k(\theta))'\mathbf{E}\{X_0(\theta)X_0'(\theta)\} - \pi(\theta)\pi'(\theta).$$

Since $\{X_k(\theta)\}$ is geometrically ergodic and starts at $k = -\lfloor 1/\mu \rfloor$, $X_0(\theta)$ has distribution $\pi(\theta)$, so $\mathbf{E}\{X_0(\theta)X_0'(\theta)\} = D(\theta)$. Then using the fact that $\pi(\theta) = D(\theta)\mathbb{1}$, it

follows that $\mathbf{E}\xi_k(\theta)\xi_0'(\theta) = (A^k(\theta) - \mathbb{1}\pi'(\theta))'D(\theta)$. Thus it is easily checked that

$$\lim_{\mu\to 0}\sum_{k=1}^{\lfloor 1/\mu\rfloor}\mathbf{E}\xi_k(\theta)\xi_0'(\theta) = \lim_{\mu\to 0}\sum_{k=1}^{\lfloor 1/\mu\rfloor}\left(A^k(\theta) - \mathbb{1}\pi'(\theta)\right)' D(\theta) = (Z(\theta) - I)'D(\theta); \tag{5.13}$$

see also [6, p. 226], where it was shown that $\lim_{\mu\to 0}\sum_{k=1}^{\lfloor 1/\mu\rfloor}(A^k(\theta)(\theta) - \mathbb{1}\pi'(\theta)) = Z(\theta) - I$. Similarly,

$$\begin{aligned}&\lim_{\mu\to 0}\sum_{k=-\lfloor 1/\mu\rfloor}^{-1}\mathbf{E}\xi_k(\theta)\xi_0'(\theta) = D(\theta)(Z(\theta) - I),\\ &\mathbf{E}\xi_0(\theta)\xi_0'(\theta) = D(\theta) - \pi(\theta)\pi'(\theta).\end{aligned} \tag{5.14}$$

The expression (5.12) and the limits in (5.13) and (5.14) yield (5.11). □

5.3. Weak limit via a martingale problem solution. To obtain the desired weak convergence result, we work with the pair $(v^{\mu,N}(\cdot), \theta^\mu(\cdot))$. By virtue of the tightness and Prohorov's theorem, we can extract a weakly convergent subsequence (still denoted by $(v^{\mu,N}(\cdot), \theta^\mu(\cdot))$ for simplicity) with limit $(v^N(\cdot), \theta(\cdot))$. We will show that the limit is a switching diffusion.

To proceed with the diffusion approximation, similarly as in the proof of Theorem 4.5, we will use the martingale problem formulation to derive the desired result. For $v \in \mathbb{R}^S$, $\theta \in \mathcal{M}$, and any twice continuously differentiable function $f(\cdot,\theta)$ with compact support, consider the operator $\mathcal{L}$ defined by

$$\mathcal{L}f(v,\theta) = -\nabla f'(v,\theta)v + \frac{1}{2}\mathrm{tr}[\nabla^2 f(v,\theta)\Sigma(\theta)] + Qf(v,\cdot)(\theta), \tag{5.15}$$

where $\Sigma(\theta)$ is given by (5.10) and $\nabla^2 f(v,\theta)$ denotes $(\partial^2/\partial v_i\partial v_j)f(v,\theta)$, the mixed second-order partial derivatives. For any positive integer ℓ_0, any $t > 0$, $s > 0$, any $0 < t_j \le t$ with $j \le \ell_0$, and any bounded and continuous function $h_j(\cdot,\theta)$ for each $\theta \in \mathcal{M}$, we aim to derive an equation similar to (4.13) with the operator L_1 replaced by $\mathcal{L}$. As in the proof of Theorem 4.5, we work with the sequence indexed by μ. Choose n_μ such that $n_\mu \to \infty$ but $\delta_\mu = \mu n_\mu \to 0$. The tightness of $\{v^{\mu,N}(\cdot), \theta^\mu(\cdot)\}$ and the Skorohod representation yield that (4.14)–(4.16) hold with $\widehat{\pi}^\mu(\cdot)$ and $\widehat{\pi}(\cdot)$ replaced by $v^{\mu,N}(\cdot)$ and $v^N(\cdot)$, respectively.

LEMMA 5.4. *Assume the conditions of Lemma* 5.1 *and that* $(v^{\mu,N}(0), \theta^\mu(0))$ *converges weakly to* $(v^N(0), \theta(0))$. *Then* $(v^{\mu,N}(\cdot), \theta^\mu(\cdot))$ *converges weakly to* $(v^N(\cdot), \theta(\cdot))$, *which is a solution of the martingale problem with operator* $\mathcal{L}^N$ *given by*

$$\mathcal{L}^N f(v,\theta) = -\nabla f'(v^N,\theta)v^N\tau^N(v^N) + \frac{1}{2}\mathrm{tr}[\nabla^2 f(v^N,\theta)\Sigma(\theta)] + Qf(v^N,\cdot)(\theta), \tag{5.16}$$

or equivalently $v^N(\cdot)$ *satisfies*

$$dv^N(t) = -v^N(t)\tau^N(v^N(t)) + \Sigma^{1/2}(\theta(t))dw, \tag{5.17}$$

where $w(\cdot)$ *is a standard* S*-dimensional Brownian motion and* $\Sigma(\theta)$ *is given by* (5.10).

Proof. In view of (5.8), the term $\sum_{k=t/\mu}^{(t+s)/\mu-1}[\mathbf{E}\pi(\theta_k) - \mathbf{E}\pi(\theta_{k+1})]/\sqrt{\mu} = O(\sqrt{\mu})$

can be ignored in the characterization of the limit process. Moreover,

$$\begin{aligned}&\sqrt{\mu}\sum_{k=t/\mu}^{(t+s)/\mu-1}[X_{k+1}-\mathbf{E}\pi(\theta_k)]\\&=\sqrt{\mu}\sum_{k=t/\mu}^{(t+s)/\mu-1}(X_{k+1}-\mathbf{E}X_{k+1})+\sqrt{\mu}\sum_{k=t/\mu}^{(t+s)/\mu-1}(\mathbf{E}X_{k+1}-\mathbf{E}\pi(\theta_k)).\end{aligned}$$

Since $\mathbf{E}X_{k+1}-\mathbf{E}\pi(\theta_k)\to 0$ exponentially fast owing to the elementary properties of a Markov chain, the last term above is $o(1)$ that goes to 0 as $\mu\to 0$. Thus,

(5.18)
$$v^{\mu,N}(t+s)-v^{\mu,N}(t)=-\mu\sum_{k=t/\mu}^{(t+s)/\mu-1}v_k^N\tau^N(v_k^N)+\sqrt{\mu}\sum_{k=t/\mu}^{(t+s)/\mu-1}(X_{k+1}-\mathbf{E}X_{k+1})+o(1).$$

Similarly to the argument in the proof of Theorem 4.5,

(5.19)
$$\begin{aligned}&\lim_{\mu\to 0}\mathbf{E}\prod_{j=1}^{\ell_0}h_j(v^{\mu,N}(t_j),\theta^\mu(t_j))\left[\sum_{ln_\mu=t/\mu}^{(t+s)/\mu-1}[f(v^N_{ln_\mu},\theta_{ln_\mu+n_\mu})-f(v^N_{ln_\mu},\theta_{ln_\mu})]\right]\\&=\mathbf{E}\prod_{j=1}^{\ell_0}h_j(v^N(t_j),\theta(t_j))\left[\int_t^{t+s}Qf(v^N(u),\theta(u))du\right].\end{aligned}$$

In addition,

(5.20)
$$\begin{aligned}&\lim_{\mu\to 0}\mathbf{E}\prod_{j=1}^{\ell_0}h_j(v^{\mu,N}(t_j),\theta^\mu(t_j))\left[-\sum_{ln_\mu=t/\mu}^{(t+s)/\mu-1}\frac{\delta_\mu}{n_\mu}\sum_{k=ln_\mu}^{ln_\mu+n_\mu-1}\nabla f'(v^N_{ln_\mu},\theta_{ln_\mu})v_k^N\tau^N(v_k^N)\right]\\&=\lim_{\mu\to 0}\mathbf{E}\prod_{j=1}^{\ell_0}h_j(v^{\mu,N}(t_j),\theta^\mu(t_j))\left[-\sum_{ln_\mu=t/\mu}^{(t+s)/\mu-1}\delta_\mu\nabla f'(v^N_{ln_\mu},\theta_{ln_\mu})v^N_{ln_\mu}\tau^N(v^N_{ln_\mu})\right]\\&=\mathbf{E}\prod_{j=1}^{\ell_0}h_j(v^N(t_j),\theta(t_j))\left[-\int_t^{t+s}\nabla f'(v^N(u),\theta(u))v^N(u)\tau^N(v^N(u))du\right].\end{aligned}$$

Next we note that

(5.21)
$$\begin{aligned}&\left|\mathbf{E}\prod_{j=1}^{\ell_0}h_j(v^{\mu,N}(t_j),\theta^\mu(t_j))\left[\sqrt{\mu}\sum_{ln_\mu=t/\mu}^{(t+s)/\mu-1}\nabla f'(v^N_{ln_\mu},\theta_{ln_\mu})\sum_{k=ln_\mu}^{ln_\mu+n_\mu-1}[X_{k+1}-\mathbf{E}X_{k+1}]\right]\right|\\&\le\left|\mathbf{E}\prod_{j=1}^{\ell_0}h_j(v^{\mu,N}(t_j),\theta^\mu(t_j))\left[\sqrt{\mu}\sum_{ln_\mu=t/\mu}^{(t+s)/\mu-1}|\nabla f'(v^N_{ln_\mu},\theta_{ln_\mu})|\right.\right.\\&\qquad\left.\left.\times\sum_{k=ln_\mu}^{ln_\mu+n_\mu-1}|\mathbf{E}_{ln_\mu}[X_{k+1}-\mathbf{E}X_{k+1}]|\right]\right|\\&\to 0\ \text{ as }\ \mu\to 0\end{aligned}$$

owing to the mixing property.

Finally, define

$$g_{ln_\mu} g'_{ln_\mu} = \frac{1}{n_\mu} \sum_{k=ln_\mu}^{ln_\mu+n_\mu-1} \sum_{k_1=ln_\mu}^{ln_\mu+n_\mu-1} \mathbf{E}_{ln_\mu}[X_{k+1} - \mathbf{E}X_{k+1}][X_{k_1+1} - \mathbf{E}X_{k_1+1}]'.$$

It follows that

$$\begin{aligned}
&\mathbf{E}\prod_{j=1}^{\ell_0} h_j(v^{\mu,N}(t_j), \theta^\mu(t_j)) \left[\sum_{ln_\mu=t/\mu}^{(t+s)/\mu-1} \mathrm{tr}[\nabla^2 f(v^N_{ln_\mu}, \theta_{ln_\mu})(v^N_{ln_\mu+n_\mu} - v^N_{ln_\mu}) \right. \\
&\qquad \left. \times (v^N_{ln_\mu+n_\mu} - v^N_{ln_\mu})'] \right] \\
&= \mathbf{E}\prod_{j=1}^{\ell_0} h_j(v^{\mu,N}(t_j), \theta^\mu(t_j)) \left[\sum_{j=1}^{m_0} \sum_{ln_\mu=t/\mu}^{(t+s)/\mu-1} \mathrm{tr}[\nabla^2 f(v^N_{ln_\mu}, \theta_{ln_\mu})(v^N_{ln_\mu+n_\mu} - v^N_{ln_\mu}) \right. \\
&\qquad \left. \times (v^N_{ln_\mu+n_\mu} - v^N_{ln_\mu})'] I_{\{\theta_{ln_\mu}=\overline{\theta}_j\}} \right] \\
&= \mathbf{E}\prod_{j=1}^{\ell_0} h_j(v^{\mu,N}(t_j), \theta^\mu(t_j)) \left[\sum_{j=1}^{m_0} \sum_{ln_\mu=t/\mu}^{(t+s)/\mu-1} \delta_\mu \mathrm{tr}[\nabla^2 f(v^N_{ln_\mu}, \theta_{ln_\mu}) \mathbf{E}_{ln_\mu} g_{ln_\mu} g'_{ln_\mu}] \right. \\
&\qquad \left. \times I_{\{\theta_{ln_\mu}=\overline{\theta}_j\}} \right] + \rho_\mu,
\end{aligned}$$

where $\rho_\mu \to 0$ as $\mu \to 0$. Since it is conditioned on $\theta_{ln_\mu} = \overline{\theta}_j$, $X_{k+1} - \mathbf{E}X_{k+1}$ can be approximated by a fixed-$\overline{\theta}_j$ process $X_{k+1}(\overline{\theta}_j) - \mathbf{E}X_{k+1}(\overline{\theta}_j)$, and since $X_{k+1}(\overline{\theta}_j) - \mathbf{E}X_{k+1}(\overline{\theta}_j)$ is a finite-state Markov chain with irreducible transition matrix $A(\overline{\theta}_j)$, it is ϕ-mixing, and the argument in (5.10) implies that for each $\overline{\theta}_j \in \mathcal{M}$ with $j = 1, \ldots, m_0$,

(5.22)
$$\begin{aligned}
&\frac{1}{n_\mu} \sum_{k=ln_\mu}^{ln_\mu+n_\mu-1} \sum_{k_1=ln_\mu}^{ln_\mu+n_\mu-1} \mathbf{E}_{ln_\mu}(X_{k+1}(\overline{\theta}_j) - \mathbf{E}X_{k+1}(\overline{\theta}_j))(X_{k_1+1}(\overline{\theta}_j) - \mathbf{E}X_{k_1+1}(\overline{\theta}_j))' \\
&\to \Sigma(\overline{\theta}_j) \quad \text{w.p.1 as} \quad \mu \to 0,
\end{aligned}$$

where $\Sigma(\theta)$ is defined in (5.10). By virtue of Lemma 4.3, $\theta^\mu(\cdot)$ converges weakly to $\theta(\cdot)$. As a result, by Skorohod representation, sending $\mu \to 0$ and $l\delta_\mu \to u$ leads to $\theta^\mu(\mu l n_\mu)$ converging to $\theta(u)$ w.p.1. In addition, $I_{\{\theta^\mu(l\delta_\mu)=\overline{\theta}_j\}} \to I_{\{\theta(u)=\overline{\theta}_j\}}$ w.p.1. It follows that

$$
\begin{aligned}
&\mathbf{E}\prod_{j=1}^{\ell_0} h_j(v^{\mu,N}(t_j),\theta^\mu(t_j))\left[\sum_{ln_\mu=t/\mu}^{(t+s)/\mu-1}\operatorname{tr}\big[\nabla^2 f(v^N_{ln_\mu},\theta_{ln_\mu})(v^N_{ln_\mu+n_\mu}-v^N_{ln_\mu})\right.\\
&\qquad\left.\times(v^N_{ln_\mu+n_\mu}-v^N_{ln_\mu})'\big]\right]\\
&\to \mathbf{E}\prod_{j=1}^{\ell_0} h_j(v^N(t_j),\theta(t_j))\left[\int_t^{t+s}\sum_{j=1}^{m_0}\operatorname{tr}\big[\nabla^2 f(v^N(u),\overline{\theta}_j)\Sigma(\overline{\theta}_j)\big]I_{\{\theta(u)=\overline{\theta}_j\}}du\right]\\
&= \mathbf{E}\prod_{j=1}^{\ell_0} h_j(v^N(t_j),\theta(t_j))\left[\int_t^{t+s}\operatorname{tr}\big[\nabla^2 f(v^N(u),\theta(u))\Sigma(\theta(u))\big]du\right].
\end{aligned}
\tag{5.23}
$$

In view of (5.19)–(5.23), the desired result follows. □

COROLLARY 5.5. *Under the conditions of Lemma* 5.4*, the untruncated process* $(v^\mu(\cdot),\theta^\mu(\cdot))$ *converges weakly to* $(v(\cdot),\theta(\cdot))$ *satisfying the switching diffusion equation*

$$
dv(t) = -v(t)dt + \Sigma^{1/2}(\theta(t))dw, \tag{5.24}
$$

where $w(\cdot)$ *is a standard Brownian motion and* $\Sigma(\theta)$ *is given by* (5.10).

Proof. The uniqueness of the associated martingale problem can be proved similarly to that of [22, Lemma 7.18]. The rest of the proof follows from a similar argument as in [16, Step 4, p. 285]. □

Combining Lemma 5.1, Lemma 5.4, and Corollary 5.5, we have proved the following result.

THEOREM 5.6. *Assume conditions* (M) *and* (S) *and that* $(v^\mu(0),\theta^\mu(0))$ *converges weakly to* $(v(0),\theta(0))$*. Then* $(v^\mu(\cdot),\theta^\mu(\cdot))$ *converges weakly to* $(v(\cdot),\theta(\cdot))$*, which is the solution of the martingale problem with operator defined by* (5.15)*, or equivalently, it is the solution of the system of diffusions with regime switching* (5.24).

Remark 5.7. The reason for obtaining a result such as Theorem 5.6 stems from the motivation for figuring out rates of convergence. If θ were a fixed parameter, we would obtain a diffusion limit as those in [16, Chapter 10]. As a consequence, the sequence v_n will be approximately normal. Now, our motivation is still for getting the rate of convergence. However, Theorem 5.6 reveals that v_n is an asymptotically Gaussian mixture. The mixture results from the time-varying parameter.

Remark 5.8. *Occupation measure for hidden Markov model.* The development thus far concerns recursive estimation of the occupation measure $\pi(\theta_n)$, given exact measurements of the conditional Markov sequence $\{X_n\}$. The above results can be extended to the hidden Markov model (HMM) case where the process $\{X_n\}$ is observed in noise as $\{Y_n\}$, where

$$
Y_n = X_n + \zeta_n. \tag{5.25}
$$

Assume that $\{\zeta_n\}$ satisfies the standard noise assumptions of an HMM [8, 13], i.e., it is a mutually independent and identically distributed (i.i.d.) noise process independent of X_n and θ_n. Then, given $\{Y_n\}$, to recursively estimate $\pi(\theta_n)$, the following modified version of the LMS algorithm (2.3) can be used. Replace X_{n+1} in algorithm (2.3) by Y_{n+1}. The mean square error analysis, switching ODE, and switching diffusion results of the previous sections carry over to this HMM case. More precisely, the following theorem holds.

THEOREM 5.9. *Consider the LMS algorithm* (2.3)*, where* X_{n+1} *is replaced by the HMM observation* Y_{n+1} *defined in* (5.25)*. Assume that the conditions of Theorem* 5.6 *hold, that* $\{\zeta_n\}$ *is a sequence of i.i.d. random variables with zero mean and* $E|\zeta_1|^2 < \infty$*, and that* $\{\zeta_n\}$ *is independent of* $\{X_n\}$ *and* $\{\theta_n\}$*. Then the conclusions of Theorems* 3.1*,* 4.5*, and* 5.6 *continue to hold.*

6. Application—Adaptive discrete stochastic optimization. In this section we apply the results developed in sections 3–5 to analyzing the tracking performance of an adaptive version of a discrete stochastic optimization algorithm proposed by Andradóttir [2]. Throughout this section we assume that the $\mathcal{M}$ in (2.1) is $\mathcal{M} = \mathcal{S} = \{e_1, \ldots, e_S\}$, where e_i denotes the standard unit vector. In what follows, $\mathcal{M}$ denotes the set of candidate values from which the time-varying global minimizer is chosen at each time instant (according to a slow Markov chain). $\mathcal{S}$ is the set of candidate solutions for the discrete optimization. Because we assume $\mathcal{M} = \mathcal{S}$, we do not use the notation $\mathcal{S}$ in this section. Note that the assumption that $\mathcal{M} = \mathcal{S}$ is made purely for notational convenience. Indeed, the set $\mathcal{M}$ of possible values from which the time-varying optimum is drawn can be any subset of $\mathcal{S}$.

6.1. Static discrete stochastic optimization. Consider the following discrete stochastic optimization problem:

$$\min_{\overline{\theta} \in \mathcal{M}} \mathbf{E}\{c_n(\overline{\theta})\}, \tag{6.1}$$

where for each fixed $\overline{\theta} \in \mathcal{M}$, $\{c_n(\overline{\theta})\}$ is a sequence of i.i.d. random variables with finite variance. Let $\mathcal{K} \subset \mathcal{M}$ denote the set of global minimizers for (6.1). The problem is static in the sense that the set $\mathcal{K}$ of global minima does not evolve with time.

When the expected value $\mathbf{E}\{c_n(\overline{\theta})\}$ can be evaluated analytically, (6.1) may be solved using standard integer programming techniques. A more interesting and important case motivated by applications in operations research [20] and wireless communication networks [11] is when $\mathbf{E}\{c_n(\overline{\theta})\}$ cannot be evaluated analytically and only $c_n(\overline{\theta})$ can be measured via simulation.

If a closed form solution of $\mathbf{E}\{c_n(\overline{\theta})\}$ cannot be obtained, a brute force method [18, Chapter 5.3] of solving the discrete stochastic optimization problem involves an exhaustive enumeration. It proceeds as follows: For each possible $\overline{\theta} \in \mathcal{M}$, compute the empirical average

$$\widehat{c}_N(\overline{\theta}) = \frac{1}{N} \sum_{i=1}^{N} c_i(\overline{\theta})$$

via simulation for large N, and pick out $\widehat{\theta} = \arg\min_{\overline{\theta} \in \mathcal{M}} \widehat{c}_N(\overline{\theta})$.

Since for any fixed $\overline{\theta} \in \mathcal{M}$, $\{c_n(\overline{\theta})\}$ is an i.i.d. sequence of random variables with finite variance, by virtue of Kolmogorov's strong law of large numbers, $\widehat{c}_N(\overline{\theta}) \to \mathbf{E}\{c_1(\overline{\theta})\}$ w.p.1 as $N \to \infty$. This and the finiteness of $\mathcal{M}$ imply that, as $N \to \infty$,

$$\arg\min_{\overline{\theta} \in \mathcal{M}} \widehat{c}_N(\overline{\theta}) \to \arg\min_{\overline{\theta} \in \mathcal{M}} \mathbf{E}\left\{c_1(\overline{\theta})\right\} \text{ w.p.1.} \tag{6.2}$$

In principle, the above brute force simulation method can solve the discrete stochastic optimization problem (6.1) for large N and the estimate is *consistent*, i.e., (6.2) holds. However, the method is highly inefficient since $\widehat{c}_N(\overline{\theta})$ needs to be evaluated for each $\overline{\theta} \in \mathcal{M}$. The evaluations of $\widehat{c}_N(\overline{\theta})$ for $\overline{\theta} \notin \mathcal{K}$ are wasted because they contribute nothing to the estimation of $\widehat{c}_N(\theta)$, $\theta \in \mathcal{K}$.

The idea of discrete stochastic optimization in [3] is to design an algorithm that is both *consistent* and *attracted* to the minimum. That is, the algorithm should spend more time obtaining observations $c_n(\overline{\theta})$ in areas of the state space $\mathcal{M}$ near the minimizer θ, and less so in other areas. Thus in discrete stochastic optimization the aim is to devise an *efficient* [18, Chapter 5.3] adaptive search (sampling plan), which allows us to find the maximizer with as few samples as possible by not making unnecessary observations at nonpromising values of $\overline{\theta}$.

In the papers [2] and [3], Andradóttir has proposed random search–based discrete stochastic optimization algorithms for computing the global minimizer in (6.1). In this subsection a brief outline of the assumptions and algorithm in [2] is given. Sections 6.2 and 6.3 analyze the performance of an adaptive version of the algorithm for tracking a time-varying minimum. In [2], the following stochastic ordering assumption was used.

(O) For each $e_i, e_j \in \mathcal{M}$, there exists some random variable Y^{e_i,e_j} such that for all $e_i \in \mathcal{K}$, $e_j \in \mathcal{K}$, and $e_l \in \mathcal{M}$, $l \neq i, j$,

$$\begin{aligned} &P(Y^{e_j,e_i} > 0) \geq P(Y^{e_i,e_j} > 0), \qquad P(Y^{e_l,e_i} > 0) \geq P(Y^{e_l,e_j} > 0), \\ &P(Y^{e_i,e_l} \leq 0) \geq P(Y^{e_j,e_l} \leq 0). \end{aligned} \tag{6.3}$$

Roughly speaking, this assumption ensures that the algorithm is more likely to jump towards a global minimum than away from it; see [2] for details. Some examples on how to choose Y^{e_i,e_j} are given in [2]. For example, suppose $c_n(\overline{\theta}) = \overline{\theta} + w_n(\overline{\theta})$ in (6.1) for each $\overline{\theta} \in \mathcal{M}$, where $\{w_n(\overline{\theta})\}$ has a symmetric continuous probability density function with zero mean. In this case simply choose $Y^{e_i,e_j} = c_n(e_i) - c_n(e_j)$. It is easily established that such a Y^{e_i,e_j} satisfies assumption (O). In [10] a stochastic comparison algorithm is presented for this example.

The static discrete stochastic optimization algorithm presented in [2] is as follows.

ALGORITHM 1 (static discrete stochastic optimization algorithm).

a. `Step 0:` (Initialization) At time $n = 0$, select starting point $X_0 \in \mathcal{M}$. Set $\widehat{\pi}_0 = X_0$, and select $\widehat{\theta}_0^* = X_0$.

b. `Step 1:` (Random search) At time n, sample $\widetilde{X}_n$ with uniform distribution from $\mathcal{M} - \{X_n\}$.

c. `Step 2:` (Evaluation and acceptance) Generate observation $Y^{X_n,\widetilde{X}_n}$. If $Y^{X_n,\widetilde{X}_n} > 0$, set $X_{n+1} = \widetilde{X}_n$; else, set $X_{n+1} = X_n$.

d. `Step 3:` (LMS algorithm for updating occupation probabilities of X_n) Construct $\widehat{\pi}_{n+1}$ as

$$\widehat{\pi}_{n+1} = \widehat{\pi}_n + \frac{1}{n}(X_{n+1} - \widehat{\pi}_n).$$

e. `Step 4:` (Compute estimate of the solution) $\widehat{\theta}_n^* = e_{i^*}$, where

$$i^* = \arg \max_{i \in \{1,\ldots,S\}} \widehat{\pi}_{n+1}^*;$$

set $n \to n+1$ and go to Step 1 ($\widehat{\pi}_{n+1}^i$ denotes the ith component of the S-dimensional vector $\widehat{\pi}_{n+1}$).

The main convergence results in [2] for the above algorithm can be summarized as follows.

THEOREM 6.1. *Under assumption* (O), *the sequence* $\{X_n\}$ *generated by Algorithm* 1 *is a homogeneous, aperiodic, irreducible Markov chain with state space* $\mathcal{M}$.

Furthermore, for sufficiently large n, $\{X_n\}$ spends more time in $\mathcal{K}$ than other states; i.e., if $\theta = e_i$ is a global minimizer of (6.1), then the stationary distribution $\pi(\theta)$ of $\{X_n\}$ satisfies $\pi^i(\theta) \geq \pi^j(\theta)$, $e_j \in \mathcal{M} - \mathcal{K}$, where $\pi^i(\theta)$ denotes the ith component of $\pi(\theta)$.

The theorem shows that $\widehat{\theta}_n^*$ is attracted to and converges almost surely to an element in $\mathcal{K}$.

6.2. Adaptive discrete stochastic optimization algorithm. Motivated by problems in spreading code optimization of CDMA wireless networks [11], we consider a variant of Algorithm 1 where the optimal solution $\theta \in \mathcal{M}$ of (6.1) is time-varying. Denote this time-varying optimal solution as θ_n. We subsequently refer to θ_n as the *true parameter* or *hypermodel*. Tracking such time-varying parameters is at the very heart of applications of adaptive SA algorithms. We propose the following adaptive algorithm.

ALGORITHM 2 (adaptive discrete stochastic optimization algorithm).

a. `Steps 0-2:` identical to Algorithm 1.

b. `Step 3:` (Constant step-size) Replace Step 3 of Algorithm 1 with a fixed-step-size algorithm, i.e.,

$$\widehat{\pi}_{n+1} = \widehat{\pi}_n + \mu\,(X_{n+1} - \widehat{\pi}_n)\,, \tag{6.4}$$

where the step size μ is a small positive constant.

c. `Step 4:` identical to Algorithm 1.

Note that as long as $0 < \mu < 1$, $\widehat{\pi}_n$ is guaranteed to be a probability vector. Intuitively, the constant step size μ introduces exponential forgetting of the past occupation probabilities and permits tracking of slowly time-varying θ_n. The rest of this section is devoted to obtaining bounds on the error probability of the estimate $\widehat{\theta}_n^*$ generated by Algorithm 2.

6.3. Convergence analysis of adaptive discrete SA algorithm. In adaptive filtering (e.g., LMS), a typical method for analyzing the tracking performance of an adaptive algorithm is to postulate a *hypermodel* for the variation in the true parameter $\{\theta_n\}$. Since $\theta_n \in \mathcal{M}$ and $\mathcal{M}$ is a finite state space, it is reasonable to describe $\{\theta_n\}$ as a slow Markov chain on $\mathcal{M}$ for the subsequent analysis. Henceforth, we assume that (M) holds for $\{\theta_n\}$. Note that the hypermodel assumption is used only for the analysis and does not enter the actual algorithm implementation; see Algorithm 2.

Theorem 6.1 says that for fixed $\theta_n = \theta$ the sequence $\{X_n\}$ generated by Algorithm 2 is a conditional Markov chain (conditioned on θ_n); i.e., assumption (S) of section 2 holds. The update of the occupation probabilities (6.4) is identical to (2.3). Thus the behavior of the sequence $\{\widehat{\pi}_n\}$ generated by Algorithm 2 exactly fits the model of section 2 with $m_0 = S$. In particular, the mean squares analysis of section 3, the limit system of switching ODEs, and switching diffusion limit of section 5 hold.

Owing to the discrete nature of the underlying parameter θ_n, it makes sense to give bounds on the probability of error of the estimates $\widehat{\theta}_n^*$ generated by Step 4 of Algorithm 2. Define the error event E and probability of error $P(E)$ as

$$E = \{\widehat{\theta}_n^* \neq \theta_n\}, \quad P(E) = P(\widehat{\theta}_n^* \neq \theta_n). \tag{6.5}$$

Clearly E depends on n and the step size μ; we suppress the n here for notational simplicity. When we wish to emphasize the n- and μ-dependence, we write it as E_n^μ. Based on the mean square error of Theorem 3.1, the following result holds.

THEOREM 6.2. *Under conditions* (M) *and* (S), *if* $\mu = \varepsilon$, *then there is an* n_1 *such that for all* $n \geq n_1$ *the error probability of the estimate* $\widehat{\theta}_n^*$ *generated by Algorithm* 2 *satisfies*

$$P(E) = P(E_n^\mu) \leq K\mu^{1-2\gamma}, \quad 0 < \gamma < \frac{1}{2}, \tag{6.6}$$

where K *is a positive constant independent of* μ *and* ε.

The above result can be used to check the consistency: As $\mu \to 0$, the probability of error $P(E)$ of the tracking algorithm goes to zero. The constant K can be explicitly determined; however, it is highly conservative.

Proof. The estimate of the maximum generated by the discrete stochastic optimization algorithm at time n is $\widehat{\pi}_n^* = \arg\max_j \widehat{\pi}_n^j$ (where $\widehat{\pi}_n^j$ denotes the jth component of the S-dimensional vector $\widehat{\pi}_n$). Thus the error event E in (6.5) is equivalent to $E = \{\arg\max_i \pi^i(\theta_n) \neq \arg\max_j \widehat{\pi}_n^j\}$. Then clearly the complement event $\bar{E} = \{\arg\max_i \pi^i(\theta_n) = \arg\max_j \widehat{\pi}_n^j\}$ satisfies

$$\begin{aligned}\bar{E} &\supseteq \left\{\left|\max_i \pi^i(\theta_n) - \max_j \widehat{\pi}_n^j\right| \leq \min_{i,j} |\pi^i(\theta_n) - \widehat{\pi}_n^j|\right\} \\ &\supseteq \left\{\left|\max_i \pi^i(\theta_n) - \max_j \widehat{\pi}_n^j\right| \leq L\right\},\end{aligned}$$

where

$$L \leq \min_{i,j} |\pi^i(\theta_n) - \widehat{\pi}_n^j| \tag{6.7}$$

is a positive constant. Then the probability of no error is

$$P(\bar{E}) = P\left(\arg\max_i \pi^i(\theta_n) = \arg\max_j \widehat{\pi}_n^j\right) > P\left(\left|\max_i \pi^i(\theta_n) - \max_j \widehat{\pi}_n^j\right| \leq L\right)$$

for any sufficiently small positive number L. Then, using the above equation and Theorem 3.1,

$$\begin{aligned}P(E) &\leq P\left(\left|\max_i \pi^i(\theta_n) - \max_j \widehat{\pi}_n^j\right| > L\right) \\ &\leq P\left(\max_i |\pi^i(\theta_n) - \widehat{\pi}_n^i| > L\right).\end{aligned} \tag{6.8}$$

Applying Chebyshev's inequality to (3.1) yields, for any i,

$$P(|\pi^i(\theta_n) - \widehat{\pi}_n^i| > L) \leq \frac{1}{L^2} K\mu$$

for some constant K. Thus (6.8) yields

$$P\left(\max_i |\pi^i(\theta_n) - \widehat{\pi}_n^i| > L\right) \leq \frac{1}{L^2} K\mu. \tag{6.9}$$

It only remains to pick a sufficiently small L. Choose $L = \mu^\gamma$, where $0 < \gamma < \frac{1}{2}$ is arbitrary. It is clear that, for sufficiently small μ, L satisfies (6.7). Then (6.9) yields $P(E) \leq K\mu^{1-2\gamma}$. □

Using the diffusion approximation Corollary 5.5 and Theorem 5.6, a sharper upper bound for the error probability can be obtained as follows. First, without loss of generality we may order the states $\overline{\theta}_i \in \mathcal{M}$ so that the covariances $\Sigma(\overline{\theta})$ are, in ascending order,

$$\Sigma(\overline{\theta}_1) \le \Sigma(\overline{\theta}_2) \le \cdots \le \Sigma(\overline{\theta}_S), \tag{6.10}$$

where $\Sigma(\overline{\theta}_i) \le \Sigma(\overline{\theta}_j)$ (resp., $\Sigma(\overline{\theta}_i) < \Sigma(\overline{\theta}_j)$) means that $\Sigma(\overline{\theta}_i) - \Sigma(\overline{\theta}_j)$ is nonnegative definite (resp., positive definite). Note that $\Sigma(\overline{\theta}_i)$ is explicitly computable using (5.11). Define

$$e^{ji} \stackrel{\text{def}}{=} e_j - e_i, \quad \sigma^{ji}(\overline{\theta}) \stackrel{\text{def}}{=} \sqrt{e^{ji,\prime}\Sigma(\overline{\theta})e^{ji}}. \tag{6.11}$$

THEOREM 6.3. *Assume that conditions* (M) *and* (S) *hold and that* $\mu = \varepsilon$. *Then for sufficiently large* n *the error probability of the estimate* $\widehat{\theta}_n$ *generated by Algorithm* 2 *satisfies*

(6.12)
$$P(E) = \sum_{i=1}^{S} P(\theta_n = \overline{\theta}_i)P(E|\theta_n = \overline{\theta}_i) = \sum_{i=1}^{S} z^i(\mu n)P(E|\theta_n = \overline{\theta}_i) + O(\mu + \exp(-k_0 n)),$$

$$\begin{aligned} P(E|\theta_n = \overline{\theta}_i) \le \sum_{\substack{j=1 \\ j\ne i}}^{S} \Bigg[& I(e^{ji,\prime}\bar{\pi}(\mu n) \le 0)\Phi^c\left(\frac{-e^{ji,\prime}\bar{\pi}(\mu n)/\sqrt{\mu}}{\sigma^{ji}(\overline{\theta}_1)/2}\right) \\ & + I(e^{ji,\prime}\bar{\pi}(\mu n) > 0)\Phi^c\left(\frac{-e^{ji,\prime}\bar{\pi}(\mu n)/\sqrt{\mu}}{\sigma^{ji}(\overline{\theta}_S)/2}\right)\Bigg], \end{aligned} \tag{6.13}$$

where $z^i(\cdot), \bar{\pi}(\cdot)$ *are defined in* (5.2), *and* $\sigma^{ji}(\cdot)$ *are defined in* (6.11), *which can be computed using* (5.11) *and* $\Phi^c(\cdot) = 1 - \Phi(\cdot)$, *with* $\Phi(\cdot)$ *being the standard normal distribution function.*

Proof. Clearly $P(E) = \sum_{i=1}^{S} P(\theta_n = \overline{\theta}_i)P(E|\theta_n = \overline{\theta}_i)$. Then (5.2) yields (6.12). Now

$$\begin{aligned} P(E|\theta_n = \overline{\theta}_i) &= P\left(\arg\max_j \widehat{\pi}_n^j \ne e_i | \theta_n = \overline{\theta}_i\right) \\ &= P\left(\bigcup_{\substack{j=1 \\ j\ne i}}^{S} \{\widehat{\pi}_n^j - \widehat{\pi}_n^i > 0\} \mid \theta_n = \overline{\theta}_i\right) \\ &\le \sum_{\substack{j=1 \\ j\ne i}}^{S} P\left(\widehat{\pi}_n^j - \widehat{\pi}_n^i > 0 \mid \theta_n = \overline{\theta}_i\right) \quad \text{(union bound)}. \end{aligned}$$

Upper bounds for each of the $S-1$ terms in the above summation will now be constructed.

Using (5.1), with $\bar{\pi}(\mu n)$ defined in (5.2),

$$\widehat{\pi}_n = \mathbf{E}\{\pi(\theta_n)\} + \sqrt{\mu}v_n = \bar{\pi}(\mu n) + \sqrt{\mu}v_n + O(\mu + \exp(-k_0 n)), \tag{6.14}$$

where $v(t)$, the limit of the interpolation of v_n, satisfies the switching diffusion (5.24), and $\Sigma(\overline{\theta}_i)$ are in ascending order as in (6.10).

Define scalar processes β_n^{ji} and $\beta^{ji}(t)$ as $\beta_n^{ji} = e^{ji,\prime} v_n$ and $\beta^{ji}(t) = e^{ji,\prime} v(t)$. Then $\beta^{ji}(t)$ satisfies the real-valued switching diffusion

$$d\beta^{ji}(t) = -\beta^{ji}(t)dt + \sigma_{ji}(\theta(t))db(t),$$

where $\sigma^{ji}(\theta(t))$ is defined in (6.11) and $b(t)$ is a real-valued standard Brownian motion.

Owing to (6.14), $\widehat{\pi}_n^j - \widehat{\pi}_n^i = e^{ji,\prime}\widehat{\pi}_n = e^{ji,\prime}\bar{\pi}(\mu n) + \sqrt{\mu}\beta_n^{ji} + O(\mu + \exp(-k_0 n))$. Since the $O(\mu + \exp(-k_0 n))$ does not contribute to the limit in distribution, we drop it henceforth. We have

$$P\left(\widehat{\pi}_n^j - \widehat{\pi}_n^i > 0 \mid \theta_n = \overline{\theta}_i\right) = P\left(\beta_n^{ji} > \frac{-e^{ji,\prime}\bar{\pi}(\mu n)}{\sqrt{\mu}} \;\middle|\; \theta_n = \overline{\theta}_i\right). \tag{6.15}$$

Since the process β_n^{ji} is a Gaussian mixture and the limiting process $\beta^{ji}(t)$ is a switching diffusion, it is difficult to explicitly compute the right-hand side of (6.15). However, it can be upper-bounded by considering the Gaussian diffusion processes $\underline{\beta}^{ji}(t)$ and $\overline{\beta}^{ji}(t)$, which are defined as follows:

$$\begin{aligned} d\underline{\beta}^{ji}(t) &= -\underline{\beta}^{ji}(t)dt + \sigma_{ji}(\overline{\theta}_1)db(t), \quad \underline{\beta}^{ji}(0) = \beta^{ji}(0), \\ d\overline{\beta}^{ji}(t) &= -\overline{\beta}^{ji}(t)dt + \sigma_{ji}(\overline{\theta}_S)db(t), \quad \overline{\beta}^{ji}(0) = \beta^{ji}(0). \end{aligned}$$

Due to the ordering of the positive definite matrices $\Sigma(\overline{\theta}_i)$ in (6.10), the scalars $\sigma_{ji}(\overline{\theta}_i)$ satisfy

$$\sigma_{ji}(\overline{\theta}_1) \le \sigma_{ji}(\overline{\theta}_2) \le \cdots \le \sigma_{ji}(\overline{\theta}_S). \tag{6.16}$$

To proceed, we claim the following result and postpone the proof until later.

LEMMA 6.4. *For any $a > 0$, $P(\underline{\beta}^{ji}(t) \le a) \ge P(\beta^{ji}(t) \le a|\theta(t) = \overline{\theta}_i) \ge P(\overline{\beta}^{ji}(t) \le a)$. For any $a \le 0$, $P(\underline{\beta}^{ji}(t) \le a) \le P(\beta^{ji}(t) \le a|\theta(t) = \overline{\theta}_i) \le P(\overline{\beta}^{ji}(t) \le a)$.*

Lemma 6.4 implies that

$$P(\beta^{ji}(t) > a|\theta(t) = \overline{\theta}_i) \le I(a > 0)P(\underline{\beta}^{ji}(t) > a) + I(a \le 0)P(\overline{\beta}^{ji}(t) > a). \tag{6.17}$$

Since $\underline{\beta}^{ji}(t)$ and $\overline{\beta}^{ji}(t)$ are real-valued diffusions and are stable, their stationary covariances are easily computed as $\underline{\sigma}^2 = \sigma_{ji}^2(\overline{\theta}_1)/2$ and $\overline{\sigma}^2 = \sigma_{ji}^2(\overline{\theta}_S)/2$, respectively. Thus, asymptotically $\underline{\beta}^{ji}(t)$, $\overline{\beta}^{ji}(t)$ are Gaussian random variables with zero mean and variance $\sigma_{ji}^2(\overline{\theta}_1)/2$ and $\sigma_{ji}^2(\overline{\theta}_S)/2$, respectively. Then (6.17) yields

$$P(\beta^{ji}(t) > a|\theta(t) = \overline{\theta}_i) \le I(a > 0)\Phi^c\left(\frac{a}{\sigma_{ji}(\overline{\theta}_1)/2}\right) + I(a \le 0)\Phi^c\left(\frac{a}{\sigma_{ji}(\overline{\theta}_S)/2}\right).$$

Thus for sufficiently large n and sufficiently small $\mu > 0$,

$$P(\beta_n^{ji} > a|\theta_n = \overline{\theta}_i) \le I(a > 0)\Phi^c\left(\frac{a}{\sigma^{ji}(\overline{\theta}_1)/2}\right) + I(a \le 0)\Phi^c\left(\frac{a}{\sigma^{ji}(\overline{\theta}_S)/2}\right).$$

Using this in (6.15) proves the theorem. □

Proof of Lemma 6.4. Let $t_1 < t_2 < \cdots < t_N \le t$ denote the sequence of jump times of the Markov chain $\{\theta(t)\}$. Let $\mathcal{G}_t$ denote the σ-algebra generated by $\{\theta(s) : s < t, \theta(t)\}$. Then

$$\beta^{ji}(t) = e^{-t}\left[\sigma_{ji}(\theta(0))\int_0^{t_1-} e^{\tau}db(\tau) + \sigma_{ji}(\theta(t_1))\int_{t_1}^{t_2-} e^{\tau}db(\tau) + \cdots + \sigma_{ji}(\theta(t_N))\int_{t_N}^{t} e^{\tau}db(\tau)\right],$$

$$\underline{\beta}^{ji}(t) = e^{-t}\left[\sigma_{ji}(\overline{\theta}_1)\int_0^{t_1-} e^{\tau}db(\tau) + \sigma_{ji}(\overline{\theta}_1)\int_{t_1}^{t_2-} e^{\tau}db(\tau) + \cdots + \sigma_{ji}(\overline{\theta}_1)\int_{t_N}^{t} e^{\tau}db(\tau)\right],$$

where $\underline{\beta}^{ji}(t)$ is a zero mean scalar Gaussian variable. Conditioned on $\mathcal{G}_t$, $\beta^{ji}(t)$ is a zero mean scalar Gaussian random variable. Since $\sigma_{ji}(\overline{\theta}_1) \le \sigma_{ji}(\theta(t))$ for all t by (6.16), clearly $\mathbf{E}\{\underline{\beta}^{ji}(t)\}^2 \le \mathbf{E}\{\beta^{ji}(t)\}^2$. Hence for $x > 0$, $\mathbf{E}\{I(\underline{\beta}^{ji}(t) \le x)\} > \mathbf{E}\{I(\beta^{ji}(t) \le x)|\mathcal{G}_t, \theta(t)\}$. Taking $\mathbf{E}\{\cdot|\theta(t)\}$ on both sides and using the fact that $\underline{\beta}^{ji}(t)$ is independent of $\theta(t)$ yields $P(\underline{\beta}^{ji}(t) \le x) > P(\beta^{ji}(t) \le x|\theta(t))$. The result for $\overline{\beta}^{ji}(t)$ is established similarly. □

Remark 6.5. First, Markov chain Monte Carlo–based simulation methods can be used to evaluate the probability of error of the algorithm. In addition, a Gaussian approximation–based heuristic expression can be obtained for the probability error bounds of Algorithm 2 in lieu of Theorem 6.3. Consider a real-valued switching diffusion process

$$dx = -xdt + \sigma(\theta(t))db,$$

where $\theta(t)$ is the limit Markov chain as in section 5. The negative term $-x$ implies that the system is stable. Thus, by virtue of an argument as in [16, p. 323], the covariance is given by

$$\mathbf{E}x(t)x(0) = \mathbf{E}\left(\int_{-\infty}^{t}\exp(-(t-s))\sigma(\theta(s))db(s)\right)\left(\int_{-\infty}^{0}\exp(-s)\sigma(\theta(s))db(s)\right).$$

Assume in addition that the generator Q of the Markov chain $\theta(t)$ (the one given in condition (M)) is irreducible, which implies (see [22]) that, except for zero, all other eigenvalues are on the left half of the complex plan. As a result, the stationary covariance exists and is given by

$$\widetilde{\sigma}^2 = \mathbf{E}\sum_{l=1}^{S}\int_0^{\infty}\exp(-2s)\sigma^2(\overline{\theta}_l)I_{\{\theta(s)=\overline{\theta}_l\}}ds. \tag{6.18}$$

This covariance may be computed via the Monte Carlo method. Using $\widetilde{\sigma}^2$, an approximation of the probability of error for Algorithm 2 can be computed.

REFERENCES

[1] M. Alrefaei and S. Andradóttir, *A simulated annealing algorithm with constant temperature for discrete stochastic optimization*, Management Sci., 45 (1999), pp. 748–764.

[2] S. Andradóttir, *A global search method for discrete stochastic optimization*, SIAM J. Optim., 6 (1996), pp. 513–530.

[3] S. Andradóttir, *Accelerating the convergence of random search methods for discrete stochastic optimization*, ACM Trans. Model. Comput. Simul., 9 (1999), pp. 349–380.
[4] A. Benveniste, M. Metivier, and P. Priouret, *Adaptive Algorithms and Stochastic Approximations*, Springer-Verlag, New York, 1990.
[5] P. Billingsley, *Convergence of Probability Measures*, Wiley, New York, 1968.
[6] P. Bremaud, *Markov Chains*, Springer-Verlag, New York, 1999.
[7] H.-F. Chen, *Stochastic Approximation and Its Applications*, Kluwer Academic Publishers, Dordrecht, The Netherlands, 2002.
[8] Y. Ephraim and N. Merhav, *Hidden Markov processes*, IEEE Trans. Inform. Theory, 48 (2002), pp. 1518–1569.
[9] S. B. Gelfand and S. K. Mitter, *Simulated annealing with noisy or imprecise energy measurements*, J. Optim Theory Appl., 62 (1989), pp. 49–62.
[10] W.-B. Gong, Y.-C. Ho, and W. Zhai, *Stochastic comparison algorithm for discrete optimization with estimation*, SIAM J. Optim., 10 (1999), pp. 384–404.
[11] V. Krishnamurthy, X. Wang, and G. Yin, *Spreading code optimization and adaptation in CDMA via discrete stochastic approximation*, IEEE Trans. Inform. Theory, to appear.
[12] V. Krishnamurthy, X. Wang, and G. Yin, *Adaptive spreading code optimization in multiantenna multipath fading channels in CDMA*, in Proceedings of the IEEE Conference on Communications (ICC), Anchorage, AK, 2003, pp. 2445–2449.
[13] V. Krishnamurthy and G. Yin, *Recursive algorithms for estimation of hidden Markov models and autoregressive models with Markov regime*, IEEE Trans. Inform. Theory, 48 (2002), pp. 458–476.
[14] H. J. Kushner, *Approximation and Weak Convergence Methods for Random Processes, with Applications to Stochastic Systems Theory*, MIT Press, Cambridge, MA, 1984.
[15] H. J. Kushner and D. S. Clark, *Stochastic Approximation Methods for Constrained and Unconstrained Systems*, Springer-Verlag, 1978.
[16] H. J. Kushner and G. Yin, *Stochastic Approximation Algorithms and Recursive Algorithms and Applications*, 2nd ed., Springer-Verlag, New York, 2003.
[17] L. Ljung, *Analysis of recursive stochastic algorithms*, IEEE Trans. Automat. Control, 22 (1977), pp. 551–575.
[18] G. Pflug, *Optimization of Stochastic Models: The Interface between Simulation and Optimization*, Kluwer Academic Publishers, Norwell, MA, 1996.
[19] V. Solo and X. Kong, *Adaptive Signal Processing Algorithms—Stability and Performance*, Prentice–Hall, Englewood Cliffs, NJ, 1995.
[20] J. R. Swisher, S. H. Jacobson, P. D. Hyden, and L. W. Schruben, *A survey of simulation optimization techniques and procedures*, in Proceedings of the 2000 Winter Simulation Conference, Orlando, FL, 2000, pp. 119–128.
[21] D. Yan and H. Mukai, *Stochastic discrete optimization*, SIAM J. Control Optim., 30 (1992), pp. 594–612.
[22] G. Yin and Q. Zhang, *Continuous-Time Markov Chains and Applications: A Singular Perturbation Approach*, Springer-Verlag, New York, 1998.
[23] G. Yin and Q. Zhang, *Singularly perturbed discrete-time Markov chains*, SIAM J. Appl. Math., 61 (2000), pp. 834–854.
[24] G. Yin, Q. Zhang, and G. Badowski, *Discrete-time singularly perturbed Markov chains: Aggregation, occupation measures, and switching diffusion limit*, Adv. Appl. Probab., 35 (2003), pp. 449–476.

SIAM J. OPTIM.
Vol. 14, No. 4, pp. 1216–1236

PATTERN SEARCH METHODS FOR USER-PROVIDED POINTS: APPLICATION TO MOLECULAR GEOMETRY PROBLEMS*

PEDRO ALBERTO†, FERNANDO NOGUEIRA†, HUMBERTO ROCHA‡, AND LUÍS N. VICENTE§

Abstract. This paper deals with the application of pattern search methods to the numerical solution of a class of molecular geometry problems with important applications in molecular physics and chemistry. The goal is to find a configuration of a cluster or a molecule with minimum total energy.

The minimization problems in this class of molecular geometry problems have no constraints, and the objective function is smooth. The difficulties arise from the existence of several local minima and, especially, from the expensive function evaluation (total energy) and the possible nonavailability of first-order derivatives.

We introduce a pattern search approach that attempts to exploit the physical nature of the problem by using energy lowering geometrical transformations and to take advantage of parallelism without the use of derivatives. Numerical results for a particular instance of this new class of pattern search methods are presented, showing the promise of our approach.

The new pattern search methods can be used in any other context where there is a user-provided scheme to generate points leading to a potential objective function decrease.

Key words. pattern search methods, expensive function evaluations, parallel computing, user-provided points, molecular geometry, geometrical transformations

AMS subject classifications. 49M37, 65K05, 65Z05, 81V55, 90C30, 90C56, 90C90, 92E99

DOI. 10.1137/S1052623400377955

1. Introduction.

The motivation behind the study of the geometrical arrangement of atoms in a molecule or cluster is its close relation to their chemical and physical properties (e.g., optical response). For example, patterns in the structure of related systems can give a powerful insight into their physical properties. This is the case, for instance, of atomic clusters of different sizes of a single element, or of different elements in the same group of the periodic table. In most cases, clear and unambiguous structural information is difficult to obtain experimentally; theory then plays a particularly important role.

The stable configurations of atoms in any material can be found by minimization of the total energy of the system with respect to the atomic positions. The most stable structure is the one with the lowest total energy. The theoretical procedure can be seen as two separate problems: obtaining the total energy for a given configuration and minimizing it with respect to the atomic coordinates. Only the second problem is to be addressed in this work. There are several geometrically distinct structures (isomers) (i.e., structures with the same number of atoms but different shapes) for which the total energy is locally minimized. As some of these can be simultaneously

*Received by the editors September 8, 2000; accepted for publication (in revised form) December 30, 2003; published electronically August 4, 2004. This work was supported by the FCT under grant Praxis/P/FIS/14195/1998.

http://www.siam.org/journals/siopt/14-4/37795.html

†Departamento de Física da Universidade de Coimbra, 3004-516 Coimbra, Portugal; Centro de Física Computacional (pedro@teor.fis.uc.pt, fnog@teor.fis.uc.pt).

‡Universidade Católica, Pólo de Viseu, 3504-505 Viseu, Portugal (hrocha@mat.uc.pt).

§Departamento de Matemática da Universidade de Coimbra, 3001-454 Coimbra, Portugal (lnv@mat.uc.pt). Support for this author was also provided by Centro de Matemática da Universidade de Coimbra, by the FCT under grant POCTI/35059/MAT/2000, by the European Union under grant IST-2000-26063, and by Fundação Calouste Gulbenkian.

present in an experiment, it is sometimes desirable not only to find the lowest energy structure but also to find other low-lying local minima. The number of these local energy minima grows exponentially with the number of atoms, making it hard to find the lowest energy configuration of a moderately sized cluster or molecule,[1] even when using two-body potentials that give rise to smooth energy surfaces. For Lennard–Jones clusters, it has been found that the number of isomers grows from 2 for a 6-atom cluster to 988 for a 13-atom cluster [14], although realistic potentials yield fewer local minima.

With the exception of noble gases, Lennard–Jones potentials provide very unrealistic descriptions of physical systems. We are interested in more realistic approximations such as the local density plane-wave total energy calculation [18] briefly described in Appendix B. The expensive numerical minimization of the total energy calculated with this method motivated the work reported in this paper.

Methods commonly used to minimize the total energy include simulated annealing, steepest descent and other gradient-based methods, and genetic algorithms. Good results have been obtained by coupling some of these. An example is the so-called Langevin dynamics method, proposed some years ago by Biswas and Hamann [3]. This minimization method is a combination of simulated annealing and gradient techniques, and has proved to be very efficient for small molecules. The total energy gradient gives the internal forces on the atoms, which can be used to "guide" the annealing process, i.e., to introduce a bias in the minimization process, turning it into a "smart simulated annealing," as the system does not evolve at random. Despite the use of the gradient, the Langevin dynamics method retains the possibility of moving away from local minima that are not global. But this approach has several drawbacks. On one hand, it is not by itself parallelizable. On the other hand, it is developed to run for a given fixed number of iterations, instead of incorporating an autonomous stopping criterion. Furthermore, the method requires the possibly expensive calculation of the gradient. Another popular method that combines simulated annealing and gradient techniques is the method of Car and Parrinello [6] that also shares these numerical inconveniences.

In many cases, obtaining the gradient of the total energy can be too time-consuming. Only in the simpler (least accurate) methods of calculating the total energy of a cluster is the gradient available at moderate cost. Moreover, there are situations where the gradient is not available [33]. Therefore, many interesting problems in physics are being tackled using methods where no calculation of the total energy gradient is required. Among the methods which do not require the computation of the gradient are pattern search methods. In this paper, we develop a class of pattern search methods suited for molecular geometry problems and apply it to sodium clusters to determine the geometry that minimizes the total energy. (Sodium clusters are a typical, well-known test system in cluster physics.) This paper does not address the local refinement that could be achieved by applying local optimization techniques after the conformational searching has been applied, in order to identify more precisely the global optimizer; pattern search methods are used only in the conformational searching.

Pattern search methods are an instance of direct search methods where the step directions are not modified at the end of each iteration. Examples of pattern search methods are the coordinate search with fixed step lengths, evolutionary

[1] The terms cluster and molecule will be used without any distinction being made between them. In all cases they should be interpreted as referring simply to a collection of atoms.

operation using factorial designs [5], the original pattern search algorithm of Hooke and Jeeves [16], and the multidirectional search algorithm of Dennis and Torczon [10] (see also [28]), also referred to as the parallel direct search (PDS) method. A unified framework for pattern search methods was proposed by Torczon [29] and improved by Audet and Dennis [2] (see also the essay [21]). Surveys of other derivative free methods, including other direct search methods for unconstrained optimization (such as the well-known Nelder–Mead algorithm [23]) can be found in [8, 27, 30, 32].

The application of pattern search methods to molecular geometry is not new. Meza and Martinez [22] have compared PDS, genetic algorithms, and simulated annealing using Lennard–Jones potentials, concluding that PDS could also be used in conformational searching, and showing that it performed as well as genetic algorithms and substantially better than simulated annealing for large molecules. Pattern search methods have also been combined with evolutionary techniques (see the work by Hart [12, 13]), and the resulting evolutionary pattern search method compared favorably with evolutionary algorithms.

This paper is divided as follows. We start in section 2 by introducing pattern search methods in a quite general framework. In section 3 we introduce our new class of pattern search methods for user-provided points: section 3.1 presents a family of positive bases with desirable uniform directionality properties, and section 3.2 combines the pattern generated by these positive bases with user-provided points and develops the new class of pattern search methods. The user-provided points computation is illustrated by introducing geometrical transformations with physical meaning in the context of molecular geometry (see Appendix A). In section 4 we show numerical results with an implementation of our pattern search methods for user-provided points in molecular geometry problems. Finally, in section 5 we draw conclusions and present prospects for future work. In Appendix B we provide a brief description of the total energy evaluation and comment on its numerical complexity.

This new class of pattern search methods can also be applied in other contexts where the user can provide a scheme to compute points that may lead to an objective function decrease. We have implemented this class of pattern search methods for general user-provided points as well as for the molecular geometry context described above. The codes and their documentation can be downloaded from the web site http://www.mat.uc.pt/∼lnv/psm/. Both versions have been implemented in Fortran 95. The parallel version uses the parallelization protocol MPI; see [1] for more details.

2. Pattern search methods and positive bases. We use $\|\cdot\|$ and $\langle\cdot,\cdot\rangle$ to represent the Euclidean norm and inner product, respectively. By abuse of notation, if A is a matrix, $a \in A$ means that the vector a is a column of A. It will be also convenient to assume that $[a_1 \cdots a_r]$ represents not only the matrix with r columns, but also, depending on the context, the set of r vectors $\{a_1, \ldots, a_r\}$. The identity matrix is denoted by I and its ith column by e_i. Finally, we write e to represent a vector of ones with appropriate size.

2.1. Positive bases. We present a few basic properties of positive bases from the theory of positive linear dependence developed by Davis [9] (see also Lewis and Torczon [20]). The *positive span*[2] of a set of vectors $[v_1 \cdots v_r]$ is the convex cone

$$\{v \in \mathbb{R}^n : \ v = \alpha_1 v_1 + \cdots + \alpha_r v_r, \quad \alpha_i \geq 0, \ i = 1, \ldots, r\}.$$

[2]Strictly speaking, we should have written *nonnegative* instead of positive, but we decided to follow the notation in [9, 20]. We also note that by *span* we mean *linear span*.

The set $[v_1 \cdots v_r]$ is said to be *positively dependent* if one of the vectors is in the convex cone positively spanned by the remaining vectors, i.e., if one of the vectors is a positive combination of the others; otherwise, the set is *positively independent*. A *positive basis* is a positively independent set whose positive span is $\mathbb{R}^n$. Alternatively, a positive basis for $\mathbb{R}^n$ can be defined as a set of nonzero vectors of $\mathbb{R}^n$ whose positive combinations span $\mathbb{R}^n$, but no proper set does. The following theorem [9] indicates that a positive spanning set contains at least $n+1$ vectors in $\mathbb{R}^n$.

THEOREM 2.1. *If $[v_1 \cdots v_r]$ positively spans $\mathbb{R}^n$, then it contains a subset with $r-1$ elements that spans $\mathbb{R}^n$.*

It can also be shown that a positive basis cannot contain more than $2n$ elements [9]. Positive bases with $n+1$ and $2n$ elements are referred to as *minimal* and *maximal* positive bases, respectively.

We now present three necessary and sufficient characterizations for a set that spans $\mathbb{R}^n$ to also span $\mathbb{R}^n$ positively [9].

THEOREM 2.2. *Let $[v_1 \cdots v_r]$, with $v_i \neq 0$ for all $i \in \{1, \ldots, r\}$, span $\mathbb{R}^n$. Then the following are equivalent:*

(i) *$[v_1 \cdots v_r]$ positively spans for $\mathbb{R}^n$.*
(ii) *For every $i = 1, \ldots, r$, $-v_i$ is in the convex cone positively spanned by the remaining $r-1$ vectors.*
(iii) *There exist real scalars $\alpha_1, \ldots, \alpha_r$ with $\alpha_i > 0$, $i \in \{1, \ldots, r\}$, such that $\sum_{i=1}^{r} \alpha_i v_i = 0$.*
(iv) *For every nonzero vector $b \in \mathbb{R}^n$, there exists an index i in $\{1, \ldots, r\}$ for which $b^\top v_i > 0$.*

The following result provides a simple mechanism for generating different positive bases. The proof can be found in [20].

THEOREM 2.3. *Suppose $[v_1 \cdots v_r]$ is a positive basis for $\mathbb{R}^n$ and $B \in \mathbb{R}^{n \times n}$ is a nonsingular matrix. Then $[Bv_1 \cdots Bv_r]$ is also a positive basis for $\mathbb{R}^n$.*

From Theorems 2.2 and 2.3, we can easily deduce the following corollary.

COROLLARY 2.1. *Let $B = [b_1 \cdots b_n] \in \mathbb{R}^{n \times n}$ be a nonsingular matrix. Then $[B \ -\sum_{i=1}^{n} b_i]$ is a positive basis for $\mathbb{R}^n$.*

A trivial consequence of this corollary is that $[I \ -e]$ is a positive basis.

2.2. Pattern search methods. We present pattern search methods for unconstrained optimization problems of the form

$$\min \ f(x), \quad x \in \mathbb{R}^n,$$

and briefly describe their main convergence properties. Pattern search methods are iterative methods generating a sequence of iterates $\{x_k\}$. Given the current iterate x_k, at each iteration k, the next point x_{k+1} is chosen from a finite number of candidates on a given *mesh* M_k. The next iterate, if iteration k is *successful*, must provide a decrease on the objective function: $f(x_{k+1}) < f(x_k)$.

In order to define the mesh M_k, let us consider a set $\mathcal{V}$ of m positive bases. For convenience, let us abuse notation and also denote by $\mathcal{V}$ the matrix whose columns correspond to the vectors in the m positive bases. The number of columns of $\mathcal{V}$, denoted by $|\mathcal{V}|$, is the sum of the number of vectors in all positive bases. The mesh at iteration k is then defined as

$$M_k \ = \ \{x_k + \Delta_k \mathcal{V} z : \ z \in W \subseteq \mathbb{Z}^{|\mathcal{V}|}\}, \tag{2.1}$$

where $\Delta_k > 0$ is the mesh size parameter. Possible choices for W are

$$W = \mathbb{Z}^{|\mathcal{V}|}, \qquad W = \mathbb{N}^{|\mathcal{V}|}.$$

The choice we actually use in our implementation is

$$W = \{ne_i : n \in \mathbb{N},\ i = 1, \ldots, |\mathcal{V}|\}. \tag{2.2}$$

The mechanism of pattern search methods is best explained by considering two phases at every iteration. The first phase, or step, consists of a finite search on the mesh, with the goal of finding a new iterate that decreases the value of the objective function at the current iterate. This step, called the *search step*, is free of any other rules, as long as it searches only a finite number of points in the mesh. If the search step is unsuccessful, a second phase or step, called the *poll step*, is performed around the current iterate with the goal of decreasing the objective function.

The poll step follows stricter rules and appeals to the concept of a positive basis described in the previous section. In this step the candidate for a new iterate x_{k+1} is chosen in the *mesh neighborhood* around x_k

$$\mathcal{N}(x_k) \ = \ \{x_k + \Delta_k v : \text{ for all } v \in V_k(x_k)\},$$

where $V_k(x_k)$ is a positive basis chosen from the finite set $\mathcal{V}$ of positive bases. This set $\mathcal{V}$ of positive bases is specified *a priori*, but the choice of each $V_k(x_k) \in \mathcal{V}$ may depend on k and x_k. Note that the poll step also searches points in the mesh since every column v of any of the positive bases in $\mathcal{V}$ is of the form $\mathcal{V}z$ with $z = e_i$ for a given $i \in \{1, \ldots, |\mathcal{V}|\}$.

We now have all the ingredients to describe pattern search methods.

ALGORITHM 2.1 (pattern search methods).

0. *Initialization.* Choose a rational number $\tau > 1$ and an integer number $m_{max} \geq 1$. Choose $x_0 \in \mathbb{R}^n$ and $\Delta_0 \in \mathbb{R}_+$. Set $k = 0$.
1. *Search step (in current mesh).* With the goal of decreasing $f(x_k)$, try to obtain x_{k+1}^{trial} by evaluating f at a finite number of points in M_k. If $x_{k+1}^{trial} \in M_k$ is found satisfying $f(x_{k+1}^{trial}) < f(x_k)$, then set $x_{k+1} = x_{k+1}^{trial}$, and go to step 3, expanding M_k. (The search step and iteration are declared successful.)
2. *Poll step (in mesh neighborhood given by the positive basis).* This step is reached only if the search step is unsuccessful. If $f(x_k) \leq f(x)$ for every x in the mesh neighborhood $\mathcal{N}(x_k)$, go to step 4, shrinking M_k. (The poll step and iteration are declared unsuccessful.) Otherwise, choose a point $x_{k+1} \in \mathcal{N}(x_k)$ such that $f(x_{k+1}) < f(x_k)$ and go to step 3, expanding M_k. (The poll step and iteration are declared successful.)
3. *Mesh expansion (at successful iterations).* Let $\Delta_{k+1} = \tau^{m_k^+} \Delta_k$ (with $0 \leq m_k^+ \leq m_{max}$). Increase k by one, and move to step 1 for a new iteration.
4. *Mesh reduction (at unsuccessful iterations).* Let $\Delta_{k+1} = \tau^{m_k^-} \Delta_k$ (with $-m_{max} \leq m_k^- \leq -1$). Increase k by one, and move to step 1 for a new iteration.

The search step provides the flexibility for a global search and influences the quality of the local minimizer or stationary point found by the method. The poll step is applied when the search step fails to produce a better point. The poll step attempts to perform a local search in a mesh neighborhood that, for a sufficient small mesh parameter Δ_k, is guaranteed to provide a function reduction, unless the current iterate is at a stationary point (a fact that can be inferred by Theorem 2.2.iv with $b = -\nabla f(x_k)$). So, if the poll step also fails, the mesh parameter Δ_k must be decreased.

An interesting feature of pattern search methods is the simple way in which they can be parallelized. The poll step and the search step can be implemented by

requiring different processors to evaluate the objective function at different points; their strategies can actually depend on the number of processors available.

Pattern search methods, as described above, share the following convergence result, provided the following assumption is made on the mesh: each column i of $\mathcal{V}$ is given by $G\bar{z}_i$, where $G \in \mathbb{R}^{n\times n}$ is a nonsingular generating matrix and $\bar{z}_i$ is an integer vector in $\mathbb{Z}^n$.

THEOREM 2.4. *Suppose that the level set* $L(x_0) = \{x \in \mathbb{R}^n : \ f(x) \leq f(x_0)\}$ *is compact and that* f *is continuously differentiable in an open set containing* $L(x_0)$. *Then*

$$\liminf_{k\longrightarrow+\infty} \|\nabla f(x_k)\| \ = \ 0,$$

and there exists at least one limit point x_* *such that* $\nabla f(x_*) = 0$.

Furthermore, if $\lim_{k\longrightarrow+\infty} \Delta_k = 0$, $\|x_{k+1} - x_k\| \leq C\Delta_k$ *for some constant* $C > 0$ *independent of the iteration counter* k, *and* $x_{k+1} = \operatorname{argmin}_{x\in\mathcal{N}(x_k)} f(x)$ *in the poll step, then*

$$\lim_{k\longrightarrow+\infty} \|\nabla f(x_k)\| \ = \ 0,$$

and every limit point x_* *satisfies* $\nabla f(x_*) = 0$.

The proof can be found, for instance, in [2, 20, 29].

We note finally that the condition $x_{k+1} = \operatorname{argmin}_{x\in\mathcal{N}(x_k)} f(x)$ can be implemented in the poll step and that the condition $\|x_{k+1} - x_k\| \leq C\Delta_k$ is verified for some positive constant C if the choice of z in (2.1) is limited to a bounded set.

The results of Theorem 2.4 concern the ability of pattern search methods to converge globally (i.e., from arbitrary points) to local minimizers candidates. We recall, despite the nonexistence of any supporting theory, that there is numerical evidence about the capability of pattern search methods to compute global minimizers (see the papers [12, 13, 22] and the numerical experiments reported in this paper).

3. Pattern search methods for user-provided points: Application to molecular geometry problems. Having described pattern search methods in a general framework, we turn now to their application to the situation where one would like to take advantage of a user-provided points calculation, like the one we will describe in the context of molecular geometry problems. Our goal is to develop a class of pattern search methods especially tailored to these problems, where each optimization step is physically meaningful.

We accomplish our intention by identifying a set of geometrical transformations—the user-provided points—viewed as deformations of the molecular shape with physical meaning that may provide an energy lowering path. However, as we will see in Appendix A, these geometrical transformations are dependent on the data of the current configuration. In other words, they depend on each optimization point x_k, which stores the coordinates of the current configuration, and therefore they cannot themselves define a pattern and a mesh. (Asymptotically, the dependence would be on the sequence $\{x_k\}$, ruining the finiteness property of the pattern matrices.)

As we will see in section 3.1, we then define a pattern with interesting uniform directionality properties to fit the geometrical transformation procedure, or any other user-provided points calculation.

A new trial point for the search step is computed by geometrical transformation followed by a computation that determines approximately the closest point in the patterned mesh to the point calculated by geometrical transformation.

The positive basis needed to define the mesh neighborhood in the poll step is identified after a point is computed again by a geometrical transformation: among all the vectors in the set of positive bases, the one that makes the smallest angle with the vector defined by the current point and the point computed by the geometrical transformation is identified. This vector in turn identifies the positive basis to be used in the poll step.

The search and poll steps of this new class of pattern search methods for user-provided points (e.g., geometrical transformations) are described in section 3.2.

3.1. Positive bases with uniform angles. We start by introducing the pattern onto which geometrical transformations will be projected. Consider $n+1$ vectors $v_1, \ldots, v_{n+1}$ in $\mathbb{R}^n$ for which all the angles between pairs v_i, v_j $(i \neq j)$ have the same amplitude α. Assuming that the $n+1$ vectors are normalized, this requirement is expressed as

$$a \;=\; \cos(\alpha) \;= \langle v_i, v_j \rangle, \qquad i, j \in \{1, \ldots, n+1\},\ i \neq j, \tag{3.1}$$

where $a \neq 1$. One can show that $a = -1/n$. Let us assume, without loss of generality, that

$$v_{n+1} \;=\; \sum_{i=1}^{n} \alpha_i v_i \tag{3.2}$$

for some scalars $\alpha_1, \ldots, \alpha_n \in \mathbb{R}$. From (3.1) and (3.2), we obtain

$$1 = a \sum_{i=1}^{n} \alpha_i, \tag{3.3}$$

$$a = \sum_{i=1, i \neq j}^{n} a\alpha_i + \alpha_j, \qquad j = 1, \ldots, n. \tag{3.4}$$

Adding all the rows in (3.4) yields

$$na \;=\; (1 + (n-1)a) \sum_{i=1}^{n} \alpha_i. \tag{3.5}$$

From (3.3) and (3.5) we have that $na^2 + (1-n)a - 1 = 0$, and thus, since $a \neq 1$, we conclude that $a = -1/n$.

Now we seek a set of $n+1$ normalized vectors $[v_1 \cdots v_{n+1}]$ satisfying property (3.1) with $a = -1/n$. Let us first compute $v_1, \ldots, v_n$; i.e., let us compute a matrix $V = [v_1 \cdots v_n]$ such that

$$V^\top V \;=\; A,$$

where A is the matrix given by

$$A = \begin{bmatrix} 1 & -1/n & -1/n & \cdots & -1/n \\ -1/n & 1 & -1/n & \cdots & -1/n \\ \vdots & & \ddots & & \\ \vdots & & & \ddots & \\ -1/n & -1/n & -1/n & \cdots & 1 \end{bmatrix}.$$

The matrix A is symmetric and diagonally dominant with positive diagonal entries, and, therefore, it is positive definite [11]. Thus, we can make use of its eigenvalue decomposition

$$A \;=\; Q\Lambda Q^{\top},$$

where $Q \in \mathbb{R}^{n\times n}$ is orthogonal and Λ is a diagonal matrix of order n with positive diagonal entries. Given this decomposition, one can easily see that a choice for V is determined by

$$V \;=\; [v_1 \cdots v_n] \;=\; Q\Lambda^{\frac{1}{2}}Q^{\top}. \tag{3.6}$$

The vector v_{n+1} is then computed by

$$v_{n+1} \;=\; -\sum_{i=1}^{n} v_i. \tag{3.7}$$

It is obvious that $\langle v_i, v_{n+1}\rangle = -1/n$, $i = 1, \ldots, n$, and $\langle v_{n+1}, v_{n+1}\rangle = 1$.

Since V is nonsingular and v_{n+1} is determined by (3.7), we can apply Corollary 2.1 to establish that $[v_1 \cdots v_{n+1}]$ is a (minimal) positive basis.

Our goal is now to generate, from the positive basis $[v_1 \cdots v_{n+1}]$ given by (3.6)–(3.7), a set of positive bases such that: (i) the overall set of vectors captures the directionality of $\mathbb{R}^n$ as well as possible and (ii) each element of the set is itself a positive basis satisfying the uniform angle property (3.1) with $a = -1/n$. First, let us consider a "rotation" $U[e_1 \cdots e_n] = [u_1 \cdots u_n]$ of the coordinate axes $[e_1 \cdots e_n]$ given by the a priori fixed orthogonal matrix $U = [u_1 \cdots u_n]$. The first positive basis is computed by $U_1[v_1 \cdots v_{n+1}]$, where U_1 is an orthogonal matrix that "rotates" v_1 into u_1:

$$U_1 v_1 \;=\; u_1.$$

A choice for U_1 is the Householder transformation

$$U_1 \;=\; I - \pi^{-1}uu^{\top}, \quad u \;=\; v_1 - u_1, \quad \pi \;=\; \frac{1}{2}\|u\|^2.$$

The ith positive basis is then obtained by "rotating" v_1 into u_i. However, since $u_i = Ue_i$ and $e_i = U^{\top}u_i$, there is no need to compute another Householder transformation. In fact, we easily see that

$$U\mathcal{P}_{1i}U^{\top}U_1 v_1 \;=\; u_i,$$

where $\mathcal{P}_{1i}$ is the permutation matrix obtained from the identity by interchanging rows 1 and i. Thus the ith positive basis is given by $U_i[v_1 \cdots v_{n+1}]$, where U_i is the orthogonal matrix

$$U_i = U\mathcal{P}_{1i}U^{\top}U_1.$$

The desired set of positive bases is given by these n positive bases and their corresponding symmetrical counterparts:

$$\mathcal{V} \;=\; [\,U_1[v_1 \cdots v_{n+1}] \cdots U_n[v_1 \cdots v_{n+1}] \;-U_1[v_1 \cdots v_{n+1}] \cdots - U_n[v_1 \cdots v_{n+1}]\,]. \tag{3.8}$$

The number of positive bases is therefore $m = 2n$.

The vectors in $\mathcal{V}$ are reasonably well distributed by amplitude in $\mathbb{R}^n$. In Figure 3.1 we depict the mesh (2.1) with the choices of $\mathcal{V}$ and W respectively given by (3.8) and (2.2); the matrix U given above was set to the identity.

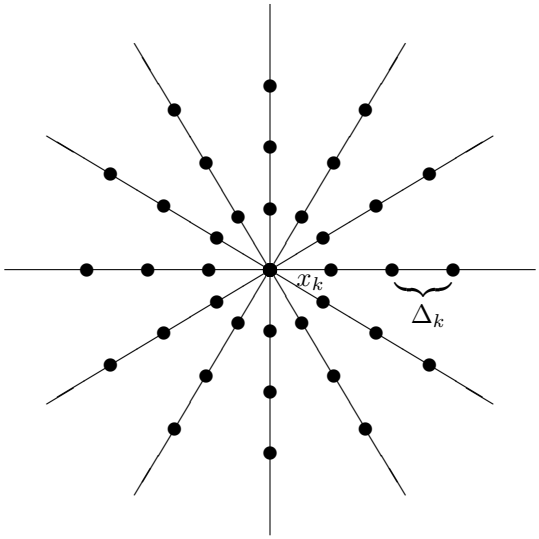


FIG. 3.1. *Mesh for $n = 2$. There are 4 uniform positive bases.*

3.2. The new pattern search framework. Finally, we combine the procedure introduced in section 3.1 with the technique described in Appendix A, and define our class of pattern search methods for molecular geometry problems. We describe how the computation of new points (by geometrical transformations) can determine a pattern search method using, for instance, the pattern described in section 3.1. The same ideas can be used in any application where the user has a scheme to provide the calculation of new points (see also [1]).

The new search and poll steps are described in a parallel environment with N_p processors. We start by showing how the computation of a trial point x_{k+1}^{trial} can be carried out in the search step.

SEARCH STEP: COMPUTATION OF x_{k+1}^{trial}. For each processor p in $\{1, \ldots, N_p\}$:

1. Compute a trial point $u_{p,k+1}^{gt}$ by a geometrical transformation.
2. Solve the integer programming problem

$$\min_{z \in W} \|u_{p,k+1}^{gt} - (x_k + \Delta_k \mathcal{V} z)\| \tag{3.9}$$

to determine a point $x_{p,k+1}^{gt}$, in M_k, closest to $u_{p,k+1}^{gt}$.

3. Set

$$x_{k+1}^{trial} \;=\; \text{argmin}_{x_{p,k+1}^{gt}} f(x_{p,k+1}^{gt})\,.$$

Using the mesh (2.1) with the choices of $\mathcal{V}$ and W respectively given by (3.8) and (2.2), as we do in our implementation, the computation of $x_{p,k+1}^{gt}$ as the solution of the integer programming problem (3.9) can be carried out with relatively little computational effort (see also Figure 3.2). In fact, it can be easily checked that the linear algebra cost is of the order of n^3, which for small n is relatively low compared to the cost of expensive function evaluations such as the total energy computed by local density plane-waves (Appendix B).

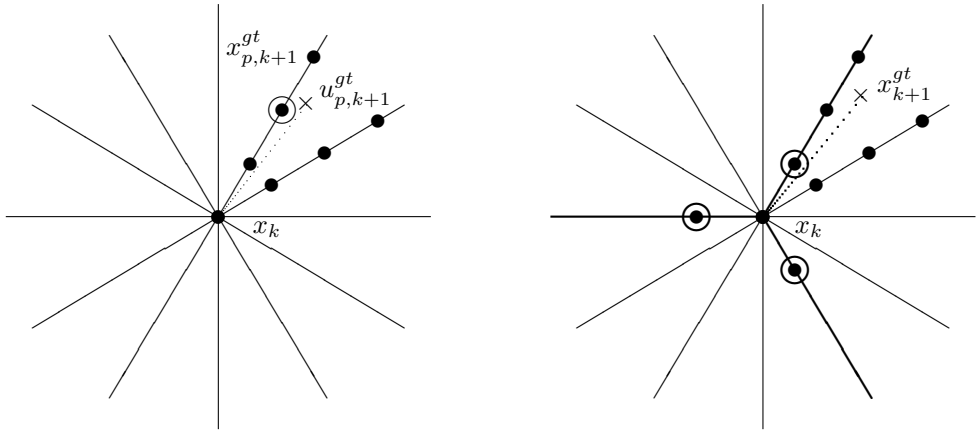


FIG. 3.2. *Search step (left) and poll step (right).*

In the poll step, the geometrical transformation technique defines the positive basis $V_k(x_k)$, which in turn defines the mesh neighborhood $\mathcal{N}(x_k)$. The procedure is described below and depicted in Figure 3.2.

POLL STEP: CHOICE OF MESH NEIGHBORHOOD $\mathcal{N}(x_k)$.

1. Compute one trial point x^{gt}_{k+1} by a geometrical transformation.
2. Determine v^{gt}_k in $\mathcal{V} = [\,U_1[v_1 \cdots v_{n+1}] \;\cdots\; U_n[v_1 \cdots v_{n+1}] \;\; -U_1[v_1 \cdots v_{n+1}] \;\cdots\; -U_n[v_1 \cdots v_{n+1}]\,]$ such that

$$\frac{\langle x^{gt}_{k+1} - x_k,\, v^{gt}_k \rangle}{\|x^{gt}_{k+1} - x_k\|} \;=\; \max_{v \in \mathcal{V}} \frac{\langle x^{gt}_{k+1} - x_k,\, v \rangle}{\|x^{gt}_{k+1} - x_k\|}.$$

3. Set $V_k(x_k)$ to the positive basis in $\mathcal{V}$ that contains v^{gt}_k, and then set $\mathcal{N}(x_k) = \{x_k + \Delta_k v : \text{for all } v \in V_k(x_k)\}$.

POLL STEP: EVALUATION OF f IN THE MESH NEIGHBORHOOD $\mathcal{N}(x_k)$.

4. List the points in $\mathcal{N}(x_k)$ by increasing order of the values of the angles between $x^{gt}_{k+1} - x_k$ and the corresponding vectors in $V_k(x_k)$.
5. Following the list given above, divided in groups of N_p points, start evaluating in parallel the function f in $\mathcal{N}(x_k)$.
   Stop if a point $x_{k+1} \in \mathcal{N}(x_k)$ is found such that $f(x_{k+1}) < f(x_k)$. In this case go to step 3, expanding M_k (poll step and iteration are declared successful).
   If $f(x_k) \le f(x)$ for every x in the mesh neighborhood $\mathcal{N}(x_k)$, go to step 4, shrinking M_k (poll step and iteration are declared unsuccessful).

Mesh expansions and reductions could also be designed to take advantage of problem information obtained from geometrical transformations.

4. Numerical experiments. In order to define a pattern search method for molecular geometry we need to be more specific about the geometrical transformations. The simplest geometrical transformations used in our calculations were the uniform expansions and compressions of the cluster in the plane perpendicular to the l-axis (Figure A.1(b)). These deformations correspond to putting $c_1 = c_2 = 0$ in

(A.1) and setting $c_3 = 1.1$ for expansions and $c_3 = 0.9$ for compressions. For the linear stretches (Figure A.1(c)), c_3 was set to 1 and $c_2 = 0.1$ or $c_2 = -0.1$ (Figure A.1(c), top and bottom, respectively). The quadratic stretches were done using $c_3 = 1$, $c_2 = 0$, and $c_1 = 0.1$ (Figure A.1(d), top) or $c_1 = -0.1$ (Figure A.1(d), bottom). The last deformation considered was the torsion of the cluster around the l-axis. This torsion was accomplished rotating atom α around the l-axis by an angle $\theta = c_2\, r_l^{\alpha,k} + c_3$, with c_2 and c_3 chosen so that the topmost atom would be rotated by $\pi/8$ clockwise (Figure A.1(e), right) or counter-clockwise (Figure A.1(e), left).

All the values mentioned above for c_1, c_2, and c_3 were the values used in the poll step of our pattern search methods. The search step should be much more aggressive than the poll step as an attempt for global search. To try to accomplish this goal, the parameters used in the search step were the poll step parameters, scaled by a factor of 5.

A random rearrangement of the atoms was also considered at every iteration, in an attempt to capture geometries very different from the current one. During the poll step, these rearrangements consisted of multiplying each coordinate of the atoms by a random value between 0.9 and 1.1; i.e., whenever a random deformation was performed, the $3N - 6$ coordinates of the cluster were scaled by a set of $3N - 6$ random values between 0.9 and 1.1. The random scaling factors used in the search step were between 0.5 and 1.5.

The mesh used in our implementation is defined by (2.1) with the choices of $\mathcal{V}$ and W respectively given by (3.8) and (2.2). The set of positive bases has $m = 2n$ uniform positive bases each with $n + 1$ vectors. To ensure that all deformations are tried in the search steps, the set of N_p deformations used is changed in a consistent way in consecutive search steps.

The stopping criterion used in our pattern search method followed the one implemented in PDS:

$$\sqrt{\frac{2(n-1)}{n}}\frac{\Delta_k}{\max\{1, \|x_k\|\}} \leq 10^{-2},$$

where $\sqrt{2(n-1)/n}\Delta_k$ is the length of the longest edge in the simplex defined in the current poll step by the corresponding uniform positive basis.

We applied our pattern search methods (PSM:MGP) to the minimization of the total energy of sodium clusters of dimension 4, 8, 16, and 32. The calculation of the total energy followed the process summarized in Appendix B. Results are given in Tables 4.1, 4.2, 4.3, and 4.4. We provide results for eight initial points, except for Na32, for which we present only two initial points. We list in these tables the number of iterations (iters), the number of total energy function evaluations (fevals), and the best value of the total energy found (f). The calculations were done in a cluster of twelve 2.266 GHz Intel Pentium IV personal computers connected through a switched full-duplex 100 Mb/s ethernet network, running LINUX, and using the message passing interface (MPI) as the parallelization protocol. We point out once again that we are dealing with expensive function evaluations: one evaluation of the total energy for the Na4 (resp., Na8, Na16, and Na32) cluster took on average 16 (resp., 59, 114, and 186) seconds of CPU time.

These preliminary results show that the method PSM:MGP is able to find a configuration nearly optimal for a significant number of initial points. The optimal value is approximately -1.698 for the Na4 cluster and -3.534 for the Na8 cluster, but these values are only attained after applying a local optimization code. The Na16

TABLE 4.1
Numerical results obtained by PSM:MGP for Na4. The numbers of processors used was $N_p = 12$.

x_0	iters	fevals	f
Na4a	24	397	-1.689
Na4b	38	589	-1.697
Na4c	42	711	-1.685
Na4d	12	193	-1.698
Na4e	16	262	-1.696
Na4f	20	312	-1.697
Na4g	28	444	-1.682
Na4h	52	875	-1.694

TABLE 4.2
Numerical results obtained by PSM:MGP for Na8. The numbers of processors used was $N_p = 12$.

x_0	iters	fevals	f
Na8a	125	3385	-3.524
Na8b	105	2913	-3.522
Na8c	101	2748	-3.528
Na8d	73	1783	-3.467
Na8e	11	293	-3.504
Na8f	169	4558	-3.521
Na8g	108	2502	-3.489
Na8h	133	3305	-3.515

TABLE 4.3
Numerical results obtained by PSM:MGP for Na16. The numbers of processors used was $N_p = 12$.

x_0	iters	fevals	f
Na16a	260	11195	-7.119
Na16b	285	11740	-7.147
Na16c	283	12091	-7.135
Na16d	371	15783	-7.159
Na16e	239	10373	-7.136
Na16f	245	10268	-7.143
Na16g	255	10846	-7.122
Na16h	304	13134	-7.138

TABLE 4.4
Numerical results obtained by PSM:MGP for Na32. The numbers of processors used was $N_p = 12$.

x_0	iters	fevals	f
Na32a	462	33417	-14.558
Na32b	477	34509	-14.580

and Na32 cluster geometries are not well established.

Due to limited access to our cluster, we are unable to provide the full results for Na32. We were able to finish only two of the runs, for two given initial configurations, which terminated with 33417 and 33509 function evaluations, respectively. We used this information, however, in a derivation of an estimate for the rate of growth in the number of (average) function evaluations in terms of the number of variables $n = 3N - 6$. The number of function evaluations (*fevals*) seems to grow with n under a rate slower than quadratic ($fevals \approx n^{1.6}$). A fit of $\log(fevals)$ to $A + B \log n$ has

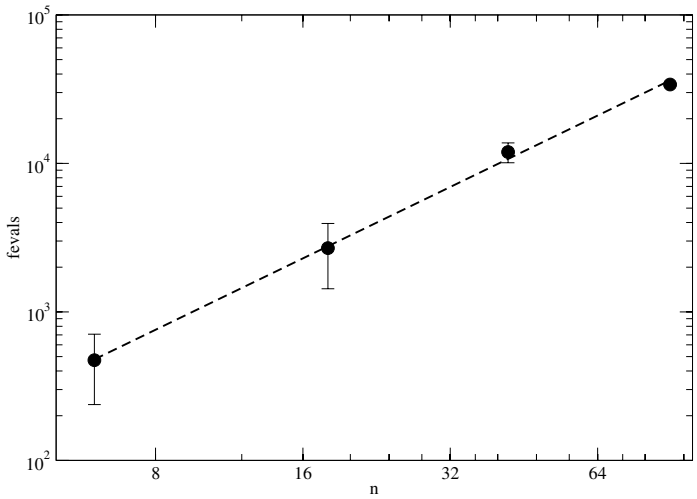


FIG. 4.1. *Scaling of the average number of function evaluations (fevals) with the number of variables (n). Dashed line is a fit of a power law, fevals* $= An^B$*, to the data (*$A = 27.336$ *and* $B = 1.5977$*).*

TABLE 4.5

*Numerical results obtained by PDS (with twelve pattern points) and PSM:MGP for Na*4*. The computation of the total energy was carried out differently from that reported in Tables* 4.1–4.4*. The numbers of processors used was* $N_p = 12$*.*

	PDS			PSM:MGP			
x_0	iters	fevals	f	iters	fevals	f	best
Na4i	26	324	-2.58191	28	533	-2.58251	PSM:MGP
Na4j	9	120	-2.58751	18	343	-2.58772	PSM:MGP
Na4k	6	84	-2.58920	4	77	-2.58868	PDS
Na4l	7	96	-2.58860	4	77	-2.58862	PSM:MGP
Na4m	25	312	-2.58098	32	609	-2.57487	PDS

given $A = 27.336$ and $B = 1.5977$, and the least squares regression error was 0.008322 (see Figure 4.1).

For the sake of comparison with other methods, we ran PSM:MGP and PDS for another set of initial configurations for Na4 and Na8. These calculations were done in a different computer system (a cluster of 24 DIGITAL/Compaq Alpha 500au Personal Workstations connected through a switched full-duplex 100 Mb/s ethernet network, running DIGITAL UNIX, and using MPI as the parallelization protocol) and with different parameters for the plane-wave code (corresponding to a different version of the application code, slower due to the use of a more accurate model for the electron-ion interactions). As a result, the total energy values presented now are not comparable to the ones reported above. A simple comparison of the total energy values obtained with both methods shows that the comparison between PDS and PSM:MGP is mildly favorable to the latter, as we indicated in the last column of Tables 4.5 and 4.6. (In one instance the values of the objective function coincided and we used as a second criterion the number of function evaluations.) Both PDS and PSM:MGP were able to find, for the Na8 cluster, the two best known local minimizers for different starting configurations.

We point out that the implementation of PSM:MGP used in these computations

TABLE 4.6

Numerical results obtained by PDS (with 36 pattern points) and PSM:MGP for Na8. The computation of the total energy was carried out differently from that reported in Tables 4.1–4.4. The numbers of processors used was $N_p = 10$.

	PDS			PSM:MGP			
x_0	iters	fevals	f	iters	fevals	f	best
Na8i	31	1135	-5.25543	35	868	-5.23153	PDS
Na8j	5	199	-5.31805	3	88	-5.31805	PSM:MGP
Na8k	24	883	-5.30026	58	1546	-5.31268	PSM:MGP
Na8l	3	118	-1.27962	3	88	-5.31022	PSM:MGP

is far from being exhaustively tuned. We did not play with the code to try to come up with the best strategies (geometrical transformations, etc.) and with the best values for the different parameters. We expect that a method like PSM:MGP has plenty of room for improvement.

5. Conclusions and future work. We designed a class of pattern search methods for molecular geometry by taking advantage of physically meaningful energy lowering geometrical transformations, and by combining them with appropriate patterns for minimization purposes. The preliminary numerical results obtained with a particular pattern search method in the class have indicated that this approach could lead to very promising algorithms for molecular geometry. We hope to obtain better numerical results by considering more elaborate search steps. In fact, our approach has the flexibility to incorporate several types of global optimization algorithms in the search step to enhance the selection of the geometrical transformations and their defining values. We have in mind, for instance, the use of evolutionary algorithms like evolutionary programming or evolutionary strategies.

The new pattern search methods can be used in any other context where there is a user-provided scheme to generate points leading to potential objective function decrease.

We plan to apply our pattern search methods to the total energy minimization of other clusters and to develop analogues of this approach in other molecular geometry contexts. We also plan to investigate patterns with similar interesting uniform directionality properties.

Appendix A. The current point x_k in the optimization process stores the atomic positions $r_i^{\alpha,k}$ of a set of N atoms, where k denotes the iteration counter and r_i^{α} is the ith coordinate of atom α $(i = 1, 2, 3)$. The set of atomic positions specifies not only the shape of the system of atoms but also its location and orientation in space. Since shapes that result from translations or rotations about a fixed point have the same energy, there are six redundant coordinates in a molecular geometry optimization process. Three of these refer to the location of the set of atoms with a certain shape in space and the other three are the angles that define the orientation of this set with respect to some fixed three-dimensional reference frame. The easiest way to get rid of these additional degrees of freedom is to fix one of the atoms at the origin of a three-dimensional reference frame, to keep another atom on one of the axes of this frame (the x-axis, for example), and to force a third atom to move only on a plane containing the above mentioned axis (the xy-plane, for example). These restrictions do not introduce constraints in shape space, they merely exclude atomic configurations representing the same system translated and/or rotated in space. Without loss of

generality, the three constrained atoms are chosen to be atoms N, $N-1$, and $N-2$. The vector x_k can then be related to the atomic positions, $r_i^{\alpha,k}$, in the following way:

$$x_k = \begin{bmatrix} r_1^{1,k} \\ r_2^{1,k} \\ r_3^{1,k} \\ r_1^{2,k} \\ \vdots \\ r_3^{N-3,k} \\ r_1^{N-2,k} \\ r_2^{N-2,k} \\ r_1^{N-1,k} \end{bmatrix}.$$

The corresponding "constraints" are:

$$r_3^{N-2,k} = r_2^{N-1,k} = r_3^{N-1,k} = r_i^{N,k} = 0, \qquad i = 1, 2, 3.$$

An optimization step, $x_k \to x_{k+1}$, can be viewed as a deformation of the molecular shape described by x_k. This deformation may not have any physical meaning, corresponding simply to a random rearrangement of the atoms. The space spanned by an algorithm where only this type of move is present is unrelated to shape space; i.e., a given path in this space is not related in a simple way to a shape space path, a path where the molecule undergoes a recognizable shape transformation. Physically meaningful deformations (as, for example, a simple uniform compression or expansion of the molecule), i.e., paths in shape space, are expected to be closer to (total energy) downhill directions than simple paths in x_k-space. In fact, a path in shape space will in general correspond to a nontrivial path in x_k-space that can even connect very distant x_k-space points.

A simple way to introduce physically meaningful and energy lowering deformations of a given molecule or cluster is to consider just stretches and twists along some direction. An obvious choice for the directions along which the cluster is to be stretched or twisted is its principal axes system.[3] In order to deform the molecule in this way it is necessary to refer the atomic positions to the principal axes system:

$$\bar{r}_i^{\alpha,k} = \sum_{j=1}^{3} \mathcal{R}_{ij}^{(k)} r_j^{\alpha,k},$$

where $\mathcal{R}^{(k)}$ rotates the reference axes to the principal axes. The deformations of the molecule can then be written as

$$\bar{r}_i^{\alpha,k+1} = \sum_{j=1}^{3} \epsilon_{ij}^{\alpha,k} \bar{r}_j^{\alpha,k}$$

[3]The principal or inertial axes system of a given molecule is the set of eigenvectors of the matrix

$$\mathcal{I}_{ij} = \sum_{\alpha=1}^{N} m_\alpha \left(\|r^\alpha\|^2 \delta_{ij} - r_i^\alpha r_j^\alpha \right),$$

where δ_{ij} is the Kronecker tensor and m_α is the mass of atom α. For convenience, we choose a reference frame whose origin is the center of mass of the molecule, i.e., where the atomic coordinates satisfy the relation $\sum_{\alpha=1}^{N} m_\alpha r^\alpha = 0$.

or, returning to the nonprincipal axes system, as

$$r_i^{\alpha,k+1} = \sum_{j,l,m=1}^{3} \mathcal{R}_{ij}^{(k)-1} \epsilon_{jl}^{\alpha,k} \mathcal{R}_{lm}^{(k)} r_m^{\alpha,k}.$$

Alternatively, using matrix notation, we can write

$$r^{\alpha,k+1} = \mathcal{R}^{(k)-1} \epsilon^{\alpha,k} \mathcal{R}^{(k)}\, r^{\alpha,k}.$$

The form for the deformations assumed above is very broad. Some simple and physically meaningful particular forms can be written simply as

$$\epsilon_{ij}^{\alpha,k} = \big[\big(c_1\,(\bar{r}_l^{\alpha,k})^2 + c_2\,\bar{r}_l^{\alpha,k} + c_3 - 1\big)(1-\delta_{jl}) + 1\big]\delta_{ij}, \tag{A.1}$$

where $l \in \{1,2,3\}$ is the label of the principal axes about which the transformations are made. The effect of these transformations is simply to put the atoms closer or farther from the principal axis l (see Figures A.1(a)–A.1(d)).

Another physically meaningful deformation that can be considered is a torsion of the molecule around some axis (Figure A.1(e)), in a way that forces different parts of the molecule to be rotated around that axis with different angles:

$$\epsilon^{\alpha,k} = \mathcal{R}_{\hat{e}_l}(\theta(\bar{r}_l^{\alpha,k})). \tag{A.2}$$

The axis $l \in \{1,2,3\}$ is the torsion axis, and $\mathcal{R}_{\hat{e}_l}(\theta)$ is a rotation by an angle θ around that axis. For example,

$$\mathcal{R}_{\hat{e}_3}(\theta) = \begin{bmatrix} \cos\theta & \sin\theta & 0 \\ -\sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

The angle θ must be a function of the l-coordinate of each atom (a quadratic function of $r_l^{\alpha,k}$, for example). Contrary to the previous forms, this type of deformation can break any axial symmetry that the molecule at iteration k might possess.

As we said before, the geometrical transformations (A.1) are performed with the center of mass of the cluster at the origin of the $\bar{r}^{\alpha,k}$ coordinates. Thus, an atom sitting on the plane containing the center of mass and perpendicular to the l-axis of this system of coordinates—where these geometrical transformations are performed—would remain unaffected by most deformations (see Figures A.1(c)–A.1(e)). The exceptions are the uniform expansions and compressions (see Figure A.1(b)).

A set of new coordinates $r_i^{\alpha,k+1}$ computed by geometrical transformation from the previous coordinates $r_i^{\alpha,k}$ (stored in x_k) can then be used as a trial point x_{k+1}^{gt} for the search and poll steps of the $k+1$ pattern search iteration.

Appendix B. We will provide a brief description of the main issues in local density plane-wave total energy calculation [18]. The Hamiltonian $\mathbf{H}$ of an N-electron system with M nuclei of charge Z_I and mass m_I can be written as

$$\mathbf{H}(r_1,\ldots,r_N,R_1,\ldots,R_M) = \sum_{i=1}^{N} \frac{\mathbf{p}_i^2}{2} + \sum_{I=1}^{M} \frac{\mathbf{P}_I^2}{2m_I} - \sum_{i,I=1}^{N,M} \frac{Z_I}{|r_i - R_I|} + \sum_{\substack{i,j=1 \\ i<j}}^{N} \frac{1}{|r_i - r_j|} + \sum_{\substack{I,J=1 \\ I<J}}^{M} \frac{Z_I Z_J}{|R_I - R_J|},$$

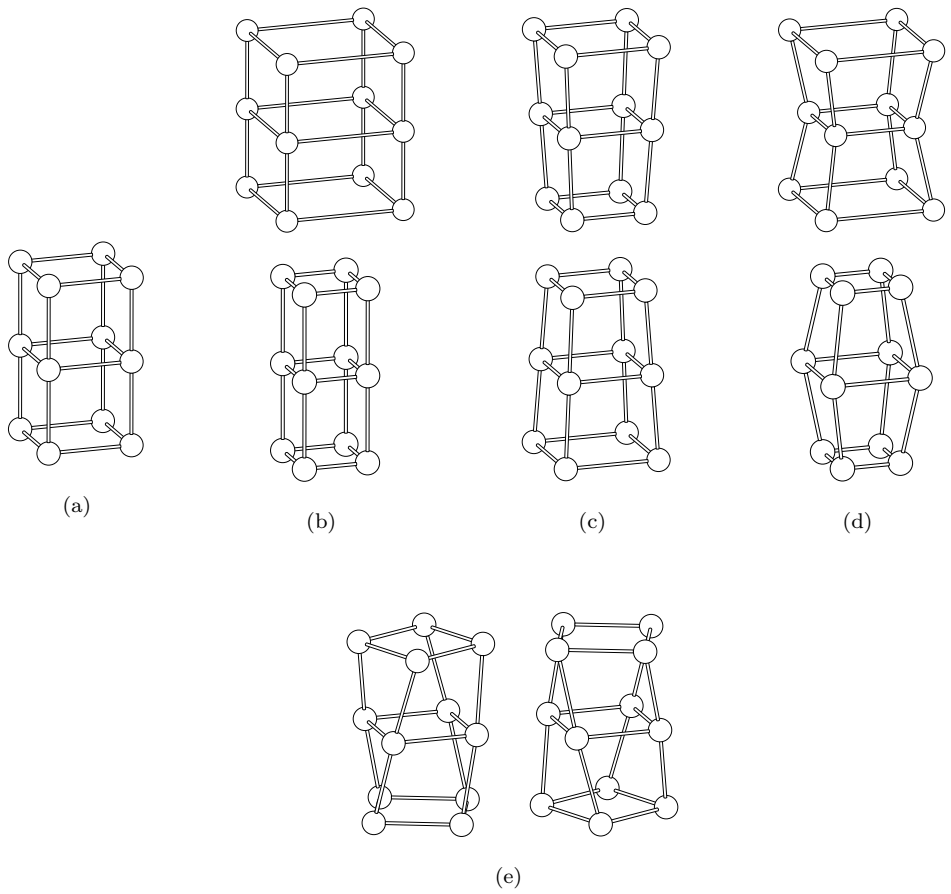


FIG. A.1. *Simple example of the deformations* (A.1) *and* (A.2). *In* (b)*, the reference molecule seen in* (a) *is expanded sideways, which corresponds to setting* $c_1 = c_2 = 0$*,* $c_3 \neq 0$ *in* (A.1)*. The parameter* c_3 *can be greater (top) or lower (bottom) than* 1*. Setting* $c_2 \neq 0$ *results in deformations similar to those in* (c) *(with* c_2 *positive (top) or negative (bottom)), while the use of a full quadratic form gives rise to deformations like those in* (d) *(with* c_1 *positive (top) or negative (bottom)). (In this example,* $\theta = c_2\, r_l^{\alpha,k} + c_3$*, with* c_2 *positive (left) or negative (right).) Panel* (e) *is an example of the deformations that can be achieved with* (A.2)*. In all these examples, the* l*-axis is the vertical axis.*

where r_i and R_I are the coordinates of the electrons and of the atomic nuclei, and $\mathbf{p}_i$ and $\mathbf{P}_I$ are their linear momenta. (Spin was not considered, for simplicity; atomic units are used throughout the calculations.)

By solving the time-independent Schrödinger equation

$$\mathbf{H}\Psi = E\Psi,$$

one obtains the set of eigenvalues (energies, E) and eigenvectors (wavefunctions, Ψ) of the system. This equation gives a good description of nonrelativistic many-electron systems subject to electric fields produced by atomic nuclei, like atoms, molecules, and solids. However, this equation is in general unsolvable. If the mass difference between electrons and nuclei is taken into account [4], the time-independent Schrödinger

equation can be separated in two equations: one for the electrons

$$\left(\sum_{i=1}^{N}\frac{\mathbf{p}_i^2}{2}-\sum_{i,I=1}^{N,M}\frac{Z_I}{|r_i-R_I|}+\sum_{\substack{i,j=1\\ i<j}}^{N}\frac{1}{|r_i-r_j|}+E_{\rm nn}\right)\Psi(r_1,\ldots,r_N;R_1,\ldots,R_M) = E(R_1,\ldots,R_M)\,\Psi(r_1,\ldots,r_N;R_1,\ldots,R_M), \tag{B.1}$$

where

$$E_{\rm nn}=\sum_{\substack{I,J=1\\ I<J}}^{M}\frac{Z_I Z_J}{|R_I-R_J|},$$

and another for the nuclei, of no interest in this context.

In (B.1), the nuclear coordinates R_i are just parameters, and the electronic wavefunctions and eigenvalues are different for each arrangement of nuclei. In order to find the lowest energy state of the system (the ground state), one can solve (B.1) for a given set of nuclear coordinates, and assume that $E(R_1,\ldots,R_M)$ is a function of the nuclear coordinates to be subsequently minimized.

Hohenberg and Kohn [15] proved a theorem that legitimizes the use of the electronic density

$$\rho(r)=N\int|\Psi(r,r_2,\ldots,r_N)|^2\,dr_2\ldots dr_N$$

as fundamental variable, instead of the wavefunction $\Psi(r_1,\ldots,r_N)$. The theorem states that any observable (e.g., the energy) is a functional of the ground state density. In particular, the ground state energy functional of an N-electron system in an external potential $v_{\rm ext}(r)$ (representing the interaction of the nuclei with the electrons, for example) can be written as

$$E_{v_{\rm ext}}[\rho]=F_{\rm HK}[\rho]+\int\rho(r)v_{\rm ext}(r)dr+E_{\rm nn}$$

where $F_{\rm HK}[\rho]$ is a universal functional, i.e., a functional that does not depend on the external potential. Therefore, $F_{\rm HK}[\rho]$ is the same for atoms, molecules, and solids. The ground state is obtained through the variational principle

$$E_*=\min_{\{\rho\}}\ E_{v_{\rm ext}}[\rho], \tag{B.2}$$

and the variational search is performed over all the admissible electronic densities.

A good approximation to the functional $F_{\rm HK}[\rho]$ was suggested by Kohn and Sham [19]. Their main hypothesis is that, for each interacting ground state density $\rho(r)$, there exists a noninteracting electron system with the same ground state density. The Kohn–Sham $F_{\rm HK}[\rho(r)]$ functional is

$$F_{\rm HK}[\rho(r)]=-\frac{1}{2}\sum_{i=1}^{N}\int\phi_i^*(r)\nabla^2\phi_i(r)dr+\frac{1}{2}\int\frac{\rho(r_1)\rho(r_2)}{|r_1-r_2|}dr_1dr_2+E_{xc}[\rho(r)],$$

with

$$\sum_{i=1}^{N}|\phi_i(r)|^2=\rho(r). \tag{B.3}$$

$E_{xc}[\rho(r)]$ is the so-called exchange and correlation functional, for which many approximations exist [7, 24, 25].

The ground state is obtained solving the Euler–Lagrange equation that results from the minimization (B.2):

$$\left[-\frac{1}{2}\nabla^2 + v_{\text{ext}}(r) + \int \frac{\rho(r)}{|r-r'|}dr' + \frac{\delta E_{xc}\left[\rho(r)\right]}{\delta\rho(r)}\right]\phi_i(r) = \epsilon_i\phi_i(r), \tag{B.4}$$

and the total energy of the system is therefore

$$E_{\text{KS}}[\rho(r)] = -\frac{1}{2}\sum_{i=1}^{N}\int \phi_i^*(r)\nabla^2\phi_i(r)dr + \frac{1}{2}\int \frac{\rho(r_1)\rho(r_2)}{|r_1-r_2|}dr_1dr_2 + E_{xc}[\rho(r)] + \int \rho(r)v_{\text{ext}}(r)dr + E_{\text{nn}}.$$

The coupled nonlinear equations (B.3)–(B.4) are the so-called Kohn–Sham equations.

To calculate the total energy of solids, a plane-wave expansion of the Kohn–Sham wavefunctions is very useful, as it takes advantage of the translation symmetry of the crystal [17, 18, 26]. For finite systems, such as atoms, molecules, and clusters, plane-waves can also be used in a supercell approach. In the supercell method, the finite system is placed in a unit cell of a fictitious crystal, and this cell is made large enough to avoid interactions between neighboring cells. However, for finite systems a very large number of plane waves is needed, as the electronic density spans only a small fraction of the total volume of the supercell. The plane-wave expansion of the wavefunctions amounts simply to Fourier transforming them and all the other quantities involved in the Kohn–Sham equations, thereby converting the differential equation (B.4) into a matrix diagonalization problem. For finite systems, as many plane waves are needed, this matrix is very large, on the order of hundreds for small clusters.

But even for extended systems, many plane-waves may be needed. The valence wavefunctions of the large Z_I atoms oscillate strongly in the vicinity of the atomic core, due to the orthogonalization to the inner electronic wavefunctions. To describe these oscillations a large number of plane-waves is required, making even more difficult the calculation of the total energy. However, the inner electrons are almost inert and are not significantly involved in bonding. This makes possible the description of an atom based solely on its valence electrons, which feel an effective potential that includes both the nuclear attraction and the repulsion of the inner electrons. This technique is the so-called pseudopotential approximation. In this work, we used the Troullier–Martins pseudopotential [31].

Although the pseudopotential approximation reduces its computational burden, the calculation of the total energy of a given system in the manner outlined above is still a very demanding task. One can deal with systems with at most a few hundred atoms. There are other methods that are significantly faster, allowing the calculation of the total energy of systems consisting of thousands of atoms. But these methods are much less accurate than the density functional method presented above. There are also some methods more accurate than this one, but they are significantly harder, prohibiting the simulation of systems with more than a few atoms.

Acknowledgments. We would like to thank Charles Audet and John Dennis for their insightful comments. We would also like to thank one referee and the associate editor for their suggestions that led to several improvements.

REFERENCES

[1] P. Alberto, F. Nogueira, H. Rocha, and L. N. Vicente, *Pattern search methods for user-provided points*, in Computational Science—ICCS 2001, Lecture Notes in Comput. Sci. 2074, V. N. Alexandrov, J. J. Dongarra, B. A. Juliano, R. S. Renner, and C. J. Kenneth Tan, eds., Springer-Verlag, Berlin, 2001, pp. 95–98.

[2] C. Audet and J. E. Dennis, Jr., *Pattern search algorithms for mixed variable programming*, SIAM J. Optim., 11 (2000), pp. 573–594.

[3] R. Biswas and D. R. Hamann, *Simulated annealing of silicon atom clusters in Langevin molecular dynamics*, Phys. Rev. B, 34 (1986), pp. 895–901.

[4] M. Born and K. Huang, *Dynamical Theory of Crystal Lattices*, Clarendon Press, Oxford, 1954.

[5] G. E. P. Box, *Evolutionary operation: A method for increasing industrial productivity*, Appl. Statist., 6 (1957), pp. 81–101.

[6] R. Car and M. Parrinello, *Unified approach for molecular dynamics and density-functional techniques*, Phys. Rev. Lett., 55 (1985), pp. 2471–2474.

[7] D. M. Ceperley and B. J. Alder, *Ground state of the electron gas by a stochastic method*, Phys. Rev. Lett., 45 (1980), pp. 566–569.

[8] A. R. Conn, K. Scheinberg, and P. L. Toint, *Recent progress in unconstrained nonlinear optimization without derivatives*, Math. Programming, 79 (1997), pp. 397–414.

[9] C. Davis, *Theory of positive linear dependence*, Amer. J. Math., 76 (1954), pp. 733–746.

[10] J. E. Dennis, Jr., and V. Torczon, *Direct search methods on parallel machines*, SIAM J. Optim., 1 (1991), pp. 448–474.

[11] G. H. Golub and C. F. Van Loan, *Matrix Computations*, 3rd ed., The John Hopkins University Press, Baltimore, MD, 1996.

[12] W. E. Hart, *A generalized stationary point convergence theory for evolutionary algorithms*, in Proceedings of the 7th International Conference on Genetic Algorithms, East Lansing, MI, 1997, Thomas Bäck, ed., Morgan Kaufmann, San Francisco, 1997, pp. 127–134.

[13] W. E. Hart, *Comparing evolutionary programs and evolutionary pattern search algorithms: A drug docking application*, in Proceedings of the Genetic and Evolutionary Computation Conference, Orlando, FL, 1999, Morgan Kaufmann, San Francisco, 1999, pp. 855–862.

[14] M. R. Hoare and J. A. McInnes, *Morphology and statistical statics of simple microclusters*, Adv. Phys., 32 (1983), pp. 791–821.

[15] P. Hohenberg and W. Kohn, *Inhomogeneous electron gas*, Phys. Rev. B, 136 (1964), pp. 864–871.

[16] R. Hooke and T. A. Jeeves, *"Direct search" solution of numerical and statistical problems*, J. ACM, 8 (1961), pp. 212–229.

[17] J. Ihm, *Total energy calculations in solid state physics*, Rep. Prog. Phys., 51 (1988), pp. 105–142.

[18] J. Ihm, A. Zunger, and M. L. Cohen, *Momentum-space formalism for the total energy of solids*, J. Phys. C: Solid State Phys., 12 (1979), pp. 4409–4422; *Errata*, J. Phys. C: Solid State Phys., 13 (1980), p. 3095.

[19] W. Kohn and L. J. Sham, *Self-consistent equations including exchange and correlation effects*, Phys. Rev. A, 140 (1965), pp. 1133–1138.

[20] R. M. Lewis and V. Torczon, *Rank Ordering and Positive Bases in Pattern Search Algorithms*, Technical report TR96-71, ICASE, NASA Langley Research Center, Hampton, VA, 1999.

[21] R. M. Lewis, V. Torczon, and M. W. Trosset, *Why pattern search works*, OPTIMA, MPS Newsletter, 59 (1998), pp. 1–7.

[22] J. C. Meza and M. L. Martinez, *On the use of direct search methods for the molecular conformation problem*, J. Comput. Chem., 15 (1994), pp. 627–632.

[23] J. A. Nelder and R. Mead, *A simplex method for function minimization*, Comput. J., 7 (1965), pp. 308–313.

[24] J. P. Perdew and Y. Wang, *Accurate and simple analytic representation of the electron-gas correlation energy*, Phys. Rev. B, 45 (1992), pp. 13244–13249.

[25] J. P. Perdew and A. Zunger, *Self-interaction correction to density-functional approximations for many-electron systems*, Phys. Rev. B, 23 (1981), pp. 5048–5079.

[26] W. E. Pickett, *Pseudopotential methods in condensed matter applications*, Comp. Phys. Rep., 9 (1989), pp. 115–198.

[27] M. J. D. Powell, *Direct search algorithms for optimization calculations*, in Acta Numer. 7, Cambridge University Press, Cambridge, UK, 1998, pp. 287–336.

[28] V. Torczon, *On the convergence of the multidirectional search algorithm*, SIAM J. Optim., 1 (1991), pp. 123–145.

[29] V. Torczon, *On the convergence of pattern search algorithms*, SIAM J. Optim., 7 (1997), pp. 1–25.
[30] M. W. Trosset, *I know it when I see it: Toward a definition of direct search methods*, SIAG/OPT Views-and-News, 9 (1997), pp. 7–10.
[31] N. Troullier and J. L. Martins, *Efficient pseudopotentials for plane-wave calculations*, Phys. Rev. B, 43 (1991), pp. 1993–2006.
[32] M. H. Wright, *Direct search methods: Once scorned, now respectable*, in Numerical Analysis 1995 (Proceedings of the 1995 Dundee Biennial Conference in Numerical Analysis), Pitman Res. Notes Math. Ser. 344, D. F. Griffiths and G. A. Watson, eds., CRC Press, Boca Raton, FL, 1996, pp. 191–208.
[33] W. Zhong, G. Overney, and D. Tománek, *Structural properties of Fe crystals*, Phys. Rev. B, 47 (1993), pp. 95–99.

SIAM J. OPTIM.
Vol. 14, No. 4, pp. 1237–1249

ON A CLASS OF MINIMAX STOCHASTIC PROGRAMS*

ALEXANDER SHAPIRO† AND SHABBIR AHMED†

Abstract. For a particular class of minimax stochastic programming models, we show that the problem can be equivalently reformulated into a standard stochastic programming problem. This permits the direct use of standard decomposition and sampling methods developed for stochastic programming. We also show that this class of minimax stochastic programs is closely related to a large family of mean-risk stochastic programs where risk is measured in terms of deviations from a quantile.

Key words. worst case distribution, problem of moments, Lagrangian duality, mean-risk stochastic programs, deviation from a quantile

AMS subject classifications. 90C15, 90C47

DOI. 10.1137/S1052623403434012

1. Introduction.

A wide variety of decision problems under uncertainty involve optimization of an expectation functional. An abstract formulation for such stochastic programming problems is

$$\operatorname*{Min}_{x\in X} \mathbb{E}_P[F(x,\omega)], \tag{1.1}$$

where $X \subseteq \mathbb{R}^n$ is the set of feasible decisions, $F : \mathbb{R}^n \times \Omega \mapsto \mathbb{R}$ is the objective function, and P is a probability measure (distribution) on the space Ω equipped with a sigma algebra $\mathcal{F}$. The stochastic program (1.1) has been studied in great detail, and significant theoretical and computational progress has been achieved (see, e.g., [18] and references therein).

In the stochastic program (1.1) the expectation is taken with respect to the probability distribution P which is assumed to be *known*. However, in practical applications, such a distribution is not known precisely and has to be estimated from data or constructed using subjective judgments. Often, the available information is insufficient to identify a unique distribution. In the absence of full information on the underlying distribution, an alternative approach is as follows. Suppose a set $\mathcal{P}$ of possible probability distributions for the uncertain parameters is known; then it is natural to optimize the expectation functional (1.1) corresponding to the "worst" distribution in $\mathcal{P}$. This leads to the following minimax stochastic program:

$$\operatorname*{Min}_{x\in X} \left\{ f(x) := \sup_{P\in\mathcal{P}} \mathbb{E}_P[F(x,\omega)] \right\}. \tag{1.2}$$

Theoretical properties of minimax stochastic programs have been studied in a number of publications. In that respect we can mention pioneering works of Žáčková [22] and Dupačová [3, 4]. Duality properties of minimax stochastic programs were thoroughly studied in Klein Haneveld [10]; for more recent publications see [19] and references therein. These problems have also received considerable attention in the

*Received by the editors August 29, 2003; accepted for publication (in revised form) February 7, 2004; published electronically August 4, 2004.
http://www.siam.org/journals/siopt/14-4/43401.html
†School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA 30332-0205 (ashapiro@isye.gatech.edu, sahmed@isye.gatech.edu). The research of the second author was supported by the National Science Foundation under grant DMI-0133943.

context of bounding and approximating stochastic programs [1, 7, 9]. A number of authors have proposed numerical methods for minimax stochastic programs. Ermoliev, Gaivoronski, and Nedeva [5] proposed a method based on the stochastic quasi-gradient algorithm and generalized linear programming. A similar approach along with computational experience is reported in [6]. Breton and El Hachem [2] developed algorithms based on bundle methods and subgradient optimization. Riis and Andersen [16] proposed a cutting plane algorithm. Takriti and Ahmed [21] considered minimax stochastic programs with binary decision variables arising in power auctioning applications, and developed a branch-and-cut scheme. All of the above numerical methods require explicit solution of the inner optimization problem $\sup_{P\in\mathcal{P}} \mathbb{E}_P[F(x,\omega)]$ corresponding to the candidate solution x in each iteration. Consequently, such approaches are inapplicable in situations where calculation of the respective expectations numerically is infeasible because the set Ω although finite is prohibitively large, or possibly infinite.

In this paper, we show that a fairly general class of minimax stochastic programs can be equivalently reformulated into standard stochastic programs (involving optimization of expectation functionals). This permits a direct application of powerful decomposition and sampling methods that have been developed for standard stochastic programs in order to solve large-scale minimax stochastic programs. Furthermore, the considered class of minimax stochastic programs is shown to subsume a large family of mean-risk stochastic programs, where the risk is measured in terms of deviations from a quantile.

2. The problem of moments. In this section we discuss a variant of the problem of moments. This will provide us with basic tools for the subsequent analysis of minimax stochastic programs.

Let us denote by $\mathcal{X}$ the (linear) space of all finite signed measures on $(\Omega,\mathcal{F})$. We say that a measure $\mu\in\mathcal{X}$ is nonnegative, and write $\mu\succeq 0$, if $\mu(A)\geq 0$ for any $A\in\mathcal{F}$. For two measures $\mu_1,\mu_2\in\mathcal{X}$ we write $\mu_2\succeq\mu_1$ if $\mu_2-\mu_1\succeq 0$. That is, $\mu_2\succeq\mu_1$ if $\mu_2(A)\geq\mu_1(A)$ for any $A\in\mathcal{F}$. It is said that $\mu\in\mathcal{X}$ is a *probability* measure if $\mu\succeq 0$ and $\mu(\Omega)=1$. For given nonnegative measures $\mu_1,\mu_2\in\mathcal{X}$ consider the set

$$\mathcal{M}:=\{\mu\in\mathcal{X}:\mu_1\preceq\mu\preceq\mu_2\}. \tag{2.1}$$

Let $\varphi_i(\omega)$, $i=0,\ldots,q$, be real valued measurable functions on $(\Omega,\mathcal{F})$ and $b_i\in\mathbb{R}$, $i=1,\ldots,q$, be given numbers. Consider the problem

$$\begin{array}{ll} \mathrm{Max}_{P\in\mathcal{M}} & \int_\Omega \varphi_0(\omega)dP(\omega) \\ \text{subject to} & \int_\Omega dP(\omega)=1, \\ & \int_\Omega \varphi_i(\omega)dP(\omega)=b_i, \quad i=1,\ldots,r, \\ & \int_\Omega \varphi_i(\omega)dP(\omega)\leq b_i, \quad i=r+1,\ldots,q. \end{array} \tag{2.2}$$

In the above problem, the first constraint implies that the optimization is performed over probability measures, the next two constraints represent moment restrictions, and the set $\mathcal{M}$ represents upper and lower bounds on the considered measures. If the constraint $P\in\mathcal{M}$ is replaced by the constraint $P\succeq 0$, then the above problem (2.2) becomes the classical problem of moments (see, e.g., [13], [20], and references therein). As we shall see, however, the introduction of lower and upper bounds on the considered measures makes the above problem more suitable for an application to minimax stochastic programming.

We make the following assumptions throughout this section:

(A1) The functions $\varphi_i(\omega)$, $i = 0, \ldots, q$, are μ_2-integrable; i.e.,

$$\int_\Omega |\varphi_i(\omega)| d\mu_2(\omega) < \infty, \quad i = 0, \ldots, q.$$

(A2) The feasible set of problem (2.2) is nonempty, and, moreover, there exists a probability measure $P^* \in \mathcal{M}$ satisfying the equality constraints as well as the inequality constraints as equalities, i.e.,

$$\int_\Omega \varphi_i(\omega) dP^*(\omega) = b_i, \quad i = 1, \ldots, q.$$

Assumption (A1) implies that $\varphi_i(\omega)$, $i = 0, \ldots, q$, are P-integrable with respect to all measures $P \in \mathcal{M}$, and hence problem (2.2) is well defined. By assumption (A2), we can make the following change of variables $P = P^* + \mu$, and hence write problem (2.2) in the form

$$\begin{array}{ll} \mathrm{Max}_{\mu \in \mathcal{M}^*} & \int_\Omega \varphi_0(\omega) dP^*(\omega) + \int_\Omega \varphi_0(\omega) d\mu(\omega) \\ \text{subject to} & \int_\Omega d\mu(\omega) = 0, \\ & \int_\Omega \varphi_i(\omega) d\mu(\omega) = 0, \quad i = 1, \ldots, r, \\ & \int_\Omega \varphi_i(\omega) d\mu(\omega) \le 0, \quad i = r+1, \ldots, q, \end{array} \tag{2.3}$$

where

$$\mathcal{M}^* := \{\mu \in \mathcal{X} : \mu_1^* \preceq \mu \preceq \mu_2^*\} \tag{2.4}$$

with $\mu_1^* := \mu_1 - P^*$ and $\mu_2^* := \mu_2 - P^*$.

The Lagrangian of problem (2.3) is

$$L(\mu, \lambda) := \int_\Omega \varphi_0(\omega) dP^*(\omega) + \int_\Omega \mathcal{L}_\lambda(\omega) d\mu(\omega), \tag{2.5}$$

where

$$\mathcal{L}_\lambda(\omega) := \varphi_0(\omega) - \lambda_0 - \sum_{i=1}^q \lambda_i \varphi_i(\omega), \tag{2.6}$$

and the (Lagrangian) dual of (2.3) is

$$\begin{array}{ll} \mathrm{Min}_{\lambda \in \mathbb{R}^{q+1}} & \{\psi(\lambda) := \sup_{\mu \in \mathcal{M}^*} L(\mu, \lambda)\} \\ \text{subject to} & \lambda_i \ge 0, \quad i = r+1, \ldots, q. \end{array} \tag{2.7}$$

It is straightforward to see that

$$\psi(\lambda) = \int_\Omega \varphi_0(\omega) dP^*(\omega) + \int_\Omega [\mathcal{L}_\lambda(\omega)]_+ d\mu_2^*(\omega) - \int_\Omega [-\mathcal{L}_\lambda(\omega)]_+ d\mu_1^*(\omega), \tag{2.8}$$

where $[a]_+ := \max\{a, 0\}$.

By the standard theory of Lagrangian duality we have that the optimal value of problem (2.3) is always less than or equal to the optimal value of its dual (2.7). It is possible to give various regularity conditions (constraint qualifications) ensuring that the optimal values of problem (2.3) and its dual (2.7) are equal to each other, i.e., that there is no duality gap between problems (2.3) and (2.7). For example, we have (by the theory of conjugate duality [17]) that there is no duality gap between (2.3) and (2.7), and the set of optimal solutions of the dual problem is nonempty and bounded, if and only if the following assumption holds:

(A3) The optimal value of (2.2) is finite, and there exists a feasible solution to (2.2) for all sufficiently small perturbations of the right-hand sides of the (equality and inequality) constraints.

We may refer to [10] (and references therein) for a discussion of constraint qualifications ensuring the "no duality gap" property in the problem of moments.

By the above discussion we have the following result.

PROPOSITION 2.1. *Suppose that the assumptions* (A1)–(A3) *hold. Then problems* (2.2) *and* (2.3) *are equivalent and there is no duality gap between problem* (2.3) *and its dual* (2.7).

Remark 1. The preceding analysis simplifies considerably if the set Ω is finite, say, $\Omega := \{\omega_1, \ldots, \omega_K\}$. Then a measure $P \in \mathcal{X}$ can be identified with a vector $p = (p_1, \ldots, p_K) \in \mathbb{R}^K$. We have, of course, that $P \succeq 0$ if and only if $p_k \geq 0$, $k = 1, \ldots, K$. The set $\mathcal{M}$ can be written in the form

$$\mathcal{M} = \left\{p \in \mathbb{R}^K : \mu_k^1 \leq p_k \leq \mu_k^2,\ k = 1, \ldots, K\right\}$$

for some numbers $\mu_k^2 \geq \mu_k^1 \geq 0$, $k = 1, \ldots, K$, and problems (2.2) and (2.3) become linear programming problems. In that case the optimal values of problem (2.2) (problem (2.3)) and its dual (2.7) are equal to each other by the standard linear programming duality without a need for constraint qualifications, and the assumption (A3) is superfluous.

Let us now consider, further, a specific case of (2.2), where

$$\mathcal{M} := \left\{\mu \in \mathcal{X} : (1 - \varepsilon_1)P^* \preceq \mu \preceq (1 + \varepsilon_2)P^*\right\}; \tag{2.9}$$

i.e., $\mu_1 = (1 - \varepsilon_1)P^*$ and $\mu_2 = (1 + \varepsilon_2)P^*$ for some reference probability measure P^* satisfying assumption (A2) and numbers $\varepsilon_1 \in [0, 1]$, $\varepsilon_2 \geq 0$. In that case the dual problem (2.7) takes the form

$$\begin{array}{ll} \mathrm{Min}_{\lambda \in \mathbb{R}^{q+1}} & \mathbb{E}_{P^*}\Big\{\varphi_0(\omega) + \eta_{\varepsilon_1,\varepsilon_2}\left[\mathcal{L}_\lambda(\omega)\right]\Big\} \\ \text{subject to} & \lambda_i \geq 0, \quad i = r + 1, \ldots, q, \end{array} \tag{2.10}$$

where $\mathcal{L}_\lambda(\omega)$ is defined in (2.6) and

$$\eta_{\varepsilon_1,\varepsilon_2}[a] := \begin{cases} -\varepsilon_1 a & \text{if } a \leq 0, \\ \varepsilon_2 a & \text{if } a > 0. \end{cases} \tag{2.11}$$

Note that the function $\eta_{\varepsilon_1,\varepsilon_2}[\cdot]$ is convex piecewise linear and $\mathcal{L}_\lambda(\omega)$ is affine in λ for every $\omega \in \Omega$. Consequently the objective function of (2.10) is convex in λ. Thus, the problem of moments (2.2) has been reformulated as a convex stochastic programming problem (involving optimization of the expectation functional) of the form (1.1).

3. A class of minimax stochastic programs. We consider a specific class of minimax stochastic programming problems of the form

$$\mathrm{Min}_{x \in X} f(x), \tag{3.1}$$

where $f(x)$ is the optimal value of the problem

$$\begin{array}{ll} \mathrm{Max}_{P \in \mathcal{M}} & \int_\Omega F(x, \omega) dP(\omega) \\ \text{subject to} & \int_\Omega dP(\omega) = 1, \\ & \int_\Omega \varphi_i(\omega) dP(\omega) = b_i, \quad i = 1, \ldots, r, \\ & \int_\Omega \varphi_i(\omega) dP(\omega) \leq b_i, \quad i = r + 1, \ldots, q, \end{array} \tag{3.2}$$

and $\mathcal{M}$ is defined as in (2.9). Of course, this is a particular form of the minimax stochastic programming problem (1.2) with the set $\mathcal{P}$ formed by probability measures $P \in \mathcal{M}$ satisfying the corresponding moment constraints.

We assume that the set X is nonempty and assumptions (A1)–(A3) of section 2 hold for the functions $\varphi_i(\cdot)$, $i = 1, \ldots, q$, and $\varphi_0(\cdot) := F(x, \cdot)$ for all $x \in X$. By the analysis of section 2 (see Proposition 2.1 and dual formulation (2.10)) we then have that the minimax problem (3.1) is equivalent to the stochastic programming problem:

$$\begin{array}{ll} \mathrm{Min}_{(x,\lambda)\in\mathbb{R}^{n+q+1}} & \mathbb{E}_{P^*}[H(x,\lambda,\omega)] \\ \text{subject to} & x \in X \ \text{ and } \ \lambda_i \geq 0, \quad i = r+1, \ldots, q, \end{array} \tag{3.3}$$

where

$$H(x,\lambda,\omega) := F(x,\omega) + \eta_{\varepsilon_1,\varepsilon_2}\left[F(x,\omega) - \lambda_0 - \sum_{i=1}^{q} \lambda_i \varphi_i(\omega)\right]. \tag{3.4}$$

Note that by reformulating the minimax problem (3.1) into problem (3.3), which is a standard stochastic program involving optimization of an expectation functional, we avoid explicit solution of the inner maximization problem with respect to the probability measures. The reformulation, however, introduces $q + 1$ additional variables.

For problems with a prohibitively large (or possibly infinite) support Ω, a simple but effective approach to attacking (3.3) is the *sample average approximation* (SAA) method. The basic idea of this approach is to replace the expectation functional in the objective by a sample average function and to solve the corresponding SAA problem. Depending on the structure of the objective function $F(x, \omega)$ and hence $H(x, \lambda, \omega)$, a number of existing stochastic programming algorithms can be applied to solve the obtained SAA problem. Under mild assumptions, the SAA method has been shown to have attractive convergence properties. For example, a solution to the SAA problem quickly converges to a solution to the true problem as the sample size N is increased. Furthermore, by repeated solutions of the SAA problem, statistical confidence intervals on the quality of the corresponding SAA solutions can be obtained. Detailed discussion of the SAA method can be found in [18, Chapter 6] and references therein.

3.1. Stochastic programs with convex objectives. In this section, we consider minimax stochastic programs (3.1) corresponding to stochastic programs where the objective function is convex. Note that if the function $F(\cdot, \omega)$ is convex for every $\omega \in \Omega$, then the function $f(\cdot)$, defined as the optimal value of (3.2), is given by the maximum of convex functions and hence is convex. Not surprisingly, the reformulation preserves convexity.

PROPOSITION 3.1. *Suppose that the function $F(\cdot, \omega)$ is convex for every $\omega \in \Omega$. Then for any $\varepsilon_1 \in [0, 1]$ and $\varepsilon_2 \geq 0$ and every $\omega \in \Omega$, the function $H(\cdot, \cdot, \omega)$ is convex and*

$$\partial H(x,\lambda,\omega) = \begin{cases} (1-\varepsilon_1)\partial F(x,\omega) \times \{\varepsilon_1 \varphi(\omega)\} & \text{if } N(x,\lambda,\omega) < 0, \\ (1+\varepsilon_2)\partial F(x,\omega) \times \{-\varepsilon_2 \varphi(\omega)\} & \text{if } N(x,\lambda,\omega) > 0, \\ \cup_{\tau\in[-\varepsilon_1,\varepsilon_2]}(1+\tau)\partial F(x,\omega) \times \{-\tau\varphi(\omega)\} & \text{if } N(x,\lambda,\omega) = 0, \end{cases} \tag{3.5}$$

where the subdifferentials $\partial H(x, \lambda, \omega)$ and $\partial F(x, \omega)$ are taken with respect to (x, λ) and x, respectively, and

$$N(x,\lambda,\omega) := F(x,\omega) - \lambda_0 - \sum_{i=1}^{q} \lambda_i \varphi_i(\omega), \quad \varphi(\omega) := (1, \varphi_1(\omega), \ldots, \varphi_q(\omega)).$$

Proof. Consider function $\psi(z) := z + \eta_{\varepsilon_1,\varepsilon_2}[z]$. We can write

$$H(x,\lambda,\omega) = \psi\big(N(x,\lambda,\omega)\big) + \lambda_0 + \sum_{i=1}^{q} \lambda_i \varphi_i(\omega).$$

For any $\omega \in \Omega$, the function $N(\cdot,\cdot,\omega)$ is convex, and for $\varepsilon_1 \in [0,1]$ and $\varepsilon_2 \geq 0$, the function $\psi(\cdot)$ is monotonically nondecreasing and convex. Convexity of $H(\cdot,\cdot,\omega)$ then follows. The subdifferential formula (3.5) is obtained by the chain rule. □

Let us now consider instances of (3.3) with a finite set of realizations of ω:

$$\begin{array}{ll} \mathrm{Min}_{(x,\lambda)\in\mathbb{R}^{n+q+1}} & \left\{ h(x,\lambda) := \displaystyle\sum_{k=1}^{K} p_k^* H(x,\lambda,\omega_k) \right\} \\ \text{subject to} & x \in X \ \text{ and } \ \lambda_i \geq 0, \quad i = r+1,\ldots,q, \end{array} \tag{3.6}$$

where $\Omega = \{\omega_1,\ldots,w_K\}$ and $P^* = (p_1^*,\ldots,p_K^*)$. The above problem can either correspond to a problem with finite support of ω or may be obtained by sampling as in the SAA method. Problem (3.6) has a nonsmooth convex objective function, and often can be solved by using cutting plane or bundle type methods that use subgradient information (see, e.g., [8]). By the Moreau–Rockafellar theorem we have that

$$\partial h(x,\lambda) = \sum_{k=1}^{K} p_k^* \partial H(x,\lambda,\omega_k), \tag{3.7}$$

where all subdifferentials are taken with respect to (x,λ). Together with (3.5) this gives a formula for a subgradient of $h(\cdot,\cdot)$, given subgradient information for $F(\cdot,\omega)$.

3.2. Two-stage stochastic programs. A wide variety of stochastic programs correspond to optimization of the expected value of a future optimization problem. That is, let $F(x,\omega)$ be defined as the optimal value function

$$F(x,\omega) := \mathrm{Min}_{y\in Y(x,\omega)}\, G_0(x,y,\omega), \tag{3.8}$$

where

$$Y(x,\omega) := \{y \in Y : G_i(x,y,\omega) \leq 0,\ i = 1,\ldots,m\}, \tag{3.9}$$

Y is a nonempty subset of a finite dimensional vector space, and $G_i(x,y,\omega)$, $i = 0,\ldots,m$, are real valued functions. Problem (1.1), with $F(x,\omega)$ given in the form (3.8), is referred to as a two-stage stochastic program, where the first-stage variables x are decided prior to the realization of the uncertain parameters, and the second-stage variables y are decided after the uncertainties are revealed. The following result shows that a minimax problem corresponding to a two-stage stochastic program is itself a two-stage stochastic program.

PROPOSITION 3.2. *If $F(x,\omega)$ is defined as in* (3.8)*, then the function $H(x,\lambda,\omega)$, defined in* (3.4)*, is given by*

$$H(x,\lambda,\omega) = \inf_{y\in Y(x,\omega)} \mathcal{G}(x,\lambda,y,\omega), \tag{3.10}$$

where

$$\mathcal{G}(x,\lambda,y,\omega) := G_0(x,y,\omega) + \eta_{\varepsilon_1,\varepsilon_2}\left[G_0(x,y,\omega) - \lambda_0 - \sum_{i=1}^{q} \lambda_i \varphi_i(\omega) \right]. \tag{3.11}$$

Proof. The result follows by noting that

$$\mathcal{G}(x,\lambda,y,\omega)=\psi\left(G_0(x,y,\omega)-\lambda_0-\sum_{i=1}^{q}\lambda_i\varphi_i(\omega)\right)+\lambda_0+\sum_{i=1}^{q}\lambda_i\varphi_i(\omega),$$

and the function $\psi(z):=z+\eta_{\varepsilon_1,\varepsilon_2}[z]$ is monotonically nondecreasing for $\varepsilon_1\le 1$ and $\varepsilon_2\ge 0$. □

By the above result, if the set $\Omega:=\{\omega_1,\ldots,\omega_K\}$ is finite, then the reformulated minimax problem (3.3) can be written as one large-scale optimization problem:

$$\begin{array}{lll}\mathrm{Min}_{x,\lambda,y_1,\ldots,y_K} & \displaystyle\sum_{k=1}^{K}p_k^*\mathcal{G}(x,\lambda,y_k,\omega_k) & \\ \text{subject to} & y_k\in Y(x,\omega_k), & k=1,\ldots,K,\\ & x\in X,\ \lambda_i\ge 0, & i=r+1,\ldots,q.\end{array}\tag{3.12}$$

A particularly important case of two-stage stochastic programs are the two-stage stochastic (mixed-integer) linear programs, where $F(x,\omega):=V(x,\xi(\omega))$ and $V(x,\xi)$ is given by the optimal value of the problem:

$$\begin{array}{ll}\mathrm{Min}_y & c^Tx+q^Ty,\\ \text{subject to} & Wy=h-Tx,\quad y\in Y.\end{array}\tag{3.13}$$

Here $\xi:=(q,W,h,T)$ represents the uncertain (random) parameters of problem (3.13), and X and Y are defined by linear constraints (and possibly with integrality restrictions). By applying standard linear programming modelling principles to the piecewise linear function $\eta_{\varepsilon_1,\varepsilon_2}$, we obtain that $H(x,\lambda,\xi(\omega))$ is given by the optimal value of the problem:

$$\begin{array}{ll}\mathrm{Min}_{y,u^+,u^-} & c^Tx+q^Ty+\varepsilon_1u^-+\varepsilon_2u^+\\ \text{subject to} & Wy=h-Tx,\\ & u^+-u^-=c^Tx+q^Ty-\varphi^T\lambda,\\ & y\in Y,\ u^+\ge 0,\quad u^-\ge 0,\end{array}\tag{3.14}$$

where $\varphi:=(1,\varphi_1(\omega),\ldots,\varphi_q(\omega))^T$. As before, if the set $\Omega:=\{\omega_1,\ldots,\omega_K\}$ is finite, then the reformulated minimax problem (3.3) can be written as one large-scale mixed-integer linear program:

$$\begin{array}{lll}\mathrm{Min}_{x,\lambda,y,u^+,u^-} & c^Tx+\displaystyle\sum_{k=1}^{K}p_k^*\left(q_k^Ty_k+\varepsilon_1u_k^-+\varepsilon_2u_k^+\right) & \\ \text{subject to} & W_ky_k=h_k-T_kx, & k=1,\ldots,K,\\ & u_k^+-u_k^-=c^Tx+q_k^Ty_k-\varphi_k^T\lambda, & k=1,\ldots,K,\\ & y_k\in Y,\ u_k^+\ge 0,\ u_k^-\ge 0, & k=1,\ldots,K,\\ & x\in X. & \end{array}\tag{3.15}$$

The optimization model stated above has a block-separable structure which can, in principle, be exploited by existing decomposition algorithms for stochastic (integer) programs. In particular, if Y does not have any integrality restrictions, then the L-shaped (or Benders) decomposition algorithm and its variants can be immediately applied (see, e.g., [18, Chapter 3]).

4. Connection to a class of mean-risk models. Note that the stochastic program (1.1) is risk-neutral in the sense that it is concerned with the optimization of an expectation objective. To extend the stochastic programming framework to a risk-averse setting, one can adopt the *mean-risk* framework advocated by Markowitz and further developed by many others. In this setting the model (1.1) is extended to

$$\operatorname*{Min}_{x\in X} \ \mathbb{E}[F(x,\omega)] + \gamma\mathcal{R}[F(x,\omega)], \tag{4.1}$$

where $\mathcal{R}[Z]$ is a dispersion statistic of the random variable Z used as a measure of risk, and γ is a weighting parameter to trade-off mean with risk. Classically, the variance statistic has been used as the risk-measure. However, it is known that many typical dispersion statistics, including variance, may cause the mean-risk model (4.1) to provide inferior solutions. That is, an optimal solution to the mean-risk model may be stochastically dominated by another feasible solution. Recently, Ogryczak and Ruszczyński [15] have identified a number of statistics which, when used as the risk-measure $\mathcal{R}[\cdot]$ in (4.1), guarantee that the mean-risk solutions are consistent with stochastic dominance theory. One such dispersion statistic is

$$h_\alpha[Z] := \mathbb{E}\big\{\alpha[Z-\kappa_\alpha]_+ + (1-\alpha)[\kappa_\alpha - Z]_+\big\}, \tag{4.2}$$

where $0 \le \alpha \le 1$ and $\kappa_\alpha = \kappa_\alpha(Z)$ denotes the α-quantile of the distribution of Z. Recall that κ_α is said to be an α-quantile of the distribution of Z if $Pr(Z < \kappa_\alpha) \le \alpha \le Pr(Z \le \kappa_\alpha)$, and the set of α-quantiles forms the interval $[a,b]$ with $a := \inf\{z : Pr(Z \le z) \ge \alpha\}$ and $b := \sup\{z : Pr(Z \ge z) \le \alpha\}$. In particular, if $\alpha = \frac{1}{2}$, then $\kappa_\alpha(Z)$ becomes the median of the distribution of Z and

$$h_\alpha[Z] = \tfrac{1}{2}\mathbb{E}\left|Z - \kappa_{1/2}\right|,$$

and it represents half of the mean absolute deviation from the median.

In [15], it is shown that mean-risk models (4.1), with $\mathcal{R}[\cdot] := h_\alpha[\cdot]$ and $\gamma \in [0,1]$, provide solutions that are consistent with stochastic dominance theory. In the following, we show that minimax models (3.3) provide a new insight into mean-risk models (4.1).

Consider functions $\mathcal{L}_\lambda(\omega)$ and $\eta_{\varepsilon_1,\varepsilon_2}[a]$, defined in (2.6) and (2.11), respectively. These functions can be written in the form

$$\mathcal{L}_\lambda(\omega) = Z(\omega) - \lambda_0 \ \text{ and } \ \eta_{\varepsilon_1,\varepsilon_2}[a] = (\varepsilon_1+\varepsilon_2)\left(\alpha[a]_+ + (1-\alpha)[-a]_+\right),$$

where $Z(\omega) := \varphi_0(\omega) - \sum_{i=1}^q \lambda_i\varphi_i(\omega)$ and $\alpha := \varepsilon_2/(\varepsilon_1+\varepsilon_2)$, and hence

$$\eta_{\varepsilon_1,\varepsilon_2}\left[\mathcal{L}_\lambda(\omega)\right] = (\varepsilon_1+\varepsilon_2)\left(\alpha[Z(\omega)-\lambda_0]_+ + (1-\alpha)[\lambda_0 - Z(\omega)]_+\right). \tag{4.3}$$

We obtain that for fixed λ_i, $i = 1,\dots,q$, and positive ε_1 and ε_2, a minimizer $\bar\lambda_0$ of $\mathbb{E}_{P^*}\left\{\eta_{\varepsilon_1,\varepsilon_2}\left[\mathcal{L}_\lambda(\omega)\right]\right\}$ over $\lambda_0 \in \mathbb{R}$ is given by an α-quantile of the distribution of the random variable $Z(\omega)$, defined on the probability space $(\Omega,\mathcal{F},P^*)$. In particular, if $\varepsilon_1 = \varepsilon_2$, then $\bar\lambda_0$ is the median of the distribution of Z. It follows that if ε_1 and ε_2 are positive, then the minimum of the expectation in (3.3), with respect to $\lambda_0 \in \mathbb{R}$, is attained at an α-quantile of the distribution of $F(x,\omega) - \sum_{i=1}^q \lambda_i\varphi_i(\omega)$ with respect to the probability measure P^*. In particular, if the moment constraints are not present in (3.2), i.e., $q = 0$, then problem (3.3) can be written as follows:

$$\operatorname*{Min}_{x\in X} \mathbb{E}_{P^*}\big[F(x,\omega)\big] + (\varepsilon_1+\varepsilon_2)h_\alpha[F(x,\omega)], \tag{4.4}$$

where h_α is defined as in (4.2). The above discussion leads to the following result.

PROPOSITION 4.1. *The mean-risk model* (4.1) *with* $\mathcal{R}[\cdot] := h_\alpha[\cdot]$ *is equivalent to the minimax model* (3.3) *with* $\varepsilon_1 = \gamma(1-\alpha)$, $\varepsilon_2 = \alpha\gamma$, *and* $q = 0$.

The additional term $(\varepsilon_1 + \varepsilon_2)h_\alpha[F(x,\omega)]$, which appears in (4.4), can be interpreted as a regularization term. We conclude this section by discussing the effect of such regularization.

Consider the case when the function $F(\cdot,\omega)$ is convex and piecewise linear for all $\omega \in \Omega$. This is the case, for example, when $F(x,\omega)$ is the value function of the second-stage linear program (3.13) without integrality restrictions. Consider the stochastic programming problem (with respect to the reference probability distribution P^*)

$$\operatorname*{Min}_{x\in X} \mathbb{E}_{P^*}[F(x,\omega)] \tag{4.5}$$

and the corresponding mean-risk or minimax model (4.4). Suppose that X is polyhedral, the support Ω of ω is finite, and both problems (4.4) and (4.5) have finite optimal solutions. Then from the discussion at the end of section 3, the problems (4.4) and (4.5) can be stated as linear programs. Let S_0 and $S_{\varepsilon 1,\varepsilon 2}$ denote the sets of optimal solutions of (4.5) and (4.4), respectively. Then by standard theory of linear programming, we have that, for all $\varepsilon_1 > 0$ and $\varepsilon_2 > 0$ sufficiently small, the inclusion $S_{\varepsilon_1,\varepsilon_2} \subset S_0$ holds. Consequently, the term $(\varepsilon_1 + \varepsilon_2)h_\alpha[F(x,\omega)]$ has the effect of regularizing the solution set of the stochastic program (4.5). We further illustrate this regularization with an example.

Example 1. Consider the function $F(x,\omega) := |\omega - x|$, $x, w \in \mathbb{R}$, with ω having the reference distribution $P^*(\omega = -1) = p_1^*$ and $P^*(\omega = 1) = p_2^*$ for some $p_1^* > 0$, $p_2^* > 0$, $p_1^* + p_2^* = 1$. We then have that

$$\mathbb{E}_{P^*}[F(x,\omega)] = p_1^*|1+x| + p_2^*|1-x|.$$

Let us first discuss the case where $p_1^* = p_2^* = \frac{1}{2}$. Then the set S_0 of optimal solutions of the stochastic program (4.5) is given by the interval $[-1,1]$. For $\varepsilon_2 > \varepsilon_1$ and $\varepsilon_1 \in (0,1)$, the corresponding α-quantile $\kappa_\alpha(F(x,\omega))$, with $\alpha := \varepsilon_2/(\varepsilon_1+\varepsilon_2)$, is equal to the largest of the numbers $|1-x|$ and $|1+x|$, and for $\varepsilon_2 = \varepsilon_1$ the set of α-quantiles is given by the interval with the end points $|1-x|$ and $|1+x|$. It follows that, for $\varepsilon_2 \geq \varepsilon_1$, the mean-risk (or minimax) objective function in problem (4.4),

$$f(x) := \mathbb{E}_{P^*}\big[F(x,\omega)\big] + (\varepsilon_1+\varepsilon_2)h_\alpha[F(x,\omega)],$$

is given by

$$f(x) = \begin{cases} \frac{1}{2}(1-\varepsilon_1)|1-x| + \frac{1}{2}(1+\varepsilon_1)|1+x| & \text{if } x \geq 0, \\ \frac{1}{2}(1+\varepsilon_1)|1-x| + \frac{1}{2}(1-\varepsilon_1)|1+x| & \text{if } x < 0. \end{cases}$$

Consequently, $S_{\varepsilon_1,\varepsilon_2} = \{0\}$. Note that for $x = 0$, the random variable $F(x,\omega)$ has minimal expected value and variance zero (with respect to the reference distribution P^*). Therefore it is not surprising that $x = 0$ is the unique optimal solution of the mean-risk or minimax problem (4.4) for any $\varepsilon_1 \in (0,1)$ and $\varepsilon_2 > 0$.

Suppose now that $p_2^* > p_1^*$. In that case $S_0 = \{1\}$. Suppose, further, that $\varepsilon_1 \in (0,1)$ and $\varepsilon_2 \geq \varepsilon_1$, and hence $\alpha \geq \frac{1}{2}$. Then for $x \geq 0$ the corresponding α-quantile $\kappa_\alpha(F(x,\omega))$ is equal to $|1-x|$ if $\alpha < p_2^*$, $\kappa_\alpha(F(x,\omega)) = 1+x$ if $\alpha > p_2^*$, and $\kappa_\alpha(x)$ can be any point on the interval $\big[|1-x|, 1+x\big]$ if $\alpha = p_2^*$. Consequently, for $\alpha \leq p_2^*$ and $x \geq 0$,

$$f(x) = (p_1^* + \varepsilon_2 p_1^*)(1+x) + (p_2^* - \varepsilon_2 p_1^*)|1-x|.$$

It follows then that $S_{\varepsilon_1,\varepsilon_2} = \{1\}$ if and only if $p_1^* + \varepsilon_2 p_1^* < p_2^* - \varepsilon_2 p_1^*$. Since $\alpha \le p_2^*$ means that $\varepsilon_2 \le (p_2^*/p_1^*)\varepsilon_1$, we have that for ε_2 in the interval $[\varepsilon_1, (p_2^*/p_1^*)\varepsilon_1]$, the set $S_{\varepsilon_1,\varepsilon_2}$ coincides with S_0 if and only if $\varepsilon_2 < (p_2^*/p_1^* - 1)/2$. For ε_2 in this interval we can view $\bar{\varepsilon}_2 := (p_2^*/p_1^* - 1)/2$ as the breaking value of the parameter ε_2; i.e., for ε_2 bigger than $\bar{\varepsilon}_2$ an optimal solution of the minimax problem moves away from the optimal solution of the reference problem.

Suppose now that $p_2^* > p_1^*$ and $\alpha \ge p_2^*$. Then for $x \ge 0$,

$$f(x) = (p_1^* + \varepsilon_1 p_2^*)(1 + x) + (p_2^* - \varepsilon_1 p_2^*)|1 - x|.$$

In that case the breaking value of ε_1, for $\varepsilon_1 \le (p_1^*/p_2^*)\varepsilon_2$, is $\bar{\varepsilon}_1 := (1 - p_1^*/p_2^*)/2$.

5. Numerical results. In this section we describe some numerical experiments with the proposed minimax stochastic programming model. We consider minimax extensions of two-stage stochastic linear programs with finite support of the random problem parameters. We assume that $q = 0$ (i.e., that the moment constraints are not present in the model) since, in this case, the minimax problems are equivalent to mean-risk extensions of the stochastic programs, where risk is measured in terms of quantile deviations.

Recall that, owing to the finiteness of the support, the minimax problems reduce to the specially structured linear programs (3.15). We use an ℓ_∞–trust-region based decomposition algorithm for solving the resulting linear programs. The method along with its theoretical convergence properties is described in [12]. The algorithm has been implemented in ANSI C with the GNU Linear Programming Kit (GLPK) [14] library routines to solve linear programming subproblems. All computations have been carried out on a Linux workstation with dual 2.4 GHz Intel Xeon processors and 2 GB RAM.

The stochastic linear programming test problems in our experiments are derived from those used in [11]. We consider the problems `LandS`, `gbd`, `20term`, and `storm`. Data for these instances are available from the website http://www.cs.wisc.edu/∼swright/stochastic/sampling. These problems involve extremely large numbers of scenarios (joint realizations of the uncertain problem parameters). Consequently, for each problem, we consider three instances each with 1000 sampled scenarios. The reference distribution P^* for these instances corresponds to equal weights assigned to each sampled scenario.

Recall that a minimax model with parameters ε_1 and ε_2 is equivalent to a mean-risk model (involving quantile deviations) with parameters $\gamma := \varepsilon_1 + \varepsilon_2$ and $\alpha := \varepsilon_2/(\varepsilon_1 + \varepsilon_2)$. We consider α values of 0.5, 0.7, and 0.9, and ε_1 values of 0.1, 0.3, 0.5, 0.7, and 0.9. Note that the values of the parameters ε_2 and γ are uniquely determined by $\varepsilon_2 = \alpha\varepsilon_1/(1-\alpha)$ and $\gamma = \varepsilon_1/(1-\alpha)$. Note also that some combinations of ε_1 and α are such that $\gamma > 1$, and consequently the resulting solutions are not guaranteed to be consistent with stochastic dominance.

First, for each problem, the reference stochastic programming models (with $\varepsilon_1 = \varepsilon_2 = 0$) corresponding to all three generated instances were solved. Next, the minimax stochastic programming models for the various ε_1-α combinations were solved for all instances. Various dispersion statistics corresponding to the optimal solutions (from the different models) with respect to the reference distribution P^* were computed. Table 5.1 presents the results for the reference stochastic program corresponding to the four problems. The first six rows of the table display various cost-statistics corresponding to the optimal solution with respect to P^*. The presented data is the average over the three instances. The terms "Abs Med-Dev," "Abs Dev," "Std Dev,"

"Abs SemiDev," and "Std SemiDev" stand for the statistics mean absolute deviation from the median, mean absolute deviation, standard deviation, absolute semideviation, and standard semideviation, respectively. The last two rows of the table display the average (over the three instances) number of iterations and CPU seconds required. Tables 5.2–5.4 present the results for the problem `LandS`. Each table in this set corresponds to a particular α value, and each column in a table corresponds to a particular ε_1 value. The statistics are organized in the rows as in Table 5.1. Similar results are available from the authors for the problems `gbd`, `20term`, and `storm`. In Table 5.5, we present the statistics corresponding to $\alpha = 0.7$ and $\varepsilon_1 = 0.5$ for these three problems.

For a fixed level of α, increasing ε_1 corresponds to increasing the allowed perturbation of the reference distribution in the minimax model, and to increasing the weight γ for the risk term in the mean-risk model. Consequently, we observe from the tables that this leads to solutions with higher expected costs. We also observe that the value of some of the dispersion statistics decreases, indicating a reduction in risk. Similar behavior occurs upon increasing α corresponding to a fixed level of ε_1.

A surprising observation from the numerical results is that the considered problem instances are very robust with respect to perturbations of the reference distribution P^*. Even with large perturbations of the reference distribution, the perturbations of the optimal objective function values are relatively small.

A final observation from the tables is the large variability of computational effort for the various ε_1-α combinations. This can be somewhat explained by the regularization nature of the minimax (or mean-risk) objective function as discussed in section 4. For certain ε_1-α combinations, the piecewise linear objective function may become very sharp, resulting in faster convergence of the algorithm.

TABLE 5.1
Statistics corresponding to the reference stochastic program.

	`LandS`	`gbd`	`20term`	`storm`
Expected cost	225.52	1655.54	254147.15	15498557.91
Abs Med-Dev	46.63	502.01	10022.59	304941.12
Abs Dev	46.95	539.63	10145.86	313915.60
Std Dev	59.26	715.33	12079.76	371207.13
Abs SemiDev	23.47	269.81	5072.93	156957.80
Std SemiDev	44.55	605.01	8824.36	261756.11
Iterations	47.33	57.33	275.33	5000.00
CPU seconds	0.67	0.67	32.33	2309.33

TABLE 5.2
Statistics for problem `LandS` *with* $\alpha = 0.5$.

	$\varepsilon_1 = 0.1$	$\varepsilon_1 = 0.3$	$\varepsilon_1 = 0.5$	$\varepsilon_1 = 0.7$	$\varepsilon_1 = 0.9$
Expected cost	225.57	225.74	225.99	226.39	226.95
Abs Med-Dev	45.91	45.03	44.41	43.74	43.04
Abs Dev	46.24	45.38	44.70	44.15	43.47
Std Dev	58.28	57.16	56.41	55.63	54.84
Abs SemiDev	23.12	22.69	22.39	22.08	21.73
Std SemiDev	43.78	42.97	42.48	41.97	41.45
Iterations	3357.33	3357.00	75.00	70.00	67.33
CPU seconds	196.33	195.33	1.00	1.00	1.00

TABLE 5.3
Statistics for problem `LandS` *with* $\alpha = 0.7$.

	$\varepsilon_1 = 0.1$	$\varepsilon_1 = 0.3$	$\varepsilon_1 = 0.5$	$\varepsilon_1 = 0.7$	$\varepsilon_1 = 0.9$
Expected cost	225.603	225.86	226.31	226.92	227.73
Abs Med-Dev	45.69	44.76	43.91	43.14	42.32
Abs Dev	46.01	45.11	44.28	43.54	42.76
Std Dev	57.94	56.79	55.79	54.92	54.01
Abs SemiDev	23.00	22.55	22.14	21.77	21.38
Std SemiDev	43.47	42.69	42.02	41.44	40.85
Iterations	5000.00	72.67	64.67	70.67	68.00
CPU seconds	293.00	1.33	1.00	1.00	1.00

TABLE 5.4
Statistics for problem `LandS` *with* $\alpha = 0.9$.

	$\varepsilon_1 = 0.1$	$\varepsilon_1 = 0.3$	$\varepsilon_1 = 0.5$	$\varepsilon_1 = 0.7$	$\varepsilon_1 = 0.9$
Expected cost	225.66	226.23	227.16	228.24	228.72
Abs Med-Dev	45.44	44.06	42.93	42.06	41.76
Abs Dev	45.77	44.45	43.36	42.49	42.17
Std Dev	57.61	55.95	54.64	53.62	53.26
Abs SemiDev	22.88	22.23	21.68	21.25	21.09
Std SemiDev	43.21	42.13	41.27	40.54	40.28
Iterations	65.67	63.33	59.67	60.00	1700.33
CPU seconds	1.00	1.00	1.00	1.00	95.67

TABLE 5.5
Statistics for problems `gbd`, `20term`, *and* `storm` *with* $\alpha = 0.7$ *and* $\varepsilon_1 = 0.5$.

	`gbd`	`20term`	`storm`
Expected cost	1663.67	254545.40	15499225.25
Abs Med-Dev	483.94	9220.59	303585.26
Abs Dev	523.96	9360.18	312532.81
Std Dev	702.31	11002.47	369731.47
Abs SemiDev	261.98	4680.09	156266.40
Std SemiDev	598.71	7767.96	260501.73
Iterations	71.67	281.33	1718.33
CPU seconds	1.00	34.00	807.33

Acknowledgment. The authors are indebted to two anonymous referees for constructive comments which helped to improve the manuscript.

REFERENCES

[1] J. R. BIRGE AND R. J.-B. WETS, *Computing bounds for stochastic programming problems by means of a generalized moment problem*, Math. Oper. Res., 12 (1987), pp. 149–162.

[2] M. BRETON AND S. EL HACHEM, *Algorithms for the solution of stochastic dynamic minimax problems*, Comput. Optim. Appl., 4 (1995), pp. 317–345.

[3] J. DUPAČOVÁ, *Minimax stochastic programs with nonseparable penalties*, in Optimization Techniques (Proc. Ninth IFIP Conf., Warsaw, 1979), Part 1, Lecture Notes in Control and Inform. Sci. 22, Springer-Verlag, Berlin, 1980, pp. 157–163.

[4] J. DUPAČOVÁ, *The minimax approach to stochastic programming and an illustrative application*, Stochastics, 20 (1987), pp. 73–88.

[5] YU. ERMOLIEV, A. GAIVORONSKI, AND C. NEDEVA, *Stochastic optimization problems with incomplete information on distribution functions*, SIAM J. Control Optim., 23 (1985), pp. 697–716.

[6] A. A. GAIVORONSKI, *A numerical method for solving stochastic programming problems with moment constraints on a distribution function*, Ann. Oper. Res., 31 (1991), pp. 347–370.

[7] H. Gassmann and W. T. Ziemba, *A tight upper bound for the expectation of a convex function of a multi-variate random variable*, in Stochastic Programming 84.I, Math. Programming Stud. 27, North–Holland, Amsterdam, 1986, pp. 39–53.

[8] J.-B. Hiriart-Urruty and C. Lemaréchal, *Convex Analysis and Minimization Algorithms.* II, Springer-Verlag, Berlin, 1996.

[9] P. Kall, *An upper bound for SLP using first and total second moments*, Ann. Oper. Res., 30 (1991), pp. 670–682.

[10] W. K. Klein Haneveld, *Duality in Stochastic Linear and Dynamic Programming*, Lecture Notes in Econom. and Math. Systems 274, Springer-Verlag, Berlin, 1986.

[11] J. Linderoth, A. Shapiro, and S. Wright, *The empirical behavior for sampling methods for stochastic programming*, Ann. Oper. Res., to appear.

[12] J. Linderoth and S. Wright, *Decomposition algorithms for stochastic programming on a computational grid*, Comput. Optim. Appl., 24 (2003), pp. 207–250.

[13] H. J. Landau, ed., *Moments in Mathematics*, Proc. Sympos. Appl. Math. 37, AMS, Providence, RI, 1987.

[14] A. Makhorin, *GNU Linear Progamming Kit, Reference Manual*, Version 3.2.3, http://www.gnu.org/software/glpk/glpk.html, 2002.

[15] W. Ogryczak and A. Ruszczyński, *Dual stochastic dominance and related mean-risk models*, SIAM J. Optim., 13 (2002), pp. 60–78.

[16] M. Riis and K. A. Andersen, *Applying the Minimax Criterion in Stochastic Recourse Programs*, Technical report 2002/4, Department of Operations Research, University of Aarhus, Aarhus, Denmark, 2002.

[17] R. T. Rockafellar, *Conjugate Duality and Optimization*, CBMS-NSF Regional Conference Series in Applied Mathematics 16, SIAM, Philadelphia, 1974.

[18] A. Ruszczyński and A. Shapiro, eds., *Stochastic Programming*, Handbooks in Operations Research and Management Science 10, North–Holland, Amsterdam, 2003.

[19] A. Shapiro and A. Kleywegt, *Minimax analysis of stochastic programs*, Optim. Methods Softw., 17 (2002), pp. 523–542.

[20] J. E. Smith, *Generalized Chebychev inequalities: Theory and applications in decision analysis*, Oper. Res., 43 (1995), pp. 807–825.

[21] S. Takriti and S. Ahmed, *Managing Short-Term Electricity Contracts under Uncertainty: A Minimax Approach*, Technical report, School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA, 2002.

[22] J. Žáčková, *On minimax solution of stochastic linear programming problems*, Časopis Pěst. Mat., 91 (1966), pp. 423–430.